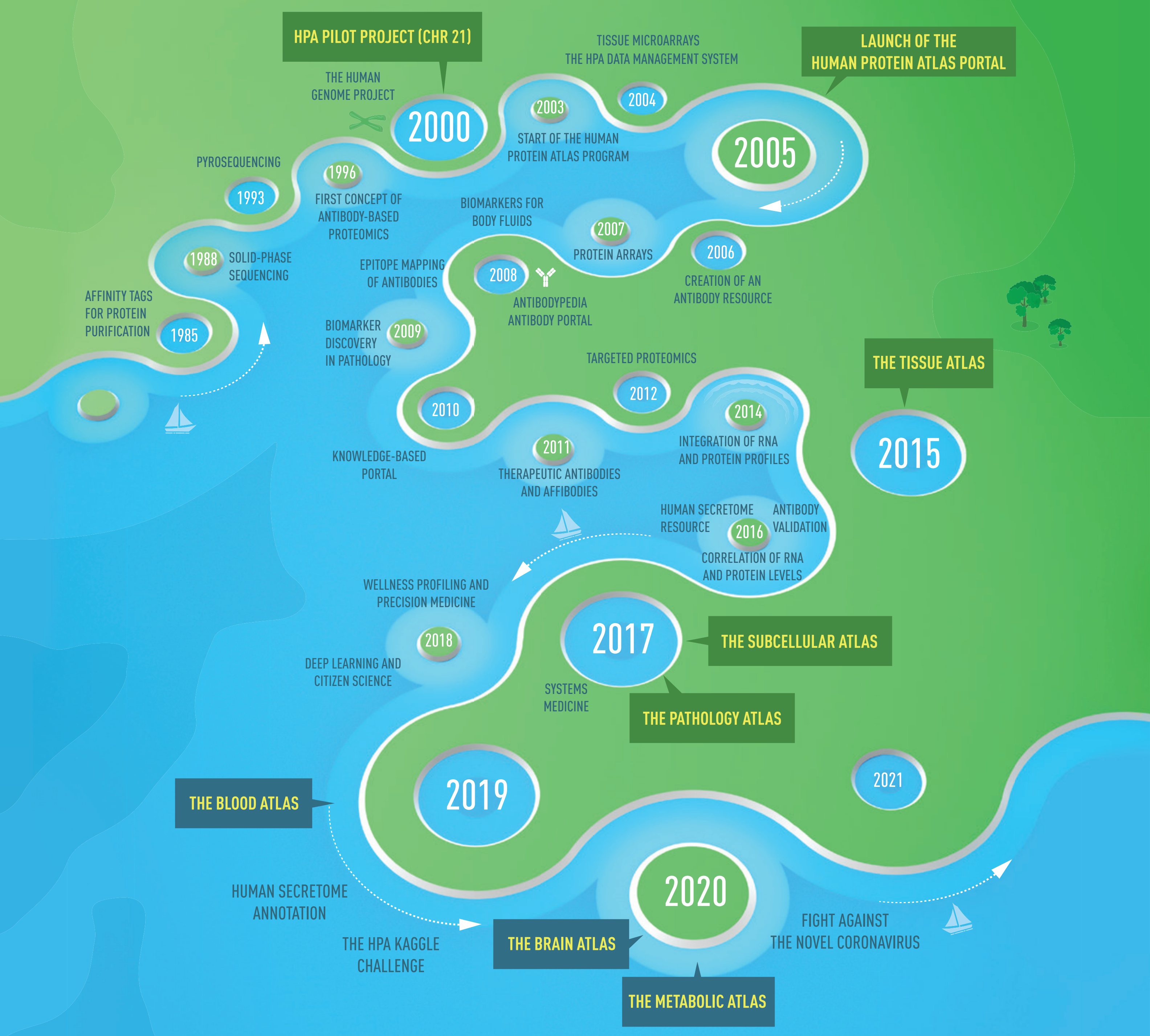
The Human Protein Atlas:
A 20-year journey into the body



1985
AFFINITY TAGS FOR PROTEIN PURIFICATION

1988
SOLID-PHASE SEQUENCING

1993
PYROSEQUENCING

1996
FIRST CONCEPT OF ANTIBODY-BASED PROTEOMICS

2000
HPA PILOT PROJECT (CHR 21)
THE HUMAN GENOME PROJECT

2003
START OF THE HUMAN PROTEIN ATLAS PROGRAM

2004
TISSUE MICROARRAYS
THE HPA DATA MANAGEMENT SYSTEM

2005
LAUNCH OF THE HUMAN PROTEIN ATLAS PORTAL

2006
CREATION OF AN ANTIBODY RESOURCE

2007
BIOMARKERS FOR BODY FLUIDS
PROTEIN ARRAYS

2008
EPITOPE MAPPING OF ANTIBODIES
ANTIBODYPEDIA ANTIBODY PORTAL

2009
BIOMARKER DISCOVERY IN PATHOLOGY

2010
KNOWLEDGE-BASED PORTAL

2011
THERAPEUTIC ANTIBODIES AND AFFIBODIES

2012
TARGETED PROTEOMICS

2014
INTEGRATION OF RNA AND PROTEIN PROFILES

2015
THE TISSUE ATLAS

2016
HUMAN SECRETOME RESOURCE
ANTIBODY VALIDATION
CORRELATION OF RNA AND PROTEIN LEVELS

2017
THE SUBCELLULAR ATLAS
THE PATHOLOGY ATLAS
SYSTEMS MEDICINE

2018
WELLNESS PROFILING AND PRECISION MEDICINE
DEEP LEARNING AND CITIZEN SCIENCE

2019
THE BLOOD ATLAS
HUMAN SECRETOME ANNOTATION
THE HPA KAGGLE CHALLENGE

2020
THE BRAIN ATLAS
THE METABOLIC ATLAS

2021
FIGHT AGAINST THE NOVEL CORONAVIRUS

Sponsored by
HPA

Produced by the *Science*/AAAS
Custom Publishing Office
Science | AAAS

Put Human Health at the
Heart of Your Research

Table of Contents

COVER IMAGE: Mattias Karlen

This booklet was produced by the *Science*/AAAS Custom Publishing Office and sponsored by The Human Protein Atlas.

Editor: Sean Sanders, Ph.D.
Proofreader/Copyeditor: Bob French
Designer: JD Huntsinger

ROGER GONCALVES,
ASSOCIATE SALES DIRECTOR
Custom Publishing
Europe, Middle East, and India
rgoncalves@science-int.co.uk
+41-43-243-1358

20 November 2020

Introductions

The Human Protein Atlas

Research Articles

Appreciations

The power of proteins

It was at a White House ceremony 20 years ago, on June 26, 2000, that President Clinton announced the successful sequencing of the human genome. Clinton was joined by Francis Collins, then director of the National Human Genome Research Institute, and Craig Venter, founder and CEO of Celera Genomics. At the time, there were high hopes that building this database of genomic knowledge would rapidly lead to new discoveries and new treatments for diseases. The heavy weight of reality, however, soon set in with the recognition that although we had the blueprint in hand, it didn't tell us much about the biochemical pathways that make cells tick or what, for instance, makes a liver cell so different from a neuron. This puzzle has been likened to having the instructions for manufacturing the millions of parts that make up a Boeing 747, but no guidance on how they all fit together.

Although the DNA code tells us *what* proteins a particular cell might make, it offers only subtle clues about where, when, and how much of that protein the cell should manufacture. Furthermore, it is these proteins, and not the DNA, that carry out cellular functions and serve as the final arbiters and gatekeepers for all biochemical processes.

The excitement of that announcement two decades ago led many to shift their focus toward genomics. Companies were even vying to copyright parts of DNA. Yet others were convinced that proteins ultimately held the key to understanding normal cellular function as well as disease-related dysfunction. Around the same time that the sequencing of the human genome was being announced, a group in Sweden led by Dr. Mathias Uhlén was embarking on the arguably more ambitious task of creating the Human Protein Atlas (HPA), a catalog of every protein in the human body. While it depended on sequence information provided by the Human Genome Project, the HPA program also augmented the genome efforts, shedding light on the apparent paradox that while only about 20,000 protein-encoding genes have been identified, it is clear that many more proteins exist. The atlas offered a free, open-access knowledge resource that gave researchers valuable insight into the expression levels and locations of all human proteins—and it continues to do so today.

In this expansive supplement, we take the reader on a journey through the 20 years since the HPA was established, stopping briefly to underscore milestones reached and to celebrate releases of updates and additions to the atlas. For the reader's convenience, we also provide reprints of the relevant journal articles associated with the most important of these milestones.

It is our hope that this publication will not only be a useful reference, but will also give readers a sense of the enormity of the task undertaken by the HPA project, the extraordinary advances that have resulted from it, and the crucial information that this database provides for researchers everywhere.

Sean Sanders, Ph.D.
Director and Senior Editor, Custom Publishing
Science/AAAS

20 years with the Human Protein Atlas

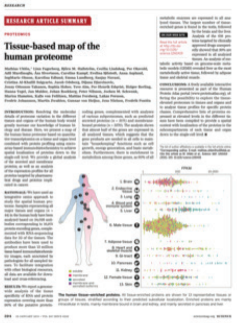
The Human Protein Atlas (HPA) journey started in 2000, the same year as the announcement of the completion of the Human Genome Project. The HPA pilot project was launched to map the proteins encoded by the genes on human chromosome 21 as a collaboration between researchers at KTH Royal Institute of Technology (Stockholm) and the startup company Affibody (Stockholm). In 2003, the program received generous support from the Knut and Alice Wallenberg Foundation (KAW), and the program was changed to an exclusively academic endeavor. The funding from KAW has continued to the present day. The project was expanded to engage research groups at KTH, Uppsala University, Karolinska Institutet, Chalmers University of Technology, and Lund University, and in addition, several international collaborations were initiated with research groups in Europe, the United States, South Korea, China, and India.

The aim of the program is to map all of the human proteins in cells, tissues, and organs, using integration of various 'omics technologies, including antibody-based imaging, mass spectrometry–based proteomics, transcriptomics, and systems biology. All the data in this knowledge resource is open-access, so that scientists in both academia and industry can freely use it to explore the human proteome. In 2005, the first version of the HPA was launched, and in 2010 the main operation of HPA was moved to the national infrastructure Science for Life Laboratory (Stockholm).

The HPA consists of various separate parts, each focusing on a particular aspect of analysis of human proteins as described in several articles in *Science* (see below). These parts include (1) the Tissue Atlas, showing the distribution of proteins across all major tissues and organs in the human body; (2) the Subcellular Atlas, showing the subcellular localization of proteins in single cells; (3) the Pathology Atlas, showing the impact of protein levels for survival of patients with cancer; (4) the Blood Atlas, showing the profiles of blood cells and proteins detectable in the blood; (5) the Brain Atlas, showing the distribution of proteins in human, mouse, and pig brain; and (6) the Metabolic Atlas, showing the presence of metabolic pathways across human tissues.

The HPA program has already contributed to thousands of publications in the field of human biology and disease and has been selected by the organization ELIXIR (www.elixireurope.org) as a European core resource, due to its fundamental importance for the wider life science community. Some of the most significant events in the history of the HPA consortium are described on the following pages, and below is a list of its major milestones during the past 20 years:

Tissue



RESEARCH ARTICLE SUMMARY

Tissue-based map of the human proteome

2015

Subcellular

RESEARCH ARTICLE SUMMARY

A subcellular map of the human proteome

2017

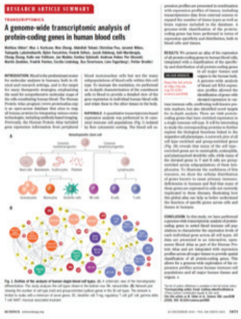
Cancer

RESEARCH ARTICLE SUMMARY

A pathology atlas of the human cancer transcriptome

2017

Blood



RESEARCH ARTICLE SUMMARY

A genome-wide transcriptomic analysis of protein-coding genes in human blood cells

2019

Brain

RESEARCH ARTICLE SUMMARY

An atlas of the protein-coding genes in the human, pig, and mouse brain

2020

2000
Pilot project: chromosome 21

2003
Funding from Knut and Alice Wallenberg Foundation

2005
Launch of first version of the Human Protein Atlas

2010
Milestone: half of the protein-coding genes analyzed

2015–2020
Milestones: launch of various databases (tissue, subcellular, cancer, blood, and brain)

Q&A

Getting to know the HPA

Q: What is the aim of the Human Protein Atlas?

Our aim is to create an open-access knowledge resource in which we have mapped the expression and location of all of the human proteins in cells, tissues, and organs, using integration of various 'omics technologies, including antibody-based imaging, mass spectrometry–based proteomics, transcriptomics, and systems biology.

Q: Is the data freely available?

Yes, all the data in the Human Protein Atlas (HPA) is open-access with no restrictions according to the Creative Commons Attribution-ShareAlike 3.0 International License.

Q: Is the data downloadable?

Yes, you can download our data in several ways and in a variety of formats. If you are interested in a single gene or a limited gene set, you can do a search on the start page, choose data from a wide range of columns, and directly download your search result in XML, RDF, TSV, or JSON format. If you are interested in genome-wide data, we have a download page.

Q: How is data from external sources used in the Atlas?

External data is both used and displayed in the Atlas. RNA expression data from the HPA, Genotype-Tissue Expression (GTEx), and Functional Annotation of the Mammalian Genome (FANTOM) portals are displayed on the gene-summary page for each gene along with the RNA specificity and distribution categories that are based on in-house RNA-seq data in combination with GTEx and FANTOM data. For all our antibody validations, UniProt protein and localization data as well as RNA-seq data from the external sources above are used as parameters. Furthermore, the Pathology Atlas contains data from The Cancer Genome Atlas (TCGA).

Q: How is the data validated?

We have a standard validation procedure for each individual application that includes concordance with available experimental data in the UniProt database, for example, but we also use some of the validation strategies described by the International Working Group for Antibody Validation (IWGAV) to provide an enhanced validation. The strategies include genetic validation, recombinant expression validation, independent antibody validation, orthogonal validation, and Capture MS validation, and are used for our Western blot, immunocytochemistry, and immunohistochemistry applications.

Q: Why are polyclonal antibodies mainly used in the Atlas instead of monoclonal antibodies?

In the HPA, we have mainly used polyclonal antibodies to analyze the location of proteins in cells, tissues, and organs. In applications such as immunohistochemistry, confocal microscopy, and Western blot, the samples encounter protein-denaturating conditions, such as heat, detergents, or solvent. Our experience is that antibodies recognizing several epitopes (polyclonal antibodies) have a better success rate than those recognizing a single epitope (monoclonal antibodies) for applications involving denaturation.

Q: Where do the HPA normal tissue samples originate?

The normal samples are from a biobank containing tissue excised from cancer patients. The samples are predominantly from noncancerous tissue surrounding the pathological tissue removed when resecting the tumor. After the samples are collected from the biobank, they are examined by a certified pathologist to ensure they are histologically normal.

Q: What are the survival scatter plots found in the Pathology Atlas?

This is a new concept for visualization of clinical data in which the survival of each patient in the study is shown as a result of the expression of a particular gene. This allows researchers to see the primary data for a particular gene and its consequences for the cancer patient summarized in a single plot. In our view, this is a good complement to the traditional Kaplan–Meier plots normally used to analyze clinical survival.

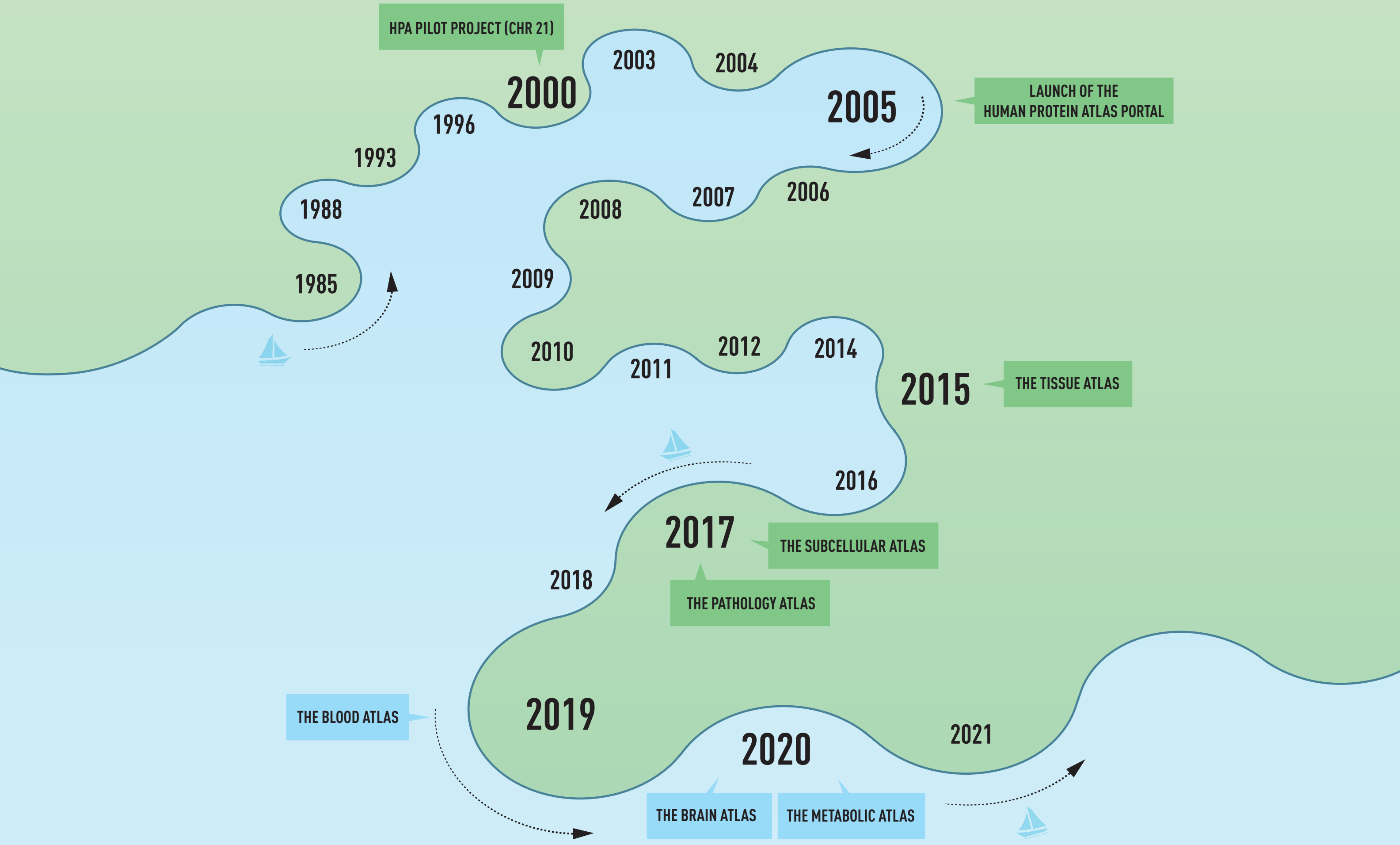
Q: Does the Atlas contain data other than human data?

Yes, the Brain Atlas contains mouse and pig data in order to give a more detailed view of brain regions, and some of the entries in the Cell Atlas contain location data for mouse cells.

Q: How can the search function be used?

The search function allows complex queries based on combining search parameters to obtain lists of genes that fit the query. One example of such a query could be, "Show me all genes enriched in the liver that encode a secreted protein."

Milestones on the HPA journey



SCIENTIFIC MILESTONES

Year	No.	Title
1985	1	Affinity tags for protein purification
1988	2	Solid-phase sequencing
1993	3	Pyrosequencing
1996	4	First concept of antibody-based proteomics
2000	5	The Human Genome Project
2000	6	Chromosome 21 pilot
2003	7	Start of the Human Protein Atlas program
2004	8	Tissue microarrays
2004	9	The HPA data management system
2005	10	Launch of the Human Protein Atlas portal
2006	11	Creation of an antibody resource
2007	12	Protein arrays
2008	13	Biomarkers for body fluids
2008	14	Epitope mapping of antibodies
2008	15	Antibodypedia antibody portal
2009	16	Biomarker discovery in pathology
2010	17	Knowledge-based portal
2011	18	Therapeutic antibodies and Affibodies
2012	19	Targeted proteomics
2014	20	Integration of RNA and protein profiles
2015	21	The Tissue Atlas
2016	22	Correlation of RNA and protein levels
2016	23	Human secretome resource
2016	24	Antibody validation
2017	25	The Subcellular Atlas
2017	26	The Pathology Atlas
2017	27	Systems medicine
2018	28	Wellness profiling and precision medicine
2018	29	Deep learning and citizen science
2019	30	Human secretome annotation
2019	31	The Blood Atlas
2019	32	The HPA Kaggle Challenge
2020	33	The Brain Atlas
2020	34	The Metabolic Atlas
2020	35	The fight against the novel coronavirus

Milestone 1

1985 Affinity tags for protein purification

Description:

The use of affinity tags for purification of recombinant fusion proteins was first described in 1983 using protein A as a purification tag. Affinity tags, including polyhistidine tags (His-tags), have since become widespread as versatile tools in bioscience, and tens of thousands of articles have been published on this concept. For the Human Protein Atlas (HPA) effort, the concept was used to generate more than 50,000 recombinant proteins with a histidine affinity tag. These proteins were in the HPA program used for immunizations to generate antibodies to most of the proteins in the human body.

Key publication:

B. Nilsson *et al.*, "Immobilization and purification of enzymes with staphylococcal protein A gene fusion vectors," *EMBO J.* **4**, 1075–1080 (1985).

Other selected publications:

M. Uhlén *et al.*, "Gene fusion vectors based on the gene for staphylococcal protein A," *Gene* **23**, 369–378 (1983).
B. Löwenadler *et al.*, "Production of specific antibodies against protein A fusion proteins," *EMBO J.* **5**, 2393–2398 (1986).
T. Moks *et al.*, "Large-scale affinity purification of human insulin-like growth factor I from culture medium of *Escherichia coli*," *Nat. Biotechnol.* **5**, 379–382 (1987).
T. Moks *et al.*, "Expression of human insulin-like growth factor I in bacteria: Use of optimized gene fusion vectors to facilitate protein purification," *Biochemistry* **26**, 5239–5244 (1987).
C. Ljungquist *et al.*, "Immobilization and affinity purification of recombinant proteins using histidine peptide fusions," *Eur. J. Biochem.* **186**, 563–569 (1989).

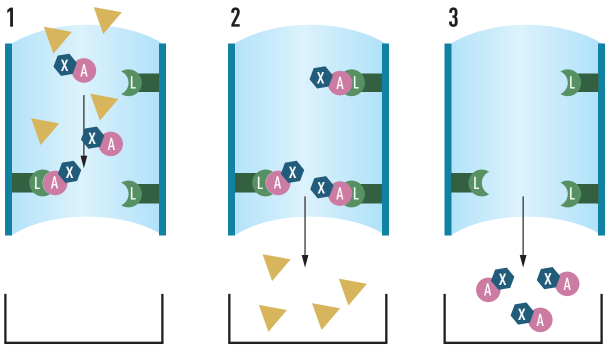


Figure legend: Basic concepts for using gene fusion to purify a gene product (X) by the use of an affinity tag (A) and an affinity ligand (L). Adapted from Uhlén *et al.*, "Gene fusions for purpose of expression: An introduction," *Methods Enzymol.* **185**, 129–143 (1990).

Key facts:

- The affinity tag concept was first described in 1983 (protein-A based)
- The most frequently used affinity tag at present is the histidine-peptide system
- More than 55,000 gene constructs with affinity tags have been generated in the HPA program
- A search for "affinity tag" at Google Scholar yields more than 30,000 publications

Milestone 2

1988 Solid-phase sequencing

Description:

The principle of solid-phase DNA sequencing was first described in 1988 based on binding of biotinylated DNA to streptavidin-coated magnetic beads and selective elution of one strand using alkali. Solid-phase methods are now frequently integrated into many next-generation DNA sequencing methods as well as numerous molecular diagnostics applications.

Key publication:

S. Ståhl *et al.*, "Solid-phase DNA sequencing using the biotin-avidin system," *Nucleic Acids Res.* **16**, 3025–3038 (1988).

Other selected publications:

T. Hultman *et al.*, "Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support," *Nucleic Acids Res.* **17**, 4937–4946 (1989).
M. Uhlén, "Magnetic separation of DNA," *Nature* **340**, 733–734 (1989).
T. Hultman *et al.*, "Solid phase in vitro mutagenesis using plasmid DNA template," *Nucleic Acids Res.* **18**, 5107–5112 (1990).
M. Uhlén *et al.*, "Semi-automated solid-phase DNA sequencing" *Trends Biotechnol.* **10**, 52–55 (1992).
A. Holmberg *et al.*, "The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures," *Electrophoresis* **26**, 501–510 (2005).

Figure:

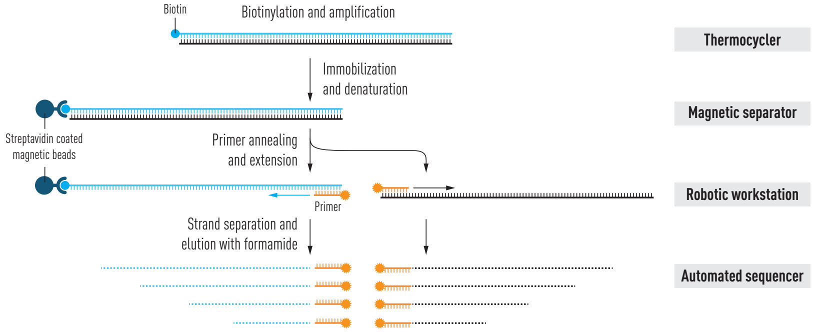

Biotin
Biotinylation and amplification
Thermocycler
Immobilization and denaturation
Magnetic separator
Streptavidin coated magnetic beads
Primer annealing and extension
Robotic workstation
Primer
Strand separation and elution with formamide
Automated sequencer

Figure legend: The principle of solid-phase sequencing using the biotin-streptavidin system. Adapted from M. Uhlén *et al.*, (1992).

Key facts:

- Solid-phase DNA sequencing using the biotin-streptavidin interaction was described in 1988
- Solid-phase methods are now used in the majority of next-generation sequencing methods
- A search for "biotin streptavidin" in Google Scholar yields more than 40,000 publications

Milestone 3

1993 Pyrosequencing

Description:

A novel "sequencing-by-synthesis" principle for DNA sequencing was developed, taking advantage of the detection of pyrophosphate release through a combination of enzymes to generate light. Pyrosequencing, first described in 1993, was further developed in the United States into the first "next-generation" DNA sequencing instruments (Roche 454 sequencers), starting a new era in genomics research. In the HPA effort, next-generation sequencing has been used for genome-wide transcriptomics profiles of the human protein-coding genes.

Key publication:

P. Nyren, B. Pettersson, M. Uhlén "Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay," *Anal. Biochem.* **208**, 171–175 (1993).

Other selected publications:

M. Ronaghi *et al.*, "Real-time DNA sequencing using detection of pyrophosphate release," *Anal. Biochem.* **242**, 84–89 (1996).
M. Ronaghi *et al.*, "A sequencing method based on real-time pyrophosphate," *Science* **281**, 363–365 (1998).
M. Margulies *et al.*, "Genome sequencing in open microfabricated high-density picolitre reactors," *Nature* **437**, 376–380 (2005).

Figure:

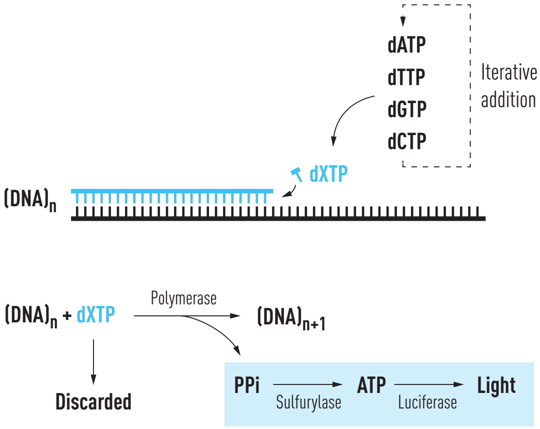


Figure legend: The principle of pyrosequencing involves iterative additions of the four nucleotides and detection of light to monitor incorporation of a nucleotide by a DNA polymerase. Adapted from M. Ronaghi *et al.* (1998).

Key facts:

- First method for pyrosequencing was published in 1993
- Pyrosequencing opened the new era of "next-generation sequencing," leading to a rapid lowering of the cost for DNA sequencing
- Used by the HPA consortium for genome-wide transcriptomics analysis
- A search for "pyrosequencing" in Google Scholar yields more than 100,000 publications
- A search for "next-generation sequencing" in Google Scholar yields more than 700,000 publications

Milestone 4

1996 First concept of antibody-based proteomics

Description:

A system for functional analysis of complementary DNA (cDNA)-encoded proteins was described in which selected portions of cDNAs are expressed as part of a fusion protein used for immunization to elicit antibodies. The concept of antibody-based proteomics was further developed and used in the HPA effort to generate more than 55,000 human recombinant proteins and 55,000 antibodies.

Key publication:

M. Larsson *et al.*, "A general bacterial expression system for functional analysis of cDNA-encoded proteins," *Prot. Expr. Purif.* **7**, 447–457 (1996).

Other selected publications:

M. Larsson *et al.*, "High-throughput protein expression of cDNA products as a tool in functional genomics," *J. Biotechnol.* **80**, 143–157 (2000).
S. Gräslund *et al.*, "A high-stringency proteomics concept aimed for generation of antibodies specific for cDNA-encoded proteins," *Biotechnol. Applied Biochem.* **35**, 75–82 (2002).
S. Gräslund *et al.*, "A novel affinity gene fusion system allowing protein A-based recovery of non-immunoglobulin gene products," *J. Biotechnol.* **99**, 41–50 (2002).
R. Falk *et al.*, "An improved dual-expression concept generating high-quality antibodies for proteomics research," *Biotechnol. Applied Biochem.* **38**, 231–239 (2003).
C. Agaton *et al.*, "Affinity proteomics for systematic profiling of chromosome 21 gene products in human tissues," *Molec. Cell. Proteomics* **2**, 405–414 (2003).

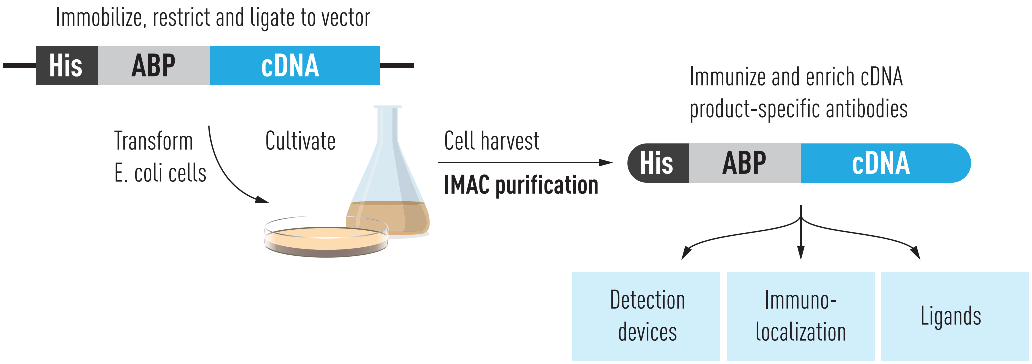


Figure legend: A flow chart representation of the basic high-stringency concept for the generation of antigen-specific antibodies to be used for antibody-based characterization of the human proteome. Adapted from Larsson *et al.* (2000). IMAC, immobilized metal affinity chromatography; His, histidine.

Key facts:

- The 1996 study outlined a new concept for systematic analysis of the human proteome
- The recombinant proteins were used for both immunization and immunocapture
- The principle was later used in the HPA program to generate 55,000 recombinant proteins
- A search for "antibody-based proteomics" in Google Scholar yields more than 2,000 publications

Milestone 5

2000 The Human Genome Project

Description:

The project to sequence the complete human genome was launched at the end of the 1980s, based on new technology for automated DNA sequencing facilitated by fluorescent detection. The project was initially controversial due to the staggering cost needed to complete it. However, several methodological advances were made in the 1990s, including the concepts of expressed sequence tags (ESTs) and whole-genome assembly based on shotgun sequencing, both first described by Craig Venter's laboratory. In addition, several technological advances were made, such as more efficient instruments for fluorescent sequencing and the introduction of automated methods for sample preparation. The solid-phase methods for sequencing (see Milestone 2) and the next-generation sequencing methods (see Milestone 3) were also described during this time, but these were not introduced to the research community until several years after the completion of the human genome sequence. In fall 2000, President Clinton held a press conference in the White House to announce that sequencing of the human genome was complete, achieved by both private and public initiatives. The descriptions of the sequencing and analysis efforts were later published in two landmark papers in 2001. In the initial publications, the number of protein-coding genes in the human genome was estimated to be around 40,000, which turned out to be a gross overestimation, and the number has since been revised down to less than 20,000. This effort has allowed the HPA to generate antibodies to proteins corresponding to nearly all of the genes predicted from the genome sequence, including those not studied previously. Thus, the HPA portal has provided the first available information about many thousands of proteins in the human genome.

Key publication:

J. C. Venter *et al.*, "The sequence of the human genome," *Science* **291**, 1304–1351 (2001).
E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nature* **409**, 860–921 (2001).

Other selected publications:

M. D. Adams *et al.*, "Complementary DNA sequencing: Expressed sequence tags and human genome project," *Science* **252**, 1651–1656 (1991).
R. D. Fleischmann *et al.*, "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd," *Science* **269**, 496–512 (1995).
G. D. Schuler *et al.*, "A gene map of the human genome" *Science* **274**, 540–546 (1996).
F. S. Collins *et al.*, "New goals for the U.S. Human Genome Project: 1998–2003," *Science* **282**, 682–689 (1998).
F. S. Collins *et al.*, "The Human Genome Project: Lessons from large-scale biology," *Science* **300**, 286–290 (2003).

Figure legend: President Clinton, flanked by J. Craig Venter and Francis Collins, in the White House on June 26, 2000 to announce the completion of a "rough draft" of the human genetic code.

Key facts:

- **The completion of the human genome project was announced at a press conference in 2000**
- **The announcement was followed by two landmark publications in 2001**
- **The publication by Venter *et al.* (2001) has been cited by more than 16,000 publications**
- **The publication by Lander *et al.* (2001) has been cited by more than 19,000 publications**
- **The original estimation of the number of protein-coding genes based on the human genome sequence was later recognized to be an overestimation and was revised from 40,000 to around 20,000**

Milestone 6

2000 Chromosome 21 pilot

Description:

A pilot study was initiated in 2000 to investigate all genes encoded by human chromosome 21. The study is the first chromosome-wide exploration in which an affinity proteomics strategy using antibodies raised against recombinant human protein fragments was used for protein profiling. The results, published in 2003, suggested that this strategy could be used to produce a proteome atlas describing distribution and expression of proteins in normal and disease tissues.

Key publication:

C. Agaton *et al.*, "Affinity proteomics for systematic protein profiling of chromosome 21 gene products in human tissues," *Mol. Cell. Proteomics* **2**, 405–414 (2003).

Other selected publications:

S. Gräslund *et al.*, "A high-stringency proteomics concept aimed for generation of antibodies specific for cDNA-encoded proteins," *Biotechnol. Appl. Biochem.* **35**, 75–82 (2002).
C. Agaton *et al.*, "Selective enrichment of monospecific polyclonal antibodies for antibody-based proteomics efforts," *J. Chromatography* **16**, 33–40 (2004).
L. Berglund *et al.*, "A whole-genome bioinformatics approach to selection of antigens for systematic antibody generation," *Proteomics* **8**, 2832–2839 (2008).
M. Uhlén *et al.*, "Antibody-based protein profiling of the human chromosome 21," *Mol. Cell Proteomics* **11**, M111.013458 (2012).

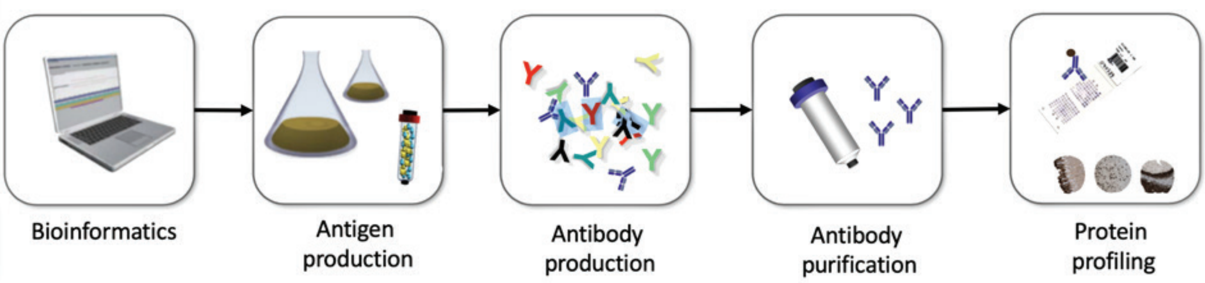


Figure legend: The pipeline to generate antibodies and protein profiles corresponding to a majority of the protein-coding genes at human chromosome 21. Bioinformatics tools were used to design recombinant protein fragments with unique epitopes. The recombinant protein fragments were produced using de novo cloning from RNA pools and expression in *Escherichia coli*. The purified recombinant protein fragments were used to immunize rabbits to generate polyclonal antibodies. The polyclonal antibodies were affinity purified using the recombinant protein fragments as affinity ligands, resulting in target protein–specific antibodies. Protein profiling was performed using tissue microarrays and target-specific polyclonal antibodies, enabling multiscale protein-expression analysis.

Key facts:

- Antibodies to more than half of the proteins encoded by the 225 genes on chromosome 21 were generated
- The antibodies were used for tissue exploration using immunohistochemistry
- The success rate of the whole concept, from in silico design to protein profiling, makes the strategy suitable for genome-wide protein profiling

Milestone 7

2003 Start of the Human Protein Atlas program

Description:

The HPA program started in summer 2003, upon receipt of funding from the Knut and Alice Wallenberg Foundation (KAW). The original funding has since been renewed multiple times by KAW. In the proposal (2002), the founders of the program asked the question: "What is it that makes a kidney a kidney? And what makes a heart a heart? All cells, regardless of whether they are in a kidney or the heart, contain exactly the same genetic material and genes. However, different genes are active in the various cells. This leads the cells to have entirely different functions. Some become nerves, others begin to produce insulin. If researchers are to understand how our bodies work, it is these differences they need to investigate, since the proteins account for all activities in the body. They build muscles and tendons, catalyze chemical reactions, send signals all over and much more."

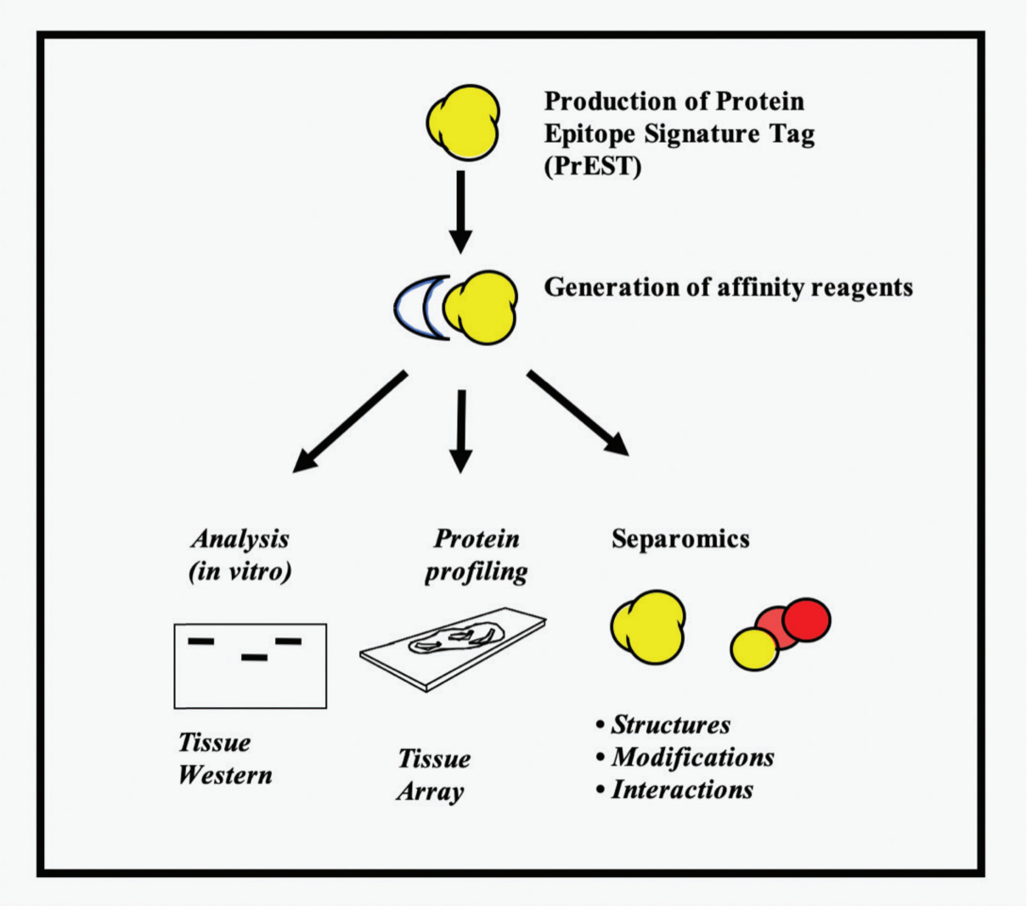


Figure legend: Figure from the original project proposal to KAW in 2002.

Key facts:

- The HPA program has been funded for 20 years by the Knut and Alice Wallenberg Foundation
- More than 500 researchers have been involved in the program throughout its history
- The project has resulted in important collaborations with researchers in Europe, the United States, India, China, and South Korea
- The project has generated 55,000 recombinant proteins, 21,000 validated antibodies, more than 10 million annotated images, and more than 500 "in-house" publications

Milestone 8

2004 Tissue microarrays

Description:

A high-throughput tissue profiling platform was set up to allow comprehensive immunohistochemistry-based analysis of protein expression patterns in normal human tissues, cancer tissue, and cell lines. Altogether, more than 700 individual microarray samples were immunohistochemically stained for each antibody. A web-based annotation system was developed to allow for evaluation and scoring of immunohistochemical staining patterns in tissues, and the manual analysis was performed by pathologists in Sweden and India. Protocols were standardized to enable a global, immunohistochemistry-based protein profiling of unprecedented magnitude.

Key publication:

C. Kampf *et al*., "Antibody-based tissue profiling as a tool for clinical proteomics," *Clin. Proteomics* **1**, 285–299 (2004).

Other selected publications:

A.-C. Andersson *et al*., "Analysis of protein expression in cell microarrays: a tool for antibody-based proteomics," *J. Histochem. Cytochem.* **54**, 1413–1423 (2006).
S. Strömberg *et al*., "A high-throughput strategy for protein profiling in cell microarrays using automated image analysis," *Proteomics* **7**, 2142–2150 (2007).
C. Kampf *et al*., "Production of tissue microarrays, immunohistochemistry staining and digitalization within the Human Protein Atlas," *J. Vis. Exp.* **63**, 3620 (2012).

Figure legend: Each antibody in the HPA program has been used to stain more than 700 tissues using the HPA microarray platform.

Key facts:

- The HPA microarrays include 44 tissue types covering all essential organs and the 20 major forms of cancer
- The HPA cell microarray has also been developed covering 68 human cell lines
- More than 10 million immunohistochemistry images have been generated and manually annotated

Milestone 9

2004 The HPA data management system

Description:

A dedicated laboratory information management system (LIMS) was developed to handle the flow of samples in the HPA workflow, allowing thousands of samples to be processed daily. The HPA LIMS is a web-based, state-of-the-art platform and handles all the steps in the program, including production, analysis, and validation. All data processed by the LIMS are stored, including the noncompressed images. Visualization of the data is made both internally and externally through the HPA portal. The LIMS system is constantly updated to meet the needs of the program.

Key publication:

L. Berglund *et al.*, "A whole-genome bioinformatics approach to selection of antigens for systematic antibody generation," *Proteomics* **8**, 2832–2839 (2008).

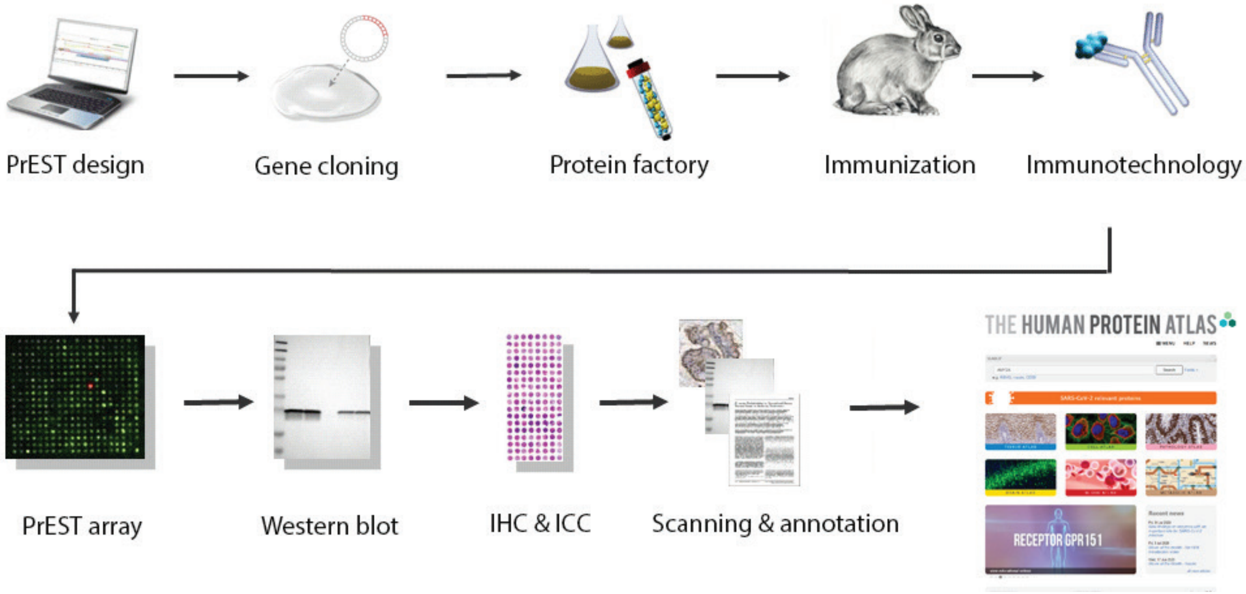


Figure legend: Schematic representation of the HPA workflow. The LIMS enables the thousands of samples processed on any given day to be followed through the workflow.

Key facts:

- The HPA LIMS system has evolved with more than 100 person-years of development invested
- Barcoding is extensively used to track the flow of samples through the different unit operations
- Tight integrations with lab instruments optimize production throughput and minimize risk of errors

Milestone 10

2005 Launch of the Human Protein Atlas portal

Description:

The first version of the HPA (www.proteinatlas.org) was based on a map of expression and localization profiles in 44 normal human tissues and 20 different cancers. The first version of this publicly available database contained approximately 400,000 high-resolution images using 700 antibodies generated in-house. Each image was annotated by a certified pathologist to provide a knowledge base for functional studies and to allow queries about protein profiles in normal and disease tissues.

Key publication:

M. Uhlén *et al.*, "A human protein atlas for normal and cancer tissues based on antibody proteomics," *Mol. Cell. Proteomics* **4**, 1920–1932 (2005).

Other selected publications:

C. Agaton *et al.*, "Selective enrichment of monospecific polyclonal antibodies for antibody-based proteomics efforts," *J. Chromatography* **16**, 33–40 (2004).
P. Nilsson *et al.*, "Towards a human proteome atlas: high-throughput generation of mono-specific antibodies for tissue profiling," *Proteomics* **5**, 4327–4337 (2005).
M. Uhlén *et al.*, "Antibody-based proteomics for human tissue profiling," *Mol. Cell. Proteomics* **4**, 384–393 (2005).
A.-C. Andersson *et al.*, "Analysis of protein expression in cell microarrays: a tool for antibody-based proteomics," *J. Histochem. Cytochem.* **54**, 1413–1423 (2006).
L. Berglund *et al.*, "A genecentric Human Protein Atlas for expression profiles based on antibodies," *Mol. Cell. Proteomics* **7**, 2019–2027 (2008).

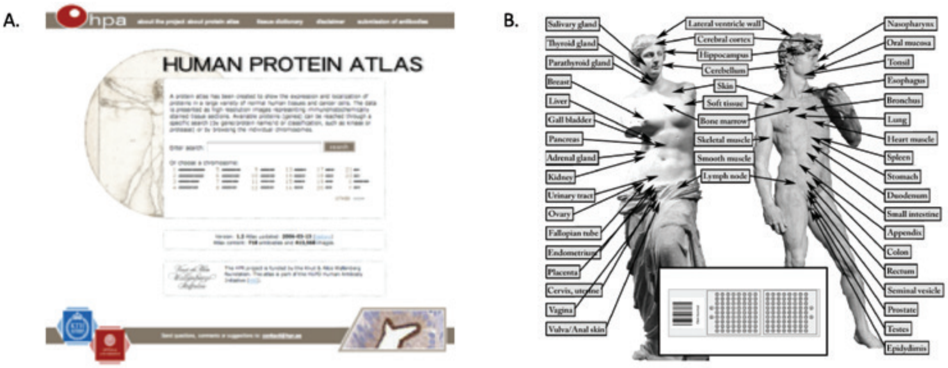


Figure legend: The launch of the HPA portal. (**A**) The home page and (**B**) the surgical specimens analyzed in this first version. In total, 144 different tissue cores representing 44 different tissue types were assembled and stained using antibody-based immunohistochemistry.

Key facts:

- The first version of the HPA was launched in 2005 at the Human Proteome Organization (HUPO) conference in Munich, Germany
- The first version included analysis using 718 antibodies against 650 human proteins
- The knowledge-based portal contained 413,568 images, all annotated by certified pathologists
- A search for "Human Protein Atlas" in Google Scholar now yields more than 10,000 publications

Milestone 11

2006 Creation of an antibody resource

Description:

All antibodies generated within the HPA effort have been made available to the research community through the Atlas Antibodies portal (www.atlasantibodies.com). All antibodies have been validated by tissue profiling in more than 700 tissue samples and annotated by a certified pathologist. Several international efforts have been initiated to provide antibodies to the research community, and HPA has participated in many of these efforts, including the European Union framework program ProteomeBinders and the U.S. National Institutes of Health Protein Capture Reagents Program.

Key publication:

M. J. Taussig *et al.*, "ProteomeBinder: Planning a European resource of affinity reagents for analysis of the human proteome," *Nat. Methods* **4**, 13–17 (2006).

Other selected publications:

O. Stoevesandt *et al.*, "European and international collaboration in affinity proteomics," *N. Biotechnol.* **29**, 511–514 (2012).
D. E. Gloriam *et al.*, "A community standard format for the representation of protein affinity reagents," *Mol. Cell. Proteomics* **9**, 1–10 (2010).

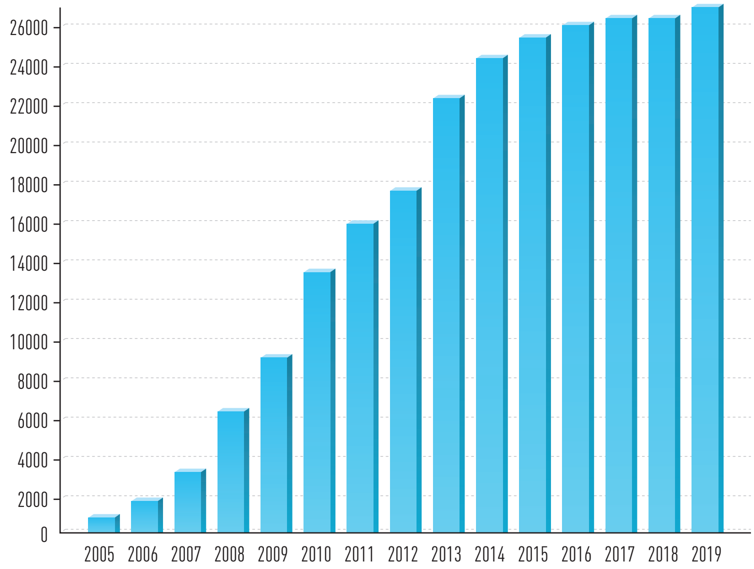


Figure legend: The accumulated number of unique antibodies used to perform the protein profiling published in the HPA each year, starting in 2005.

Key facts:

- **More than 55,000 antibodies have been generated in-house, and approximately 21,000 have passed the stringent annotation of the HPA program for use in immunohistochemistry**
- **More than 10 million tissue and cell images have been generated using this antibody resource**
- **The validation data for each antibody is shown on the "antibodies and validation" page of each gene**
- **All antibodies used in the HPA program are available to the scientific community through the Atlas Antibodies portal (www.atlasantibodies.com)**

Milestone 12

2007 Protein arrays

Description:

As part of the HPA program, several protein microarray formats were developed. All antibodies generated in the program have been analyzed with dedicated antigen arrays in which the specificity and selectivity of each antibody can be evaluated by comparing it to unrelated antigens. In addition, comprehensive protein arrays have been constructed containing more than 42,000 human recombinant protein fragments. These arrays have been used extensively for profiling autoantibody repertoires and antibody validation.

Key publication:

J. M. Schwenk *et al.*, "Determination of binding specificities in highly multiplexed bead-based assays for antibody proteomics," *Mol. Cell. Proteomics.* **6**, 125–132 (2007).

Other selected publications:

P. Nilsson *et al.*, "Towards a human proteome atlas: High-throughput generation of mono-specific antibodies for tissue profiling," *Proteomics* **5**, 4327–4337 (2005).
R. Sjöberg *et al.*, "Validation of affinity reagents using antigen microarrays," *N. Biotechnol.* **29**, 555–563 (2012).
R. Sjöberg *et al.*, "Exploration of high-density protein microarrays for antibody validation and autoimmunity profiling," *N. Biotechnol.* **33**, 582–592 (2016).
B. Ayoglu *et al.*, "Anoctamin 2 identified as an autoimmune target in multiple sclerosis," *Proc. Nat. Acad. Sci. U.S.A.* **113**, 2188–2193 (2016).
M. Neiman *et al.*, "Individual and stable autoantibody repertoires in healthy individuals," *Autoimmunity* **52**, 1–11 (2019).

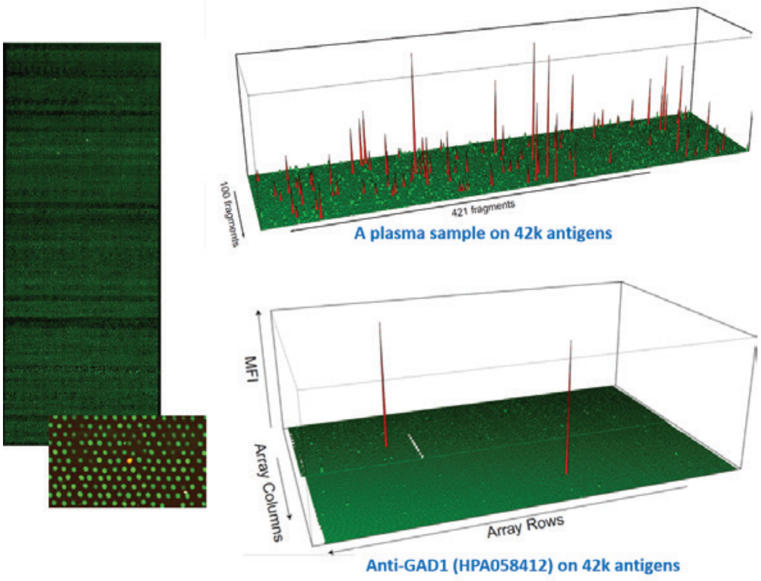


Figure legend: The HPA 42,000 (42k) protein fragment array. The autoantibody repertoire in a plasma sample is seen on the upper plot; the lower displays the results of an HPA antibody validated on the 42k array.

Key facts:

- Over 50,000 antibodies have been validated on antigen microarrays, and 76,000 tests have been performed in total
- Arrays comprising 42,000 antigens derived from 19,000 proteins have been assembled
- Over 90,000 samples have been analyzed within autoimmunity profiling-contexts since 2012
- The system has been used for population screens to detect specific antibodies to the novel coronavirus, SARS-CoV-2

Milestone 13

2008 Biomarkers for body fluids

Description:

Proteins circulating in blood and those found in other body fluids provide important information about health and disease states on both a systemic and organ-specific level. Antibodies and protein fragments generated in the HPA have been used to develop and apply multiplexed assays for discovering disease-related proteins. Together with technical and biological validation schemes, the studies conducted across different disease areas highlight the value of studying proteins as biomarkers.

Key publication:

J. M. Schwenk *et al.*, "Antibody suspension bead arrays within serum proteomics," *J. Proteome Res.* **7**, 3168–3179 (2008).

Other selected publications:

J. M. Schwenk et al., "Toward next generation plasma profiling via heat-induced epitope retrieval and array-based assays," *Mol. Cell. Proteomics* **9**, 2497–2507 (2010).
J. Bachmann *et al.*, "Affinity proteomics reveals elevated muscle proteins in plasma of children with cerebral malaria," *PLOS Pathog.* **10**, e1004038 (2014).
B. Ayoglu *et al.*, "Affinity proteomics within rare diseases: a BIO-NMD study for blood biomarkers of muscular dystrophies," *EMBO Mol. Med.* **6**, 918–936 (2014).
S. Byström *et al.*, "Affinity proteomic profiling of plasma, cerebrospinal fluid, and brain tissue within multiple sclerosis," *J. Proteome Res.* **13**, 4607–4619 (2014).
R. S. Häussler *et al.*, "Systematic development of sandwich immunoassays for the plasma secretome," *Proteomics* **19**, e1900008 (2019).

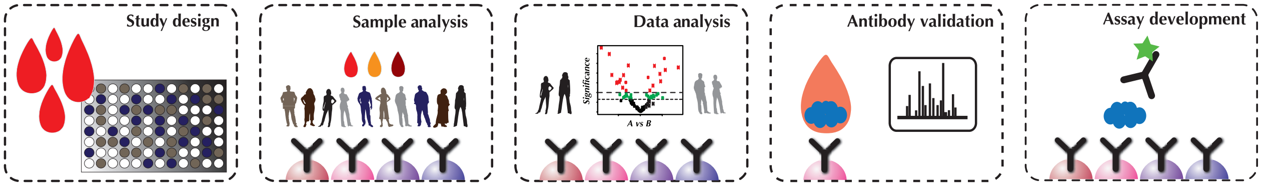


Figure legend: Translational pipeline to discover and validate proteins biomarkers in body fluids.

Key facts:

- Multiplexed immunoassays for protein profiling have been performed across multiple diseases
- Systematic assays for plasma, serum, and cerebrospinal fluid have been performed, allowing analysis of thousands of protein targets
- More than 80 peer-reviewed, body-fluid biomarker studies have been published by HPA researchers

Milestone 14

2008 Epitope mapping of antibodies

Description:

Several new technologies for analyzing the binding parameters of antibodies have been developed for antibody mapping. First, libraries using bacterial surface display have been employed to screen epitopes for antibodies of therapeutic interest. Second, microfabricated arrays with many millions of synthetic peptides have been developed and used for fine mapping of antibody epitopes. Third, suspension bead arrays have been used to allow for a flexible system for epitope mapping of both monoclonal and polyclonal antibodies.

Key publication:

J. Rockberg *et al.*, "Epitope mapping of antibodies using bacterial surface display," *Nat. Methods* **5**, 1039–1045 (2008).

Other selected publications:

J. Rockberg *et al.*, "Prediction of antibody response using recombinant human protein fragments as antigen," *Protein Sci.* **18**, 2346–2355 (2009).
J. Rockberg *et al.*, "Discovery of epitopes for targeting the human epidermal growth factor receptor 2 (HER2) with antibodies," *Mol. Oncol.* **3**, 238–247 (2009).
B. Hjelm *et al.*, "Exploring epitopes of antibodies toward the human tryptophanyl-tRNA synthetase," *N. Biotechnol.* **27**, 129–137 (2010).
Buus *et al.*, "High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays," *Mol. Cell. Proteomics* **11**, 1790–1800 (2012).
B. Forsström *et al.*, "Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays," *Mol. Cell. Proteomics* **13**, 1585–1597 (2014).

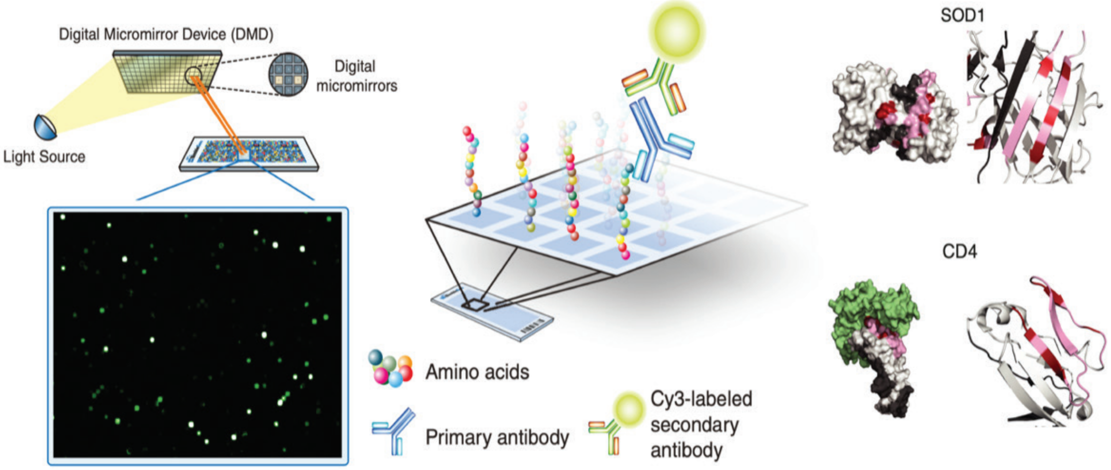


Figure legend: Peptide arrays have been developed covering all human proteins for the analysis of antibody specificity. The technology was based on parallel in situ photolithic synthesis of millions of overlapping peptides. Here, epitopes are shown for antibodies binding to SOD1 and CD4 as determined by these microarrays. Adapted from Forsström *et al.* (2014).

Key facts:

- Binding epitopes were dissected using suspension bead arrays
- Microarrays containing millions of peptides were designed and used for binding analysis
- Recombinant libraries expressed with bacterial surface display were also used for epitope mapping

Milestone 15

2008 Antibodypedia antibody portal

Description:

The portal Antibodypedia (www.antibodypedia.com) was launched in 2008 to allow sharing of information regarding validation of antibodies. The database provides a resource of publicly available antibodies to human proteins with accompanying experimental evidence supporting an individual validation score for each antibody in an application-specific manner. The resource now contains information for antibodies corresponding to over 19,000 human protein targets.

Key publication:

Björling *et al.*, "Antibodypedia, a portal for sharing antibody and antigen validation data," *Mol. Cell. Proteomics* **7**, 2019–2027 (2008).

Other selected publications:

K. Cottingham *et al.*, "Antibodypedia seeks to answer the question: 'How good is that antibody?'" *J. Proteome Res.* **7**, 4213–4213 (2008).
K. Jonasson *et al.*, "The 6th HUPO Antibody Initiative (HAI) workshop: Sharing data about affinity reagents and other recent developments. September 2009, Toronto, Canada," *Proteomics* **10**, 2066–2068 (2010).
T. Alm *et al.*, "A chromosome-centric analysis of antibodies directed toward the human proteome using Antibodypedia," *J. Proteome Res.* **13**, 1669–1676 (2014).
T. Alm *et al.*, "Introducing the Affinity Binder Knockdown Initiative: A public-private partnership for validation of affinity reagents," *EuPA Open Proteom.* **10**, 56–58 (2016).

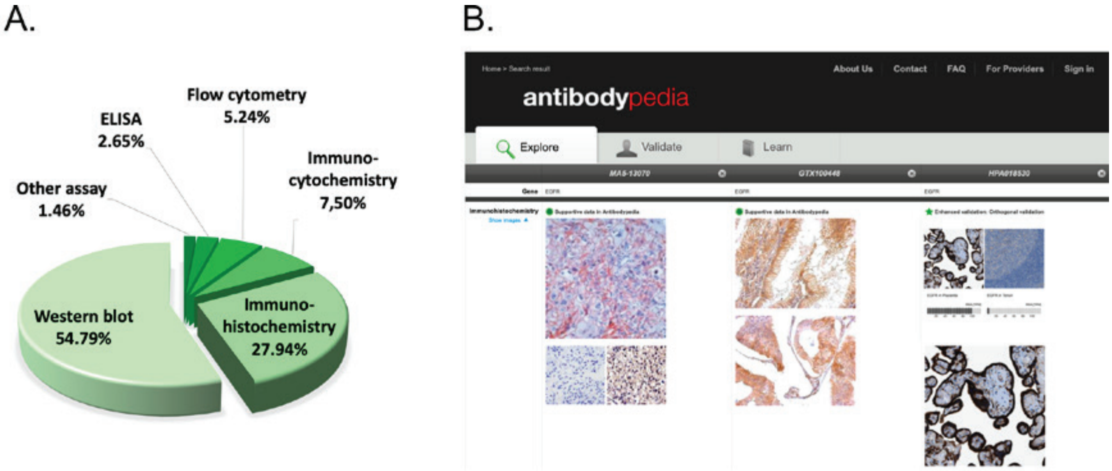


Figure legend: Antibodypedia constitutes a tool for evidence-based selection of antibodies for research. The portal enables users to (**A**) search for antibodies targeting specific proteins for selected applications and (**B**) compare side-by-side experimental evidence for various antibodies to the same target [example showing three anti-epidermal growth factor receptor (EGFR) antibodies].

Key facts:

- Antibodypedia contains information about more than 4 million antibodies
- The database contains more than 2 million validation experiments
- More than half (55%) of the antibodies produced are validated for Western blot
- Other common applications are immunohistochemistry, immunocytochemistry, and flow cytometry

Milestone 16

2009 Biomarker discovery in pathology

Description:

The resource of antibodies and protein profiling data created as part of the HPA program has been used for biomarker discovery programs using pathology-based immunohistochemistry. More than 140 scientific manuscripts have been published, coauthored by members of the HPA effort. This includes biomarkers for both cancer diagnostics and biomarkers with prognostic value to predict clinical outcome in cancer patients.

Key publication:

A. Jögi *et al.*, "Nuclear expression of the RNA-binding protein RBM3 is associated with an improved clinical outcome in breast cancer," *Mod. Pathol.* **22**, 1564–1574 (2009).

Other selected publications:

L. Jonsson *et al.*, "Low RBM3 protein expression correlates with tumour progression and poor prognosis in malignant melanoma: An analysis of 215 cases from the Malmö Diet and Cancer Study," *J. Transl. Med.* **21**, 114 (2011).
K. Magnusson *et al.*, "SATB2 in combination with cytokeratin 20 identifies over 95% of all colorectal carcinomas," *Am. J. Surg. Pathol.* **35**, 937–948 (2011).
G. Gremel *et al.*, "A systematic analysis of commonly used antibodies in cancer diagnostics," *Histopathology* **64**, 293–305 (2014).
A. Dragomir *et al.*, "The role of SATB2 as a diagnostic marker for tumors of colorectal origin: Results of a pathology-based clinical prospective study," *Am. J. Clin. Pathol.* **141**, 630–638 (2014).

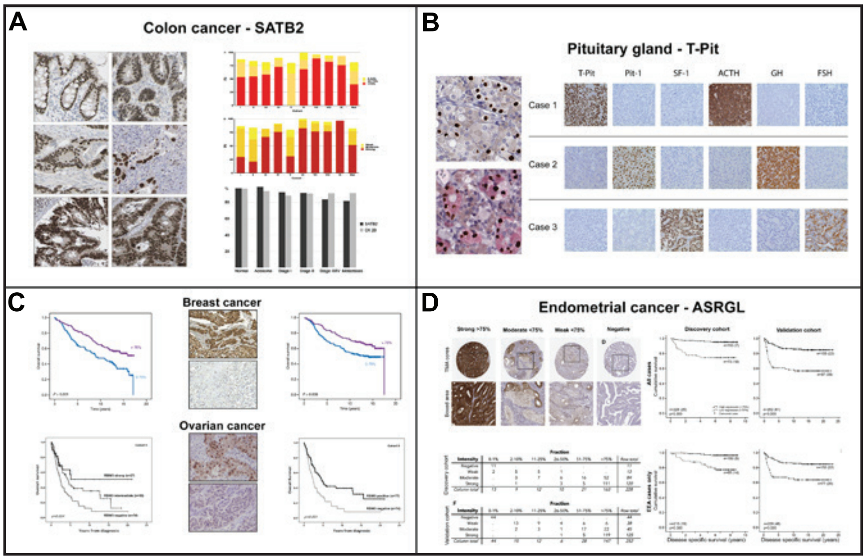


Figure legend: Examples of tissue-biomarker discovery efforts, including (**A**) SATB2 as a diagnostic marker for colorectal cancer, (**B**) T-Pit transcription factor for diagnostics of pituitary neuroendocrine tumors, (**C**) RBM3 as a prognostic cancer biomarker for breast and ovarian cancer, and (**D**) ASRGL1 as a prognostic marker for endometrial cancer.

Key facts:

- Protein expression data from the HPA has been used for cancer biomarker discovery
- Prognostic markers based on expression data were identified for several cancers
- More than 140 peer-reviewed publications in pathology have been published by HPA researchers

Milestone 17

2010 Knowledge-based portal

Description:

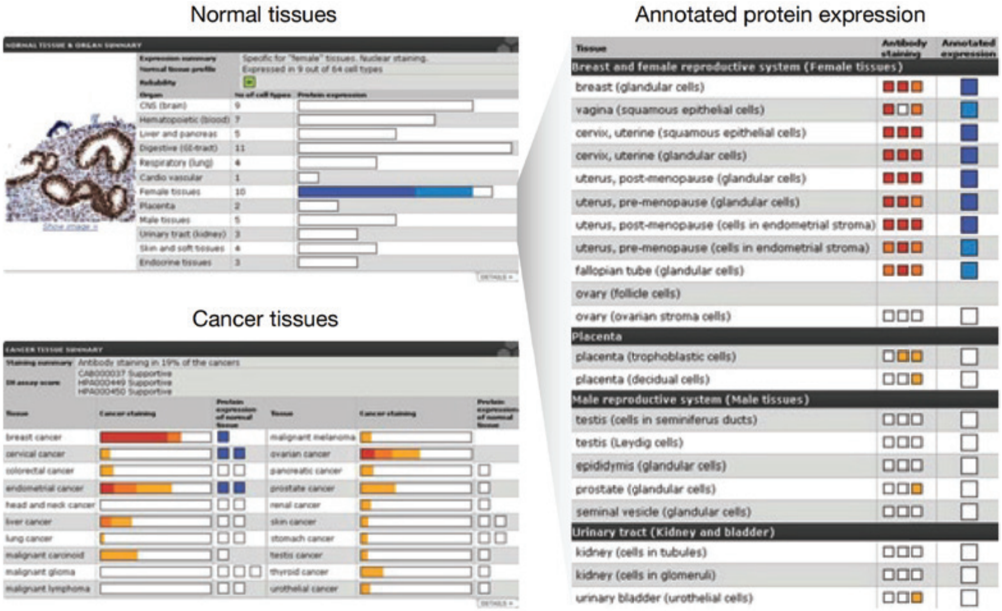
A milestone for the HPA effort was achieved with the inclusion of expression data for approximately 50% of human protein-coding genes. This was achieved in 2010 and an updated portal was launched. An important new feature was a cancer view. In addition, a new concept for subcellular localization of proteins using confocal microscopy was described.

Key publication:

M. Uhlén *et al.*, "Towards a knowledge-based Human Protein Atlas," *Nat. Biotechnol.* **28**, 1248–1250 (2010).

Other selected publications:

S. Mathivanan *et al.*, "Human Proteinpedia enables sharing of human protein data," *Nat. Biotechnol.* **26**, 164–167 (2008).
J. Bourbeillon *et al.*, "Minimum information about a protein affinity reagent (MIAPAR)," *Nat. Biotechnol.* **28**, 650–653 (2010).
M. Gry *et al.*, "Tissue-specific protein expression in human cells, tissues and organs," *J. Proteomics Bioinform.* **3**, 294–301 (2010).
P. Legrain *et al.*, "The human proteome project: Current state and future direction," *Mol. Cell. Proteomics* **10**, M111.009993 (2011).
L. Fagerberg *et al.*, "Large-scale protein profiling in human cell lines using antibody-based proteomics," *J. Proteome Res.* **10**, 4066–4075 (2011).
K. Colwill *et al.*, "A roadmap to generate renewable protein binders to the human proteome," *Nat. Methods* **8**, 551–558 (2011).



Normal tissues

NORMAL TISSUE & ORGAN SUMMARY

Expression summary	Specific for "female" tissues. Nuclear staining.
Normal tissue profile	Expressed in 9 out of 64 cell types
Reliability	🟩

Organ	No of cell types	Protein expression
CNS (brain)	9	
Hematopoietic (blood)	7	
Liver and pancreas	5	
Digestive (GI-tract)	11	
Respiratory (lung)	4	
Cardio vascular	1	
Female tissues	10	
Placenta	2	
Male tissues	5	
Urinary tract (kidney)	3	
Skin and soft tissues	4	
Endocrine tissues	3	

Show image »

DETAILS »

Cancer tissues

CANCER TISSUE SUMMARY

Staining summary	Antibody staining in 19% of the cancers
IH assay score	CAB000037 Supportive HPA000449 Supportive HPA000450 Supportive

Tissue	Cancer staining	Protein expression of normal tissue	Tissue	Cancer staining	Protein expression of normal tissue
breast cancer		🟦	malignant melanoma		
cervical cancer		🟦 🟦	ovarian cancer		
colorectal cancer		□ □	pancreatic cancer		□
endometrial cancer		🟦 🟦	prostate cancer		□
head and neck cancer		□ □	renal cancer		□
liver cancer		□ □	skin cancer		□ □
lung cancer		□ □	stomach cancer		□ □
malignant carcinoid		□	testis cancer		□
malignant glioma		□ □ □	thyroid cancer		□
malignant lymphoma		□ □	urothelial cancer		□

DETAILS »

Annotated protein expression

Tissue	Antibody staining	Annotated expression
Breast and female reproductive system (Female tissues)		
breast (glandular cells)	🟥🟥🟧	🟦
vagina (squamous epithelial cells)	🟥□🟧	🟦
cervix, uterine (squamous epithelial cells)	🟥🟥🟥	🟦
cervix, uterine (glandular cells)	🟥🟥🟥	🟦
uterus, post-menopause (glandular cells)	🟥🟥🟥	🟦
uterus, pre-menopause (glandular cells)	🟥🟥🟧	🟦
uterus, post-menopause (cells in endometrial stroma)	🟥🟥🟥	🟦
uterus, pre-menopause (cells in endometrial stroma)	🟧🟥🟧	🟦
fallopian tube (glandular cells)	🟧🟥🟧	🟦
ovary (follicle cells)		
ovary (ovarian stroma cells)	□□□	□
Placenta		
placenta (trophoblastic cells)	□🟨🟨	□
placenta (decidual cells)	□□🟨	□
Male reproductive system (Male tissues)		
testis (cells in seminiferus ducts)	□□□	□
testis (Leydig cells)	□□□	□
epididymis (glandular cells)	□□□	□
prostate (glandular cells)	□□🟨	□
seminal vesicle (glandular cells)	□□□	□
Urinary tract (Kidney and bladder)		
kidney (cells in tubules)	□□□	□
kidney (cells in glomeruli)	□□□	□
urinary bladder (urothelial cells)	□□🟨	□

Figure legend: The HPA summary view in 2010 for the human estrogen receptor 1 (ESR1), displaying expression data for the target protein in normal and cancer tissues. Adapted from M. Uhlén *et al.* (2010).

Key facts:

- The HPA contains information about 10,118 protein-coding genes in 44 tissues and organs
- This corresponds to >50% of the 19,559 human entries as defined by the Ensembl genome browser
- A new cancer view has been designed, with tumor tissues from 216 patients representing 20 cancers
- A new feature is a section for subcellular localization of proteins
- A new concept was introduced for validation of antibodies based on paired antibodies

Milestone 18

2011 Therapeutic antibodies and Affibodies

Description:

Potential biopharmaceuticals were developed by the HPA consortium to generate molecules based on different formats, including immunoglobulin G frameworks and the small Affibody scaffold suitable for protein engineering. These Affibody molecules have now entered various human clinical trials, including indications such as cancer, psoriasis, autoimmune diseases, and inflammation. Therapeutic antibodies have been developed in collaboration with the South Korean company AbClon, including therapeutic antibodies for cancer treatment and antibodies with a neutralizing effect on the novel coronavirus, SARS-CoV-2.

Key publication:

B.-K. Ko *et al.*, "Combination of novel HER2-targeting antibody 1E11 with trastuzumab shows synergistic antitumor activity in HER2-positive gastric cancer," *Mol. Oncol.* **9**, 398–408 (2015).

Other selected publications:

A. Orlova *et al.*, "Tumor imaging using a picomolar affinity HER2 binding affibody molecule," *Cancer Res.* **66**, 4339–4348 (2006).
J. Löfbom *et al.*, "Affibody molecules: Engineered proteins for therapeutic, diagnostic and biotechnological applications," *FEBS Lett.* **584**, 2670–2680 (2010).
A.-L. Volk *et al.*, "Stratification of responders towards eculizumab using a structural epitope mapping strategy," *Sci. Rep.* **6**, 31365 (2016).
S. Ståhl *et al.*, "Affibody molecules in biotechnological and medical applications," *Trends Biotechnol.* **35**, 691–712 (2017).
W. Hoyer *et al.*, "Stabilization of a β-hairpin in monomeric Alzheimer's amyloid-β peptide inhibits amyloid formation," *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5099–5104 (2008).

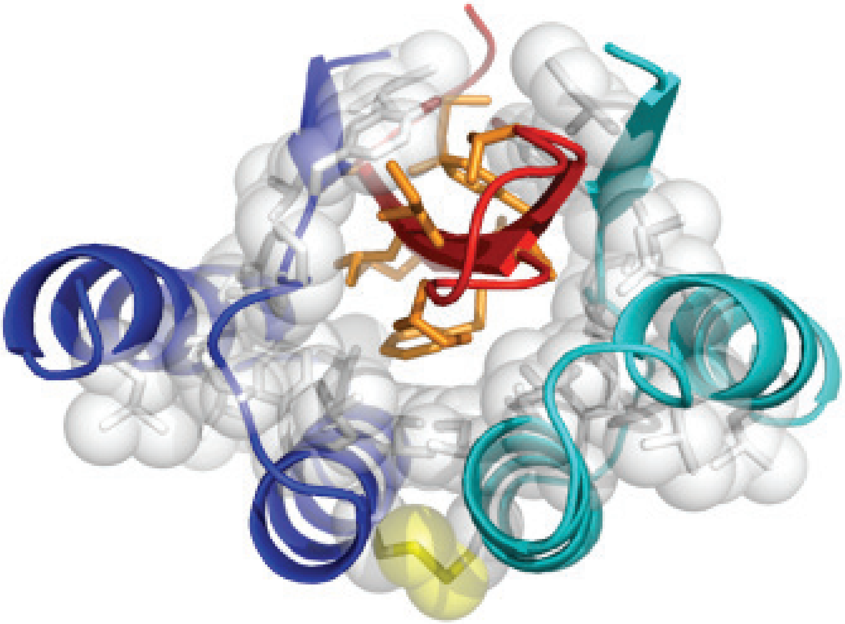


Figure legend: Affibody binding to the beta-amyloid protein involved in Alzheimer's disease. Adapted from W. Hoyer *et al.* (2008).

Key facts:

- The antibody molecule AC101 entered human clinical trials in China in 2019
- More than 200 patients have now received Affibody molecules in clinical trials
- A search for "affibody" in Google Scholar yields more than 8,000 publications

Milestone 19

2012 Targeted proteomics

Description:

The recombinant protein fragments generated within the HPA program were used to develop a new concept for targeted proteomics based on stable isotope-labeled standards to allow for multiplex quantification of proteins in tissues and blood. This work was done in collaboration with Matthias Mann and coworkers at the Max Planck Institute of Biochemistry in Martinsried, Germany. This concept was later used by the HPA group to analyze the correlation of RNA and proteins in cells and tissues; and recently, multiplex assays for blood analysis have been developed to explore protein profiles in healthy and diseased individuals.

Key publication:

M. Zeller *et al.*, "A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines," *Mol. Cell. Proteomics* **11**, O111.009613 (2012).

Other selected publications:

A. I. Lamond *et al.*, "Advancing cell biology through proteomics in space and time (PROSPECTS)," *Mol. Cell. Proteomics* **11**, O112.017731 (2012).
T. Geiger *et al.*, "Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse," *Mol. Cell. Proteomics* **12**, 1709–1722 (2013).
F. Edfors *et al.*, "Immunoproteomics using polyclonal antibodies and stable isotope-labeled affinity-purified recombinant proteins," *Mol. Cell. Proteomics* **13**, 1611–1624 (2014).
F. Edfors *et al.*, "Screening a resource of recombinant protein fragments for targeted proteomics," *J. Proteome Res.* **18**, 2706–2718 (2019).
A. Hober *et al.*, "Absolute quantification of apolipoproteins following treatment with omega-3 carboxylic acids and fenofibrate using a high precision stable isotope-labeled recombinant protein fragments based SRM assay," *Mol. Cell. Proteomics* **18**, 2433–2446 (2019).

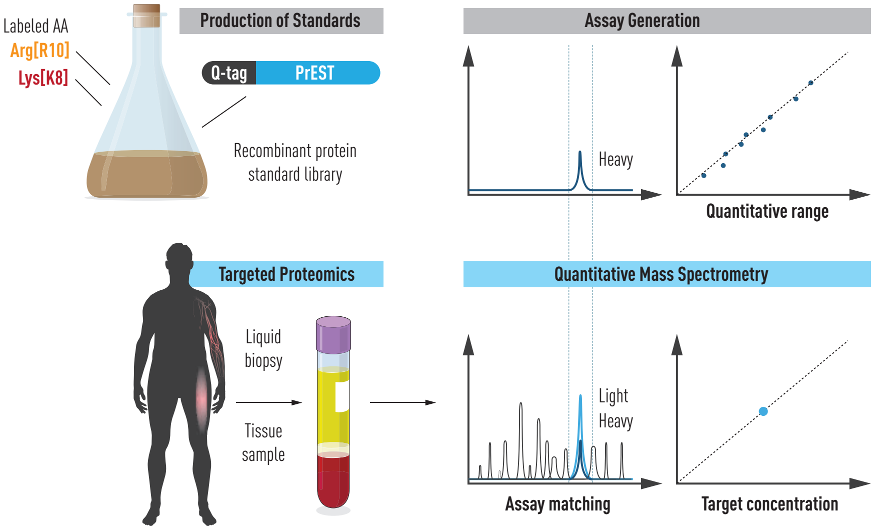


Figure legend: The principle of targeted proteomics using stable isotope-labeled recombinant protein fragments (QPRESTs).

Key facts:

- A resource of 26,840 individually purified recombinant protein fragments corresponding to more than 16,000 human protein-coding genes has been analyzed using targeted proteomics
- More than 500 stable isotope-labeled recombinant protein fragments have been used for plasma analysis
- The absolute concentration of endogenous target proteins can be determined using this method

Milestone 20

2014 Integration of RNA and protein profiles

Description:

A new version of the HPA was launched with integration of protein expression and transcriptomics data. 27 tissues and 33 cell lines were analyzed, and a new classification was introduced based on the transcriptomics profiles across all major tissues. A new version of the HPA was launched with information from 21,900 antibodies corresponding to approximately 16,600 human genes (80% of the human protein–coding genes).

Key publication:

L. Fagerberg *et al.*, "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics," *Mol. Cell. Proteomics* **13**, 397–406 (2014).

Other selected publications:

D. Djureinovic *et al.*, "The human testis-specific proteome defined by transcriptomics and antibody-based profiling" *Mol. Hum. Reprod.* **20**, 476–488 (2014).
C. Kampf *et al.*, "The human liver-specific proteome defined by transcriptomics and antibody-based profiling," *FASEB J.* **28**, 2901–2914 (2014).
G. Gremel *et al.*, "The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling," *J. Gastroenterol.* **50**, 46–57 (2014).
C. Lindskog *et al.*, "The lung-specific proteome defined by integration of transcriptomics and antibody-based profiling," *FASEB J.* **28**, 5184–5196 (2014).
C. Kampf *et al.*, "Defining the human gallbladder proteome by transcriptomics and affinity proteomics," *Proteomics* **14**, 2498–2507 (2014).

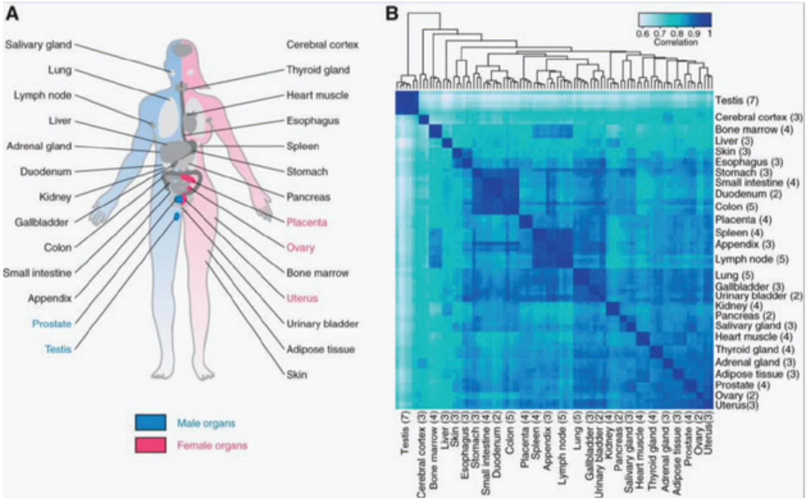


Figure legend: The human tissues and organs analyzed using transcriptomics and incorporated into the HPA 2014 update. Adapted from Fagerberg *et al.* (2014).

Key facts:

- Transcriptomics analysis across 27 organs and tissues were included in a new version of the portal
- A new classification principle was introduced based on whole-body, tissue-specific protein expression
- Data indicated that tissue-enriched genes constituted 12% of all genes (n = 2,473)
- 8% of all genes were not detected in any of the tissues analyzed

Milestone 21

2015 The Tissue Atlas

Description:

The Tissue Atlas was launched in 2015. This new atlas contained information regarding the expression profiles in human tissues and organs on both the messenger RNA (mRNA) and protein level. 76 different cell types were analyzed, corresponding to 44 normal human tissue types, and the data was presented using pathology-based annotation of protein expression levels. A genome-wide analysis of all protein-coding genes was presented with regards to whole-body expression profiles.

Key publication:

M. Uhlén *et al.*, "Tissue-based map of the human proteome," *Science* **347**, 1260419 (2015).

Other selected publications:

L. Fagerberg *et al.*, "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics," *Mol. Cell. Proteomics* **13**, 397–406 (2014).
D. Djureinovic *et al.*, "The human testis-specific proteome defined by transcriptomics and antibody-based profiling," *Mol. Hum. Reprod.* **20**, 476–488 (2014).
C. Kampf *et al.*, "The human liver-specific proteome defined by transcriptomics and antibody-based profiling," *FASEB J.* **28**, 2901–2914 (2014).

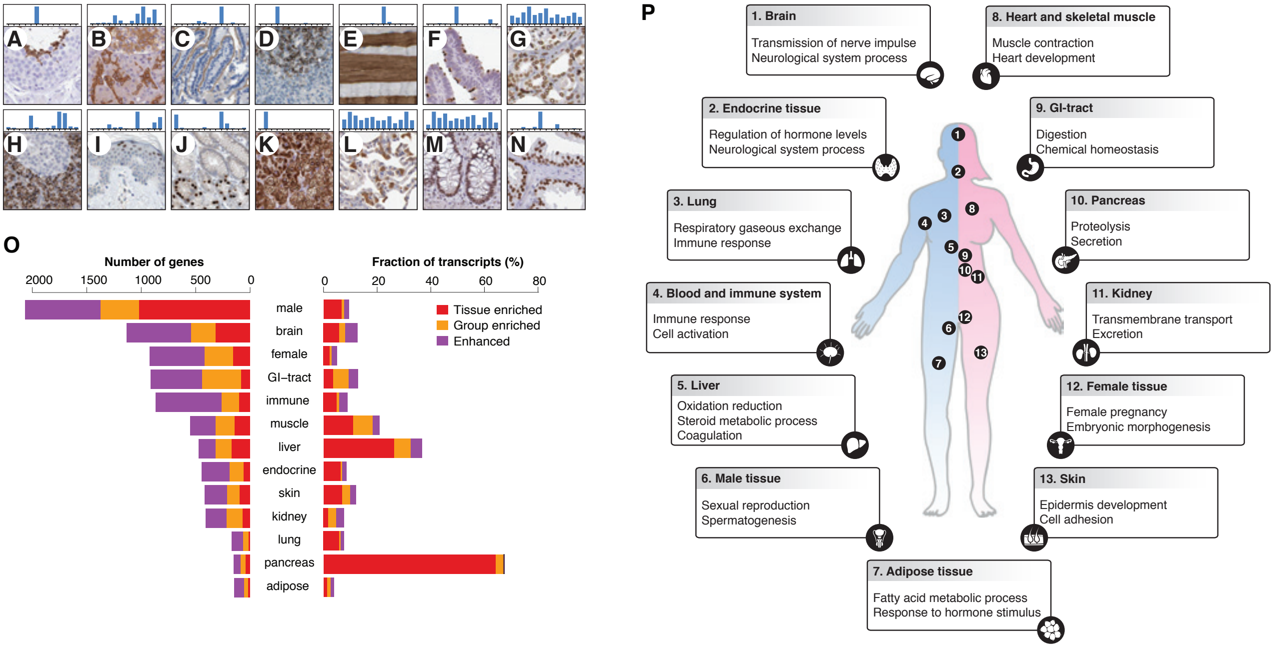


Figure legend: Global expression profiles of human genes on both the mRNA and protein level. Adaped from M. Uhlén *et al.* (2015).

Key facts:

- A genome-wide classification of all protein-coding genes was introduced with regard to tissue profiles
- Various subproteomes were analyzed, such as the druggable proteome and the housekeeping proteome
- The analysis showed that only 3% of the genes are specific for a single tissue
- The Tissue Atlas paper published in *Science* (see Key Publication, above) has been cited many thousands of times (Google Scholar)

Milestone 22

2016 Correlation of RNA and protein levels

Description:

An important task for molecular biology is to establish whether transcript levels of a given gene can be used as proxies for the corresponding protein levels. This has huge implications for the use of transcriptomics and single-cell analysis to study human biology. Contrary to many earlier studies, the analyses using targeted proteomics and next-generation transcriptomics showed that transcript and protein levels in general correlate if a gene-specific RNA-to-protein (RTP) conversion factor is introduced. The results suggest that transcriptomics can be used to predict the relative levels of the corresponding protein, thus forming an attractive link between the field of genomics and proteomics.

Key publication:

F. Edfors *et al.*, "Gene-specific correlation of RNA and protein levels in human cells and tissues," *Mol. Syst. Biol.* **12**, 883 (2016).

Other selected publications:

M. Gry *et al.*, "Correlations between RNA and protein expression profiles in 23 human cell lines," *BMC Genomics* **10**, 365 (2009).
T. Geiger *et al.*, "Initial quantitative proteomics map of 28 mouse tissues using the SILAC mouse," *Mol. Cell. Proteomics* **12**, 1709-1722 (2013).
E. Lundberg *et al.*, "Defining the transcriptome and proteome in three functionally different human cell lines," *Mol. Syst. Biol.* **6**, 450 (2010).

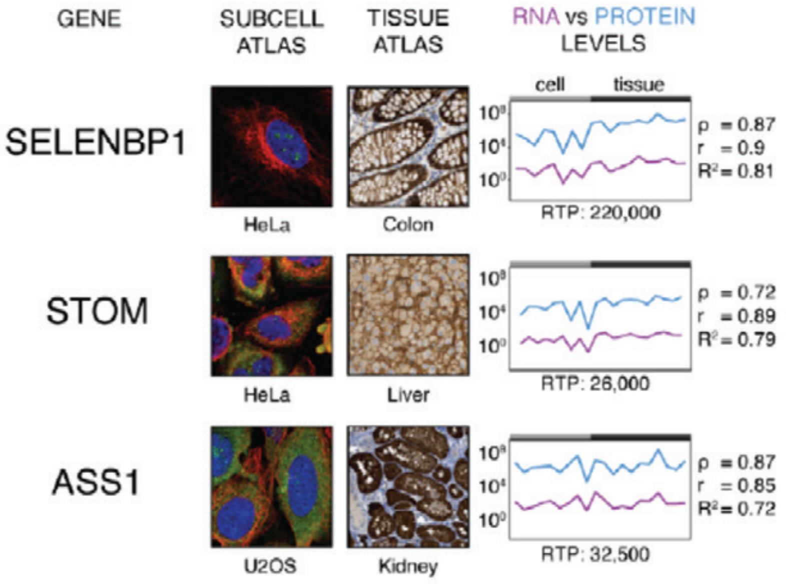


Figure legend: The correlation between levels of RNA and protein across tissues and cells. Adapted from Edfors *et al.* (2016).

Key facts:

- High genome-wide Pearson correlation between protein and RNA levels in cell lines and tissues (*r* = 0.9)
- Data suggests that mRNA levels in general can be used as a proxy for protein levels
- Transcriptomics forms an attractive link between the fields of genomics and proteomics

Milestone 23

2016 Human secretome resource

Description:

A new program was launched to generate a resource comprising most of the secreted proteins in humans. The genes coding for the proteins predicted to be secreted were constructed with synthetic biology and used for mammalian bioproduction using Chinese hamster ovary (CHO) cells. The program was initiated by a collaboration involving the Wallenberg Foundation, the Novo Nordisk Foundation, and the pharmaceutical company AstraZeneca.

Key publication:

H. Tegel *et al.*, "High throughput generation of a resource of the human secretome in mammalian cells," *New Biotechnol.* **58**, 45–54 (2020).

Other selected publications:

M. Uhlén *et al.*, "The human secretome," *Sci. Signal.* **12**, eaaz0274 (2019).
K. Jennbacken *et al.*, "Phenotypic screen with the human secretome identifies FGF16 to induce proliferation of iPSC-derived cardiac progenitor cells," *Int. J. Mol. Sci.* **20**, 6037 (2019).

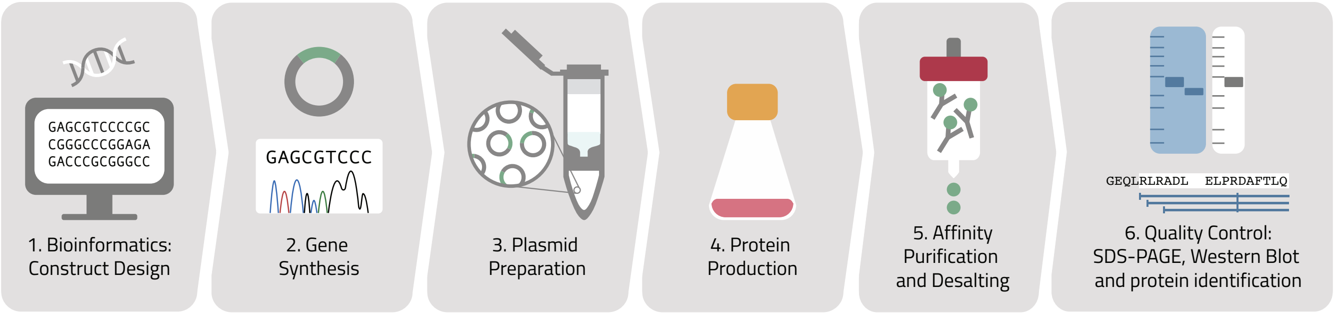

GAGCGTCCCCGC
CGGGCCCGGAGA
GACCCGCGGGCC
1. Bioinformatics: Construct Design
GAGCGTCCC
2. Gene Synthesis
3. Plasmid Preparation
4. Protein Production
5. Affinity Purification and Desalting
GEQLRLRADL ELPRDAFTLQ
6. Quality Control: SDS-PAGE, Western Blot and protein identification

Figure legend: Outline of the CHO mammalian cell factory for generation of a resource of the human secretome.

Key facts:

- More than 3,000 full-length genes have been generated with synthetic biology
- More than 1,500 proteins have been produced and recovered from the CHO culture medium
- The human secretome resource has been used for phenotypic assays in various screening platforms

Milestone 24

2016 Antibody validation

Description:

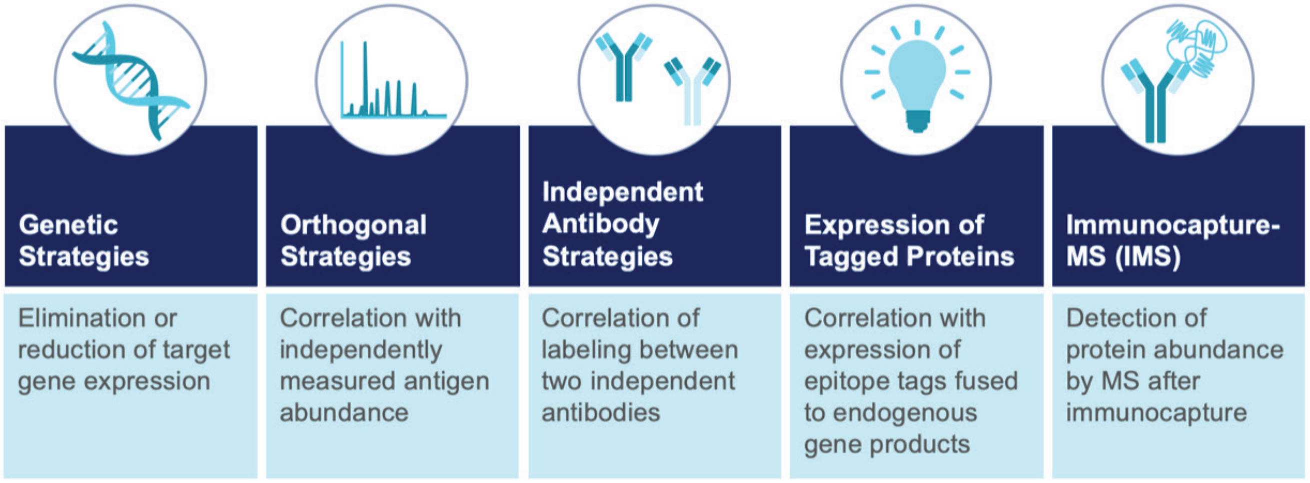
The International Working Group for Antibody Validation (IWGAV) was formed to develop approaches for validating antibodies used in common research applications and to provide guidelines that ensure antibody reproducibility. The working group recommended five conceptual "pillars" for antibody validation to be used in an application-specific manner.

Key publication:

M. Uhlén *et al.*, "A proposal for validation of antibodies," *Nat. Methods* **13**, 823–827 (2016).

Other selected publications:

F. Edfors *et al.*, "Enhanced validation of antibodies for research applications," *Nat. Commun.* **9**, 4130 (2018).
K. Sikorski *et al.*, "A high-throughput pipeline for validation of antibodies," *Nat. Methods.* **15**, 909–912 (2018).



Genetic Strategies	Orthogonal Strategies	Independent Antibody Strategies	Expression of Tagged Proteins	Immunocapture-MS (IMS)
Elimination or reduction of target gene expression	Correlation with independently measured antigen abundance	Correlation of labeling between two independent antibodies	Correlation with expression of epitope tags fused to endogenous gene products	Detection of protein abundance by MS after immunocapture

Figure legend: Proposed conceptual pillars for validation of antibodies. Adapted from Uhlén *et al.* (2016).

Key facts:

- A proposal for validation of antibodies has been suggested by IWGAV
- Many antibody providers are now using principles from the IWGAV guidelines for antibody validation
- Antibodypedia has adopted the IWGAV guidelines to improve evidence-based selection of antibodies
- More than 10,000 antibodies in the HPA have been validated using the IWGAV pillars

Milestone 25

2017 The Subcellular Atlas

Description:

The Subcellular Atlas (also called the HPA Cell Atlas) provides high-resolution insights into the spatial distribution of proteins within cells. The protein expression data was derived from antibody-based profiling using immunofluorescence confocal microscopy. The subcellular distribution of over 12,000 proteins was classified into 30 different organelles and cellular structures. A panel of 64 cell lines was also characterized using transcriptomics. A key finding was that half of all proteins localize to multiple cellular compartments.

Key publication:

P. J. Thul *et al.*, "A subcellular map of the human proteome," *Science* **356**, eaal3321 (2017).

Other selected publications:

L. Barbe *et al.*, "Toward a confocal subcellular atlas of the human proteome," *Mol. Cell. Proteomics* **7**, 499–508 (2008).
C. Stadler *et al.*, "Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells," *Nat. Methods.* **10**, 315–323 (2013).
Stadler *et al.*, "A single fixation protocol for proteome-wide immunofluorescence localization studies," *J. Proteomics* **73**, 1067–1078 (2010).

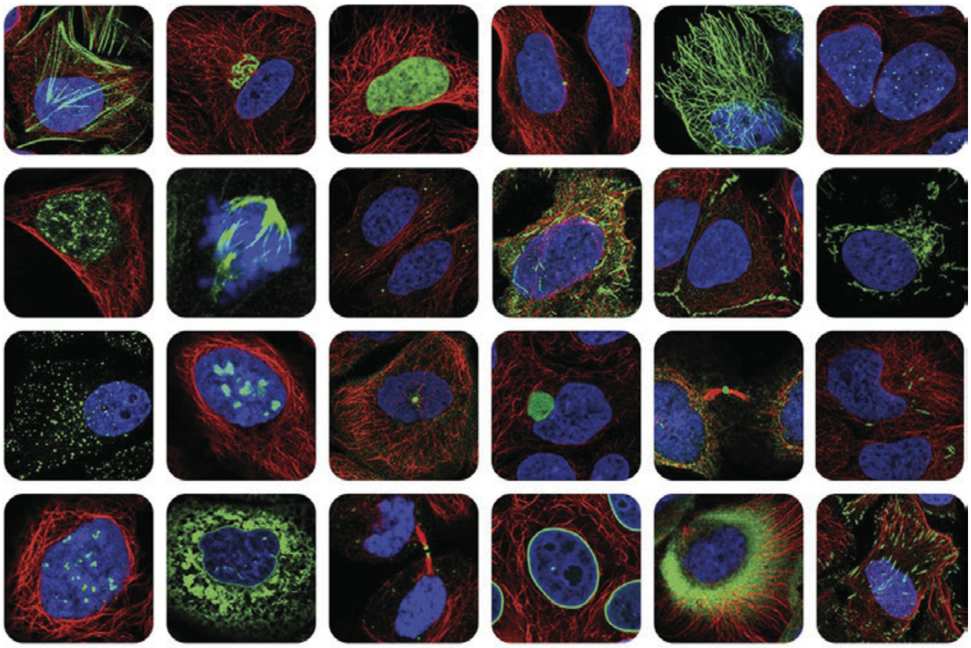


Figure legend: The subcellular locations of 12,003 proteins were determined by immunocytochemistry [also denoted as immunofluorescence (ICC-IF)] and confocal microscopy in cell lines of various origins. Adapted from Thul *et al.* (2017).

Key facts:

- The subcellular locations of over 12,000 proteins were determined using immunofluorescence and high-resolution confocal microscopy
- The Subcellular Atlas resolves the spatial distribution of the human proteome at a subcellular level
- Half of all proteins localize to multiple compartments in the cell

Milestone 26

2017 The Pathology Atlas

Description:

The Pathology Atlas provides the analysis of 17 major cancer types using data from 8,000 patients together with 5 million pathology-based images generated in-house. More than 2.5 petabytes of RNA-seq data from The Cancer Genome Atlas (TCGA) were analyzed, describing the effects of RNA and protein levels on clinical survival. Survival Scatter plots, representing a new method for showing patient survival data, were introduced.

Key publication:

M. Uhlén *et al.*, "A pathology atlas of the cancer transcriptome," *Science* **357**, eaan2507 (2017).

Other selected publications:

A. H. Larsson *et al.*, "Significant association and synergistic adverse prognostic effect of podocalyxin-like protein and epidermal growth factor receptor expression in colorectal cancer," *J. Transl. Med.* **14**, 128 (2016).
B. Glimelius *et al.*, "U-CAN: A prospective longitudinal collection of biomaterials and clinical information from adult cancer patients in Sweden," *Acta Oncol.* **57**, 187–194 (2018).
O. Casar-Borota *et al.*, "Immunohistochemistry for transcription factor T-Pit as a tool in diagnostics of corticotroph pituitary tumours," *Pituitary* **21**, 443 (2018).
S. Lee *et al.*, "TCSBN: A database of tissue and cancer specific biological networks," *Nucleic Acids Res.* **46**, D595–D600 (2018).

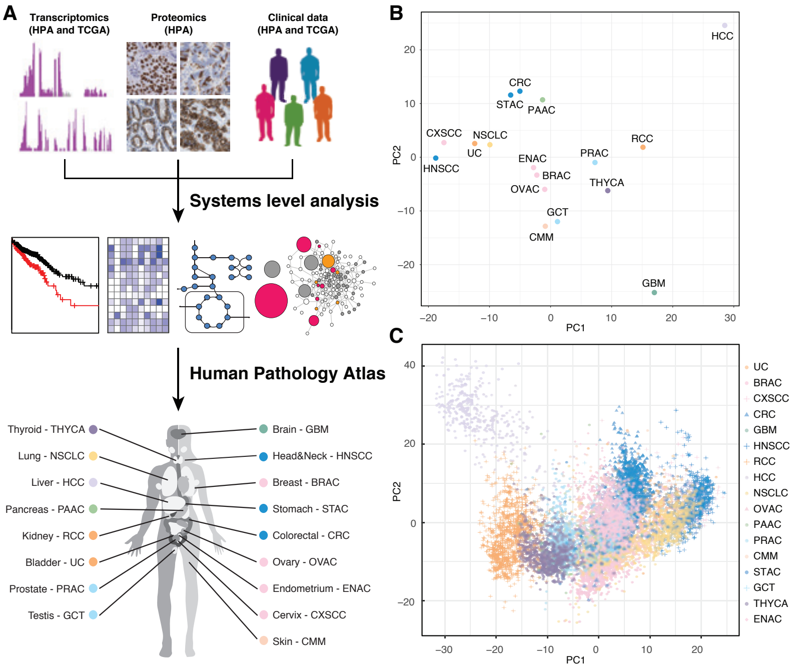


Figure legend: Analysis of the global expression patterns of protein-coding genes in human cancers.

Key facts:

- Analysis of the transcriptome in 17 major cancer types from 8,000 patients was performed
- A new concept for visualizing survival data was published: Survival Scatter plots
- More than 900,000 Survival Scatter plots are shown, covering more than 10,000 genes

Milestone 27

2017 Systems medicine

Description:

Systems medicine is an interdisciplinary subject focusing on systems of biological components and using computational models and experimental technologies such as genomics, transcriptomics, proteomics, metabolomics, and metagenomics. It includes application and development of systems biological methods with an emphasis on integration, analysis, and modeling of big data using biological networks. In the HPA, these concepts were used to study various clinically relevant diseases, including liver metabolism diseases and human cancers. In addition, several drug candidates for various clinical indications have been developed, and several clinical trials have been initiated in humans.

Key publication:

S. Lee *et al.*, "Integrated network analysis reveals an association between plasma mannose levels and insulin resistance," *Cell Metab.* **24**, 172–184 (2016).

Other selected publications:

A. Mardinoglu *et al.*, "Plasma mannose levels are associated with incident type 2 diabetes and cardiovascular disease," *Cell Metab.* **26**, 281–283 (2017).
A. Mardinoglu *et al.*, "Personal model-assisted identification of NAD^+ and glutathione metabolism as intervention target in NAFLD," *Mol. Syst. Biol.* **13**, 916 (2017).
A. Mardinoglu *et al.*, "An integrated understanding of the rapid metabolic benefits of a carbohydrate-restricted diet on hepatic steatosis in humans," *Cell Metab.* **27**, 559–571 (2018).
G. Bidkhori *et al.*, "Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes," *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11874–E11883 (2019).
C. Zhang *et al.*, "The acute effect of metabolic cofactor supplementation: A potential therapeutic strategy against non-alcoholic fatty liver disease," *Mol. Syst. Biol.* **16**, e9495 (2020).

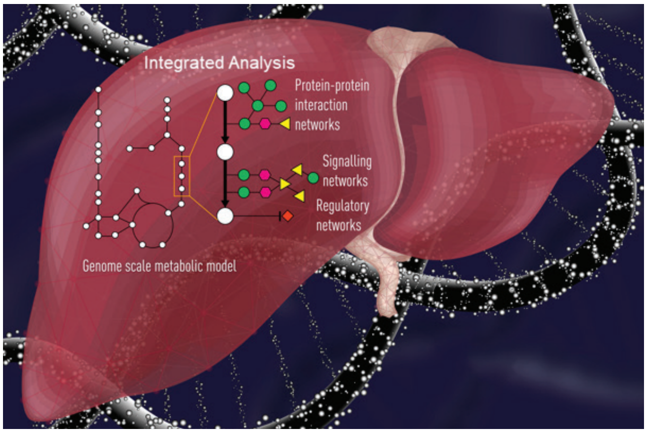

Integrated Analysis
Protein-protein interaction networks
Signalling networks
Regulatory networks
Genome scale metabolic model

Figure legend: Integrated networks, which are the combination of genome-scale metabolic networks, protein–protein interaction networks, signaling networks, and transcriptional regulatory networks, can be used for integration of 'omics data and for discovery of biomarkers and drug targets.

Key facts:

- Integration of clinical data using a holistic approach
- Discovery of potential biomarkers for stratification of patients and early detection of disease
- Identification of novel drug targets for development of efficient treatment strategies
- Systems biology-based drug repositioning for development of therapies

Milestone 28

2018 Wellness profiling and precision medicine

Description:

The resources created within the Human Protein Atlas program have been used for longitudinal analysis of both healthy and disease cohorts by combining classical clinical chemistry, advanced medical imaging, and integrative 'omics involving genomics, transcriptomics, proteomics, metabolomics, microbiomics, and high-throughput cell analysis. The focus has been on a wellness study conducted over 2 years that started in 2015, analyzing preterm children and patients with diabetes, cancer, and cardiovascular disease in collaboration with researchers at Sahlgrenska University Hospital in Gothenburg, Sweden.

Key publication:

A. Tebani *et al.*, "Integration of molecular profiles in a longitudinal wellness profiling cohort," *Nat. Commun.*, **11**, 4487 (2020).

Other selected publications:

M. Neiman *et al.*, "Individual and stable autoantibody repertoires in healthy individuals," *Autoimmunity* **52**, 1–11 (2019).
W. Zhong *et al.*, "Dramatic changes in blood protein levels during the first week of life in extremely preterm infants," *Pediatr. Res.*, online ahead of print (2020).
W. Zhong *et al.*, "Whole-genome sequence association analysis of blood proteins in a longitudinal wellness cohort," *Genome Med.* **12**, 53 (2020).
T. Dodig-Crnkovic *et al.*, "Facets of individual-specific health signatures determined from longitudinal plasma proteome profiling," *EBioMedicine* **57**, 102854 (2020).

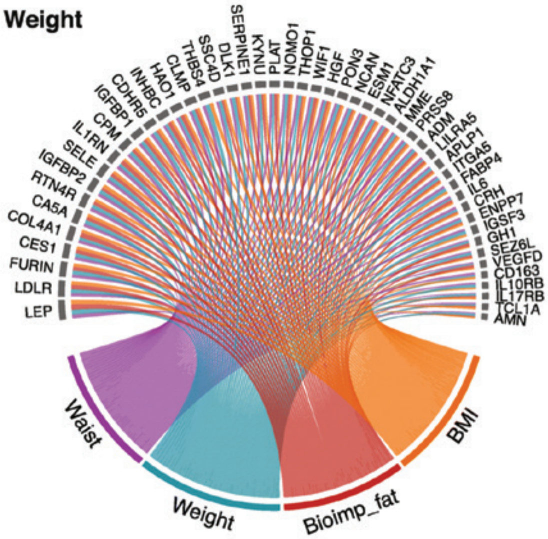


Figure legend: Chord diagram of the 50 most significant proteins related to body composition [bioimpedance fat, bioimpedance muscle, bioimpedance bone, weight, waist, and body mass index (BMI)] based on a longitudinal precision medicine study. The size of the link is defined as the absolute value of the coefficient of the corresponding effect, and proteins are sorted based on the coefficient calculated using mixed-effect modeling. Adapted from Zhong *et al.* (2020).

Key facts:

- 100 healthy individuals have been followed longitudinally during 2 years with multiple samplings
- 200 extremely preterm children have been followed with multiple sampling after birth
- More than 100 proteins have been identified in which variation in plasma levels is genetically determined

Milestone 29

2018 Deep learning and citizen science

Description:

Two approaches for large-scale classification of fluorescence microscopy images were used to analyze subcellular protein patterns in images from the HPA Cell Atlas. An image-classification task was introduced into the online science fiction video game, EVE Online. 320,000 gamers provided more than 32 million image classifications, and the data was combined with deep learning to build a tool to classify proteins into 29 subcellular localization patterns.

Key publication:

D. P. Sullivan *et al.*, "Deep learning is combined with massive-scale citizen science to improve large-scale image classification," *Nat. Biotechnol.* **36**, 820–826 (2018).

Other selected publications:

M. Peplow, "Citizen Science lures gamers into Sweden's Human Protein Atlas," *Nat. Biotechnol.* **34**, 452–453 (2016).

Figure legend: An EVE Online spaceship cruising a "universe" of cells from the HPA.

Key facts:

- Deep learning was combined with massive-scale citizen science
- Project Discovery marks the first time a citizen science task was integrated into a mainstream online computer game
- In 1 year, 320,000 players provided 32 million image classifications and spent a total of 70 working years

Milestone 30

2019 Human secretome annotation

Description:

An analysis of the human secretome was presented, including annotation of the genes encoding proteins that are predicted to be actively secreted to human blood as well as to other parts of the human body, such as the digestive system and other local compartments. The estimated concentration of the proteins detected in human blood, as determined by mass spectrometry–based proteomics or antibody-based immune assays, was also presented.

Key publication:

M. Uhlén *et al.*, "The human secretome," *Sci. Signal.* **12**, eaaz 0274 (2019).

Other selected publications:

L. Fagerberg *et al.*, "Prediction of the human membrane proteome," *Proteomics* **10**, 1141–1149 (2010).
J. M. Schwenk *et al.*, "The human plasma proteome draft of 2017: Building on the Human Plasma PeptideAtlas from mass spectrometry and complementary assays," *J. Proteome Res.* **16**, 4299–4310 (2017).

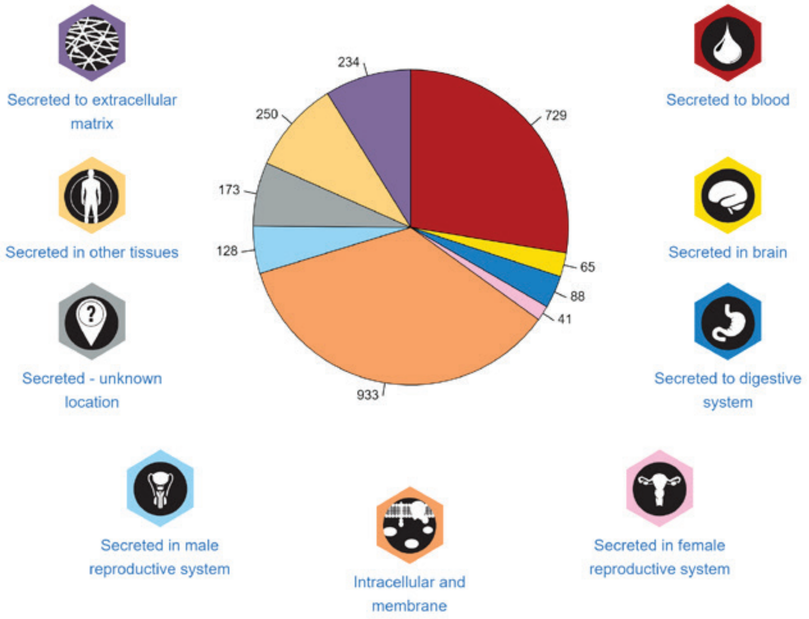


Figure legend: All genes with predicted secreted isoforms were annotated and classified according to their predicted final location in the body, with the major aim to identify proteins secreted into the blood. The largest category contains proteins that after annotation are no longer predicted to be actively secreted, and includes proteins residing in secretory pathway locations or being associated with membranes.

Key facts:

- About 13% of human genes have at least one predicted secreted isoform
- Approximately 700 proteins were predicted to be actively secreted into the blood (corresponding to less than 4% of protein coding genes)
- Another 800 proteins were predicted to be secreted locally, including in male and female reproductive tissues
- There are almost 90 proteins (mainly enzymes) predicted to be secreted into the digestive system
- Many proteins predicted to be secreted into the blood at present lack protein assays for their detection

Milestone 31

2019 The Blood Atlas

Description:

The Blood Atlas provides single-cell type information on genome-wide RNA expression profiles of human protein-coding genes in various B cells, T cells, monocytes, granulocytes, and dendritic cells. The single-cell analysis covers 18 cell types isolated through cell sorting followed by transcriptomics analysis. A genome-wide classification of the proteins with elevated expression in various immune cells was performed. The analysis also included comparisons of the proteins specific for blood in the context of proteins expressed across all tissues and organs in the human body.

Key publication:

M. Uhlén *et al.*, "A genome-wide transcriptomic analysis of protein-coding genes in human blood cells," *Science* **366**, eaax9198 (2019).

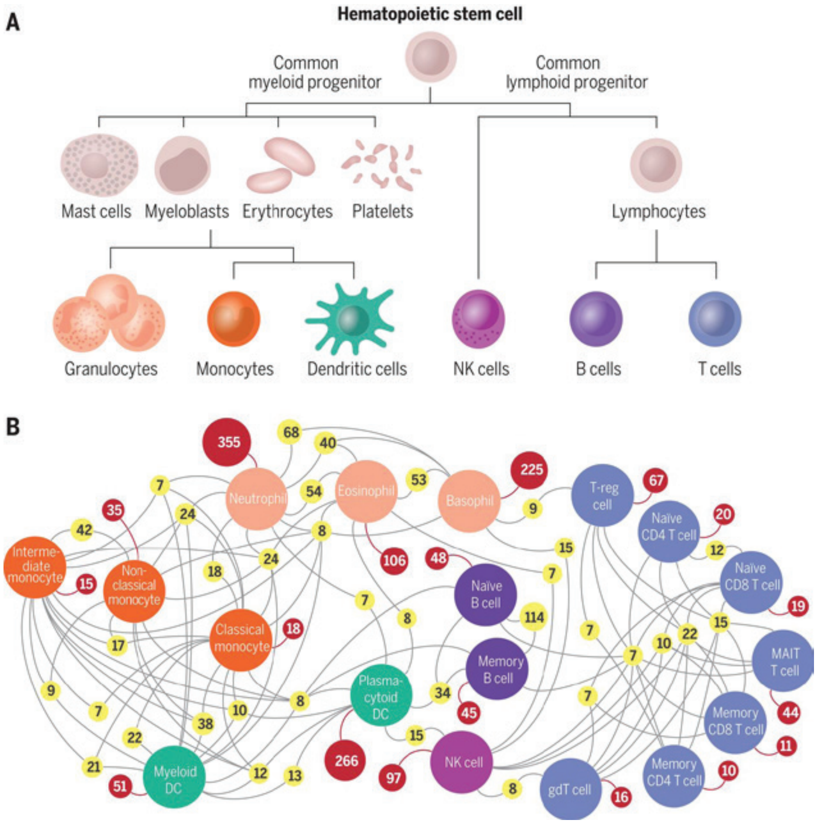


Figure legend: Classification of all blood cell type specific genes. Adapted from Uhlén *et al.* (2019).

Key facts:

- There are 1,448 protein-coding genes that have enriched expression in a single immune-cell type
- 56% of protein-coding genes have elevated expression in at least one of the analyzed tissues and cells
- Only 216 (<1%) of all genes were not detected in any of the tissues analyzed
- 224 genes associated with primary immunodeficiencies in humans were studied

Milestone 32

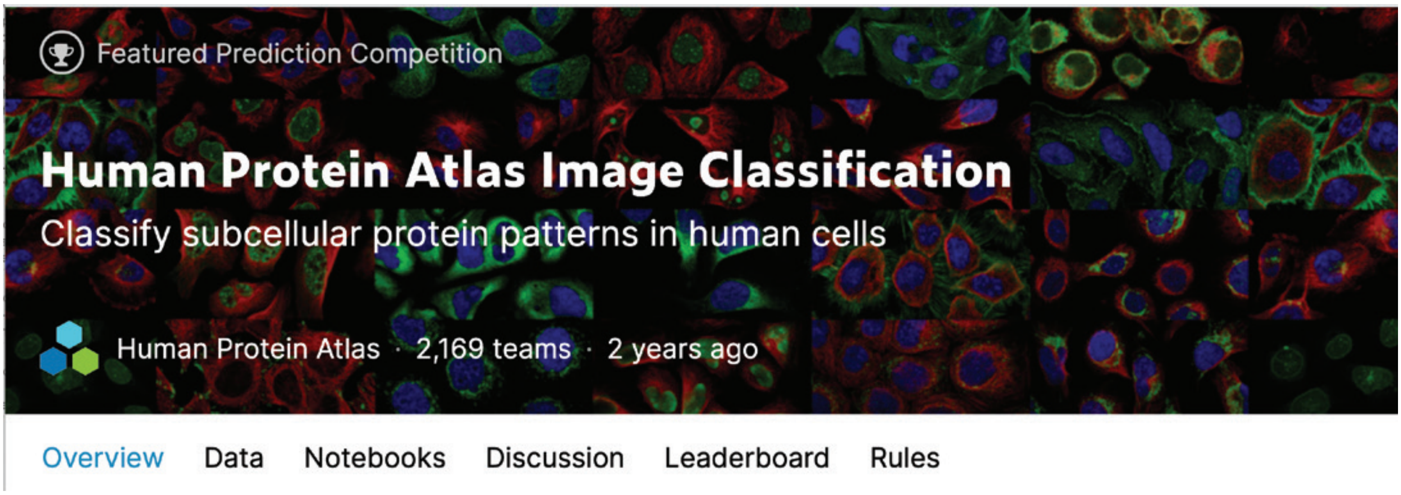
2019 The HPA Kaggle Challenge

Description:

Based on the HPA Cell Atlas image collection, a computational competition was arranged to identify deep-learning solutions for classification of subcellular protein patterns. Challenges included training on highly imbalanced classes and predicting multiple labels per image. More than 2,000 teams participated, and the winning models far outperformed our previous model. These models can be used as classifiers to annotate new images, feature extractors, or pretrained networks for a wide range of biological applications.

Key publication:

W. Ouyang *et al.*, "Analysis of the Human Protein Atlas Image Classification competition," *Nat. Methods* **16**, 1254–1261 (2019).



Featured Prediction Competition

Human Protein Atlas Image Classification

Classify subcellular protein patterns in human cells

Human Protein Atlas · 2,169 teams · 2 years ago

Figure legend: Opening page of the HPA Kaggle competition.

Key facts:

- 2,200 teams presented a diverse set of deep-learning solutions
- The competition managed multilabel classification with proteins localized to several subcellular compartments
- The top model can be used as a feature extractor to embed spatial localization in cellular models
- The top-ranking models performed better than any previously published model

Milestone 33

2020 The Brain Atlas

Description:

The Brain Atlas (www.proteinatlas.org/brain) provides the protein expression of the mammalian brain by visualization and integration of data from three mammalian species (human, pig, and mouse). Transcriptomics data is combined with affinity-based protein in situ localization down to single-cell detail. The expression data is based on human protein–coding genes and one-to-one orthologues in pig and mouse. The gene expression in the brain is visually summarized into 10 main brain regions (olfactory region, cerebral cortex, amygdala, hippocampal formation, basal ganglia, hypothalamus, thalamus, midbrain, combined pons and medulla, and cerebellum) that are used for regional expression abundance classification. In the Brain Atlas, expression information is also shown for the spinal cord, corpus callosum, retina, and pituitary gland. Each of the 10 brain regions can be reviewed and further explored on individual pages, which provide classification overviews, interactive lists and figures, and highlighted examples of regionally specialized cells and protein expression. High-resolution, stained images derived from antibody-based immunolabeling techniques are also included in the atlas.

Key publication:

E. Sjöstedt *et al.*, "An atlas of the protein-coding genes in the human, pig and mouse brain," *Science* **367**, eaay5947 (2020).

Other selected publications:

E. Sjöstedt *et al.*, "Defining the human brain proteome using transcriptomics and antibody-based profiling with a focus on the cerebral cortex," *PLOS One* **10**, e0130028 (2015).
J. Mulder *et al.*, "Tissue profiling of the mammalian central nervous system using human antibody-based proteomics," *Mol. Cell. Proteomics* **8**, 1612–1622 (2009).
J. Mulder *et al.*, "Systematically generated antibodies against human gene products: High throughput screening on sections from the rat nervous system," *Neuroscience* **146**, 1689–1703 (2007).

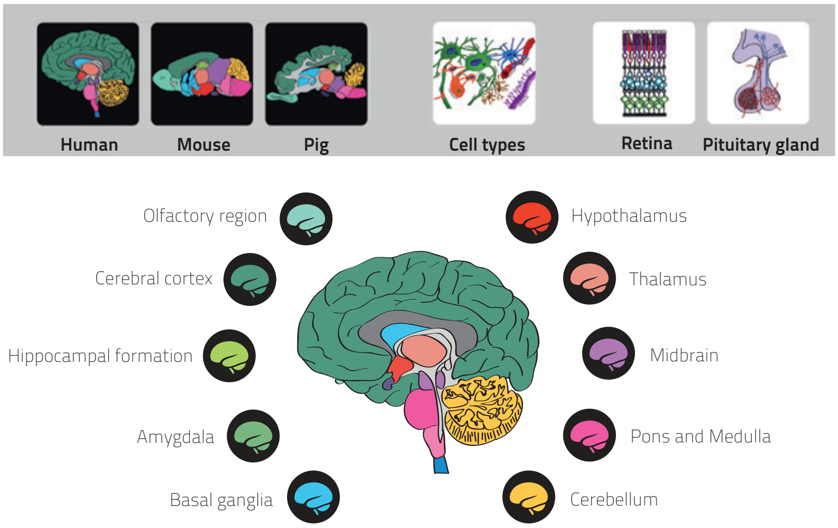


Figure legend: The HPA Brain Atlas includes gene-expression data for human protein–coding genes as well as different exploratory summary pages.

Key facts:

- The HPA Brain Atlas includes regional expression data from human, pig, and mouse
- Separate summary pages describing the analysis with examples and biological findings
- 271 genes with a protein profile of the whole mouse brain
- 815 examples of protein location in human brain or retina

Milestone 34

2020 The Metabolic Atlas

Description:

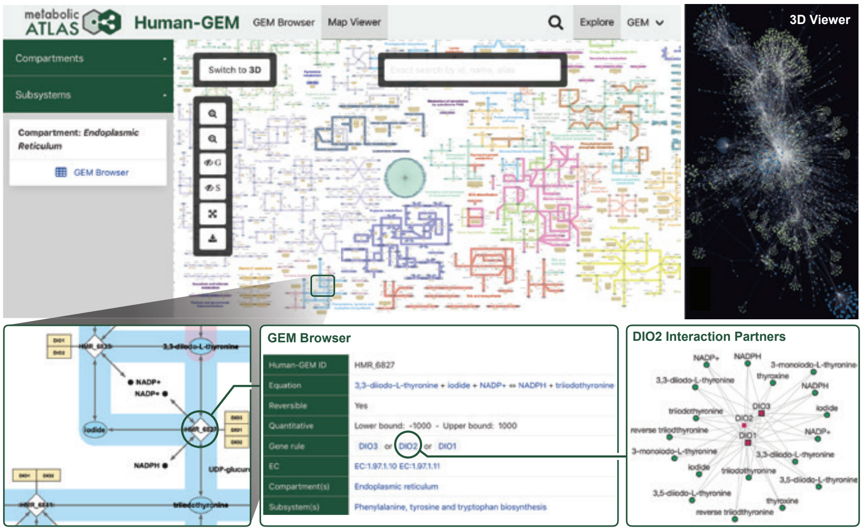
The Metabolic Atlas portion of the Tissue Atlas enables exploration of protein function and tissue-specific gene expression in the context of the human metabolic network. For proteins involved in metabolism, a metabolic summary is provided that describes the metabolic subsystems/pathways, cellular compartments, and number of reactions associated with each protein. Over 120 manually curated metabolic pathway maps facilitate the visualization of each protein's participation in different metabolic processes.

Key publication:

J. L. Robinson *et al.*, "An atlas of human metabolism," *Sci. Signal.* **13**, eaaz1482 (2020).

Other selected publications:

A. Mardinoglu *et al.*, "Integration of clinical data with a genome-scale metabolic model of the human adipocyte," *Mol. Syst. Biol.* **9**, 649 (2013).
F. Gatto *et al.*, "Chromosome 3p loss of heterozygosity is associated with a unique metabolic network in clear cell renal carcinoma," *Proc. Natl. Acad. Sci. U.S.A.* **111**, E866–E875 (2014).
A. Mardinoglu *et al.*, "Genome-scale metabolic modeling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease," *Nat. Comm.* **5**, 3083 (2014).
R. Agren *et al.*, "Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling," *Mol. Syst. Biol.* **10**, 721 (2014).



GEM Browser

Human-GEM ID	HMR_6827
Equation	3,3-diiodo-L-thyronine + iodide + NADP+ ⇔ NADPH + triiodothyronine
Reversible	Yes
Quantitative	Lower bound: -1000 - Upper bound: 1000
Gene rule	DIO3 or DIO2 or DIO1
EC	EC:1.97.1.10 EC:1.97.1.11
Compartment(s)	Endoplasmic reticulum
Subsystem(s)	Phenylalanine, tyrosine and tryptophan biosynthesis

Figure legend: The Metabolic Atlas provides an overview of all metabolic reactions operating in the human cell (yeast is also available, and other cell types will be added in the future). In addition to being a repository of data, the atlas provides a visualization of metabolism, including links between metabolites, proteins, and genes.

Key facts:

- The Human Metabolic Atlas encompass 13,417 reactions linking 4,164 unique metabolites and 3,625 genes
- The atlas provides genome-scale metabolic models for 53 healthy tissues and 33 cancers
- The genome-scale metabolic model for human cells, Human1, can be used to predict gene essentiality in many human cell lines
- The Human1 model has been expanded to include kinetic information enabling enzyme-constrained model simulations

Milestone 35

2020 The fight against the novel coronavirus

Description:

The HPA consortium has been engaged in various ways to aid in the fight against the health consequences of the novel coronavirus pandemic. The program involves efforts to increase the knowledge base on the disease and to develop diagnostic tools and therapeutic drugs to combat the pandemic. A new platform for serological assays to screen for specific antibodies to SARS-CoV-2 was developed using several of the HPA platforms, and this test has now been used to screen different cohorts in Sweden to determine the prevalence of antibodies, indicating past infection with the virus. In addition, a diagnostic laboratory has been set up by HPA researchers at the Karolinska Institutet (Stockholm) in collaboration with the Science for Life Laboratory, to expand the capabilities of viral analysis of swab tests. The lab has conducted a large portion of all tests for acute infection performed in Sweden. Furthermore, AbClon, a company founded by South Korean scientists and HPA researchers, has developed antibodies that recognize SARS-CoV-2. The aim is to generate a therapeutic antibody to combat the disease in the clinic. Finally, the HPA has provided a list of proteins implicated in the disease. The presence in the human body of the enzyme angiotensin-converting enzyme 2 (ACE2), previously proposed to be the main target for coronavirus attachment to the surface of human cells, was analyzed in more depth. The results raise questions regarding the role of ACE2 for infection of human lungs and highlight the need to further explore the route of transmission during SARS-CoV-2 infection.

Key publications:

F. Hikmet *et al*., "The protein expression profile of ACE2 in human tissues," *Mol. Syst. Biol*. **16**, e9610 (2020).
A.Rudberg *et al.,* "SARS-CoV-2 exposure, symptoms and seroprevalence in healthcare workers in Sweden," *Nat. Commun.* **11**, 5064 (2020).

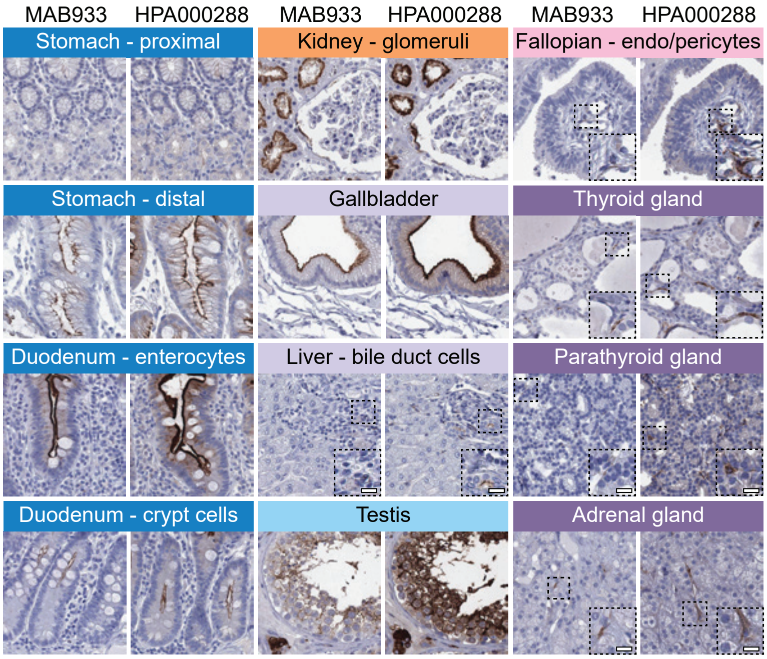


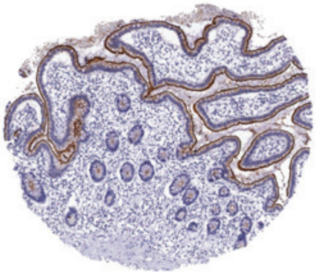
Figure legend: Cell-type specific localization of ACE2 in human tissues using validated HPA antibodies.

Key facts:

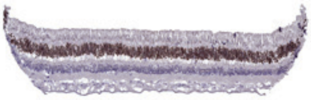
- A new concept was developed to screen for antibodies to the novel coronavirus, SARS-CoV-2, based on suspension bead arrays
- A laboratory for viral detection was set up and used to screen for virus infection across different cohorts
- Several potential antibodies to the SARS-CoV-2 spike protein were developed with neutralizing activity
- Expression analysis based on the HPA reveals that the protein ACE2 has limited expression in human lung tissue

Tissue profiling

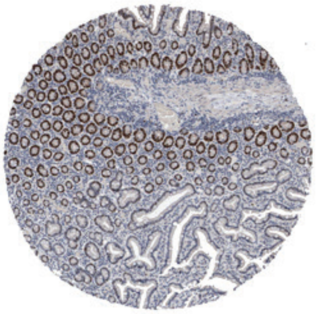
Tissue profiling is at the core of the HPA. The atlas includes protein expression data from 44 normal human tissue types, derived from antibody-based protein profiling using immunohistochemistry to determine the location and level of expression of different proteins. The expression pattern is visualized through the conversion of the chromogen 3,3'-diaminobenzidine (DAB) into a brown precipitate at the sites of antibody-target binding. Below are four examples of tissue profiles from the atlas that provide an image plus a description of the protein being detected.



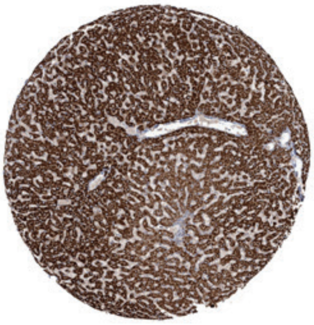
Angiotensin Converting Enzyme 2 (ACE2) in Small Intestine
ACE2 is a carboxypeptidase negatively regulating the renin-angiotensin system (RAS), which can induce vasodilation by cleaving angiotensin II. Due to the importance of ACE in cardiovascular disease, the use of ACE inhibitors for treatment of high blood pressure and heart failure, and recently the suggested involvement of ACE2 in COVID-19 disease, there has been a large interest in understanding the function and expression of ACE2 in various human organs.



Cone-Rod Homebox (CRX) in Retina
An example of a protein located in the nuclei of photoreceptor cells is cone-rod homeobox (CRX), a photoreceptor-specific transcription factor which plays a role in the differentiation of photoreceptor cells. This homeodomain protein is necessary for the maintenance of normal cone and rod function. Mutations in this gene are associated with photoreceptor degeneration.

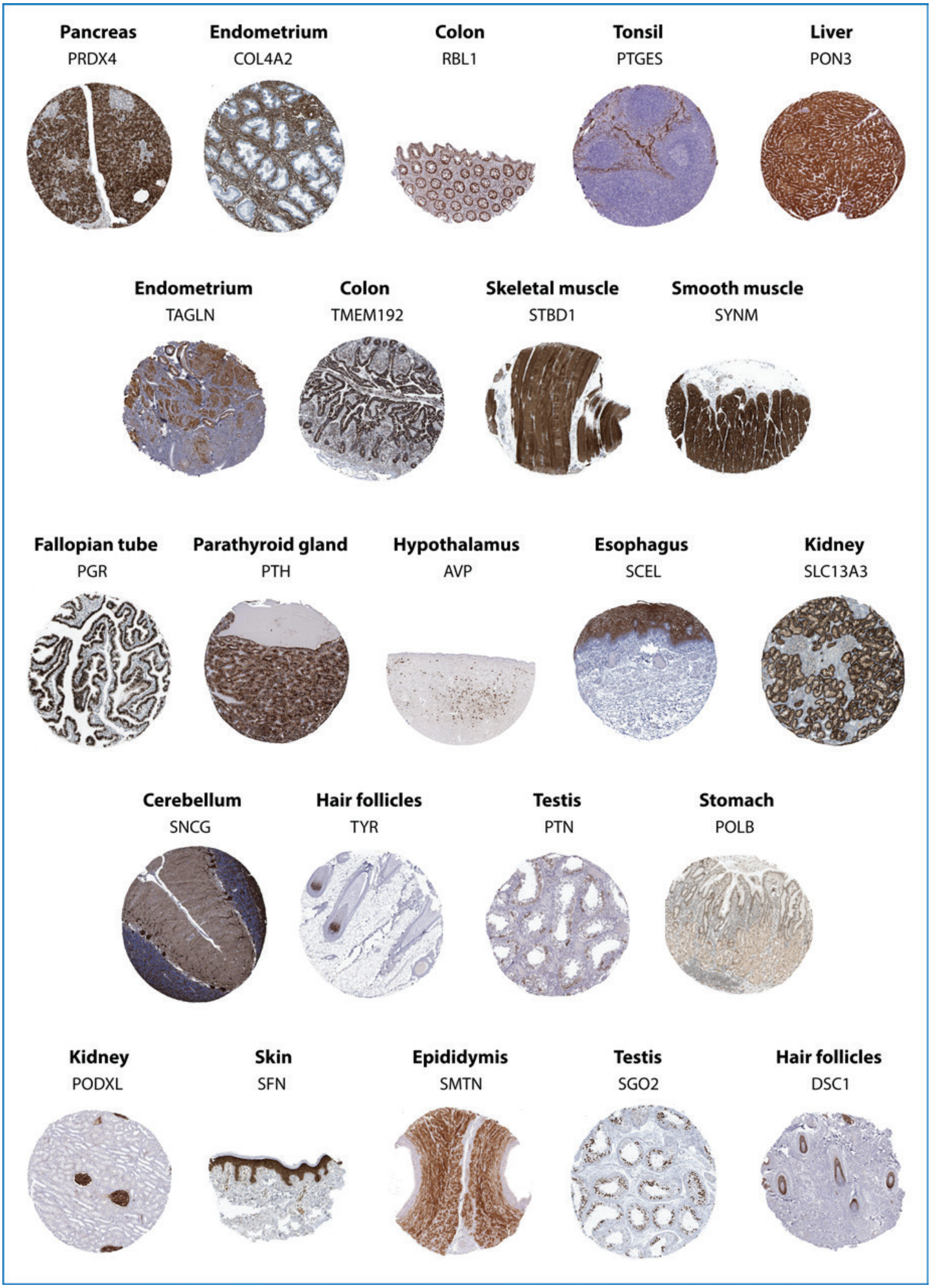


Minichromosome Maintenance Complex Component 6 (MCM6) in Duodenum
MCM6 acts as a component of the MCM complex, which is the putative replicative helicase essential for 'once per cell cycle' DNA replication initiation and elongation in eukaryotic cells. The MCM complex possesses DNA helicase activity, and may act as a DNA unwinding enzyme.



Monoglyceride Lipase (MGLL) in Liver
MGLL is a serine hydrolase that is involved in the degradation of fatty acid molecules and endogenous cannabinoids (endocannabinoids). It increases the levels of free fatty acids and contributes to the regulation of endocannabinoid signaling, nociception and the perception of pain in the nervous system.

The HPA provides more than 10 million annotated immunohistochemistry images showing staining of proteins across the human body. Below are a few examples from different tissues and organs.



Subcellular profiling

Drilling down to the next layer below tissue profiling to determine the subcellular location of proteins can provide critical clues to their function under normal and disease conditions. The Subcellular Atlas (also called the HPA Cell Atlas) provides a knowledge-based classification of protein localization into more than 30 different subcellular structures. Below is an example of proteins that localize to different subcompartments of nucleoli. The transcriptional corepressor NOC2L localizes to the entire nucleolus, while the ribosomal assembly factor NOP56 localizes specifically to nucleoli fibrillar centers, and the marker of mitosis, MKI67, is found mainly at the rim of nucleoli.

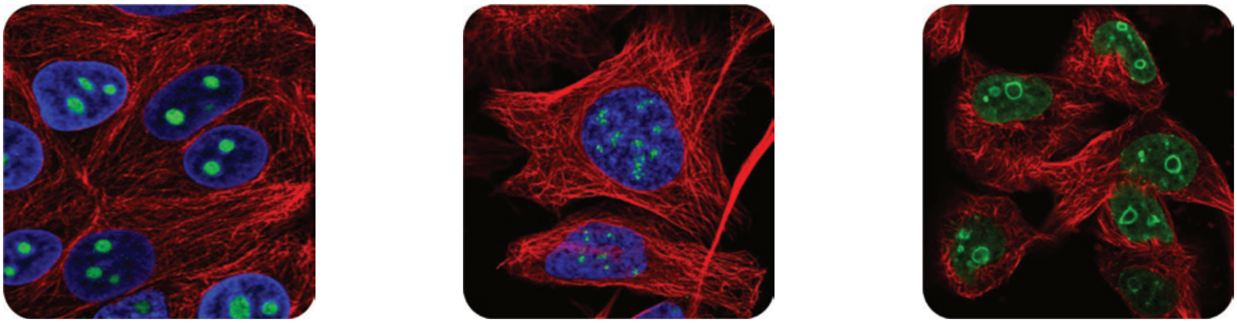


Figure 1: (L to R) NOC2L protein expression in in MCF7 cultured cells; NOP56 expression in U-2 OS cells, and MKI67 expression in U-251 MG cells.

The Subcellular Atlas provides information about single-cell variation in protein expression, both in terms of staining intensity and in spatial location. Many of the proteins that display single-cell variation have also been stained in the U-2 OS-FUCCI cell line in order to characterize potential cell-cycle dependence. These cells express fluorescent markers for the G1 phase (GMNN) and for the S/G2 phases (CDT1), enabling correlation of staining intensity with cell-cycle progression.

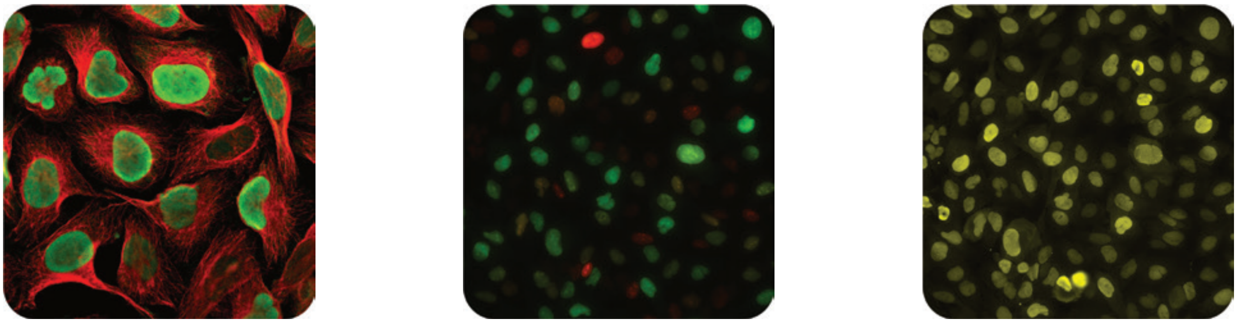


Figure 2: Top2A expression in U-2 OS cells (left) and expression of GMNN (green) and CDT1 (red) in U-2 OS-FUCCI cultured cells (center). Right panel shows Top2A expression in U-2 OS-FUCCI cells.

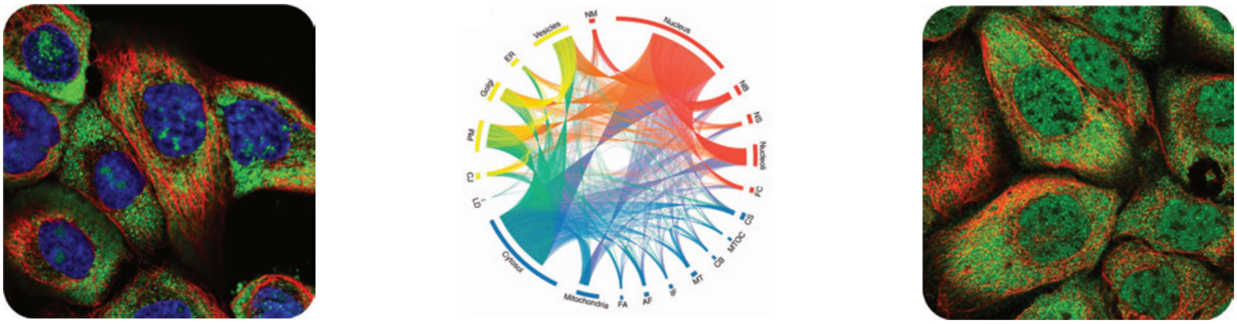


Figure 3: (L to R) RPL19 protein expression in A-431 cells; a circular plot connecting multilocalizing proteins to their subcellular compartments; CRTC3 expression in A-431 cells.

A large portion of the proteins in the Subcellular Atlas localize to more than one compartment. These multilocalizing proteins are particularly prominent in the nucleus, cytosol, and plasma membrane. Above is an example of how the ribosomal protein RPL19 localizes to nucleoli, the cytosol, and the endoplasmic reticulum. Another example, CRTC3, is a transcription coactivator that translocates from the cytosol to the nucleus upon activation. It has been identified as an interaction partner of the SARS-CoV-2 encoded RNA-dependent RNA polymerase Nsp12 [C. J. Gordon ***et al., J. Biol. Chem.*** **295, 4773–4779 (2020)].**

Brain profiling

The brain includes many different cell types and subclasses of the different cells. Protein profiling of mouse and human brain enables a detailed investigation of the protein location relative to cells and structures in the brain. Below are examples of proteins that localize to different cells and structures, with mouse brain shown on the left (fluorescent detection) and human brain on the right (chromogen detection). The cartoon in the center indicates the relevant structure being observed. The target gene name and protein location are detailed under each image. All images are available at www.proteinatlas.org/humanproteome/brain/cell+types.

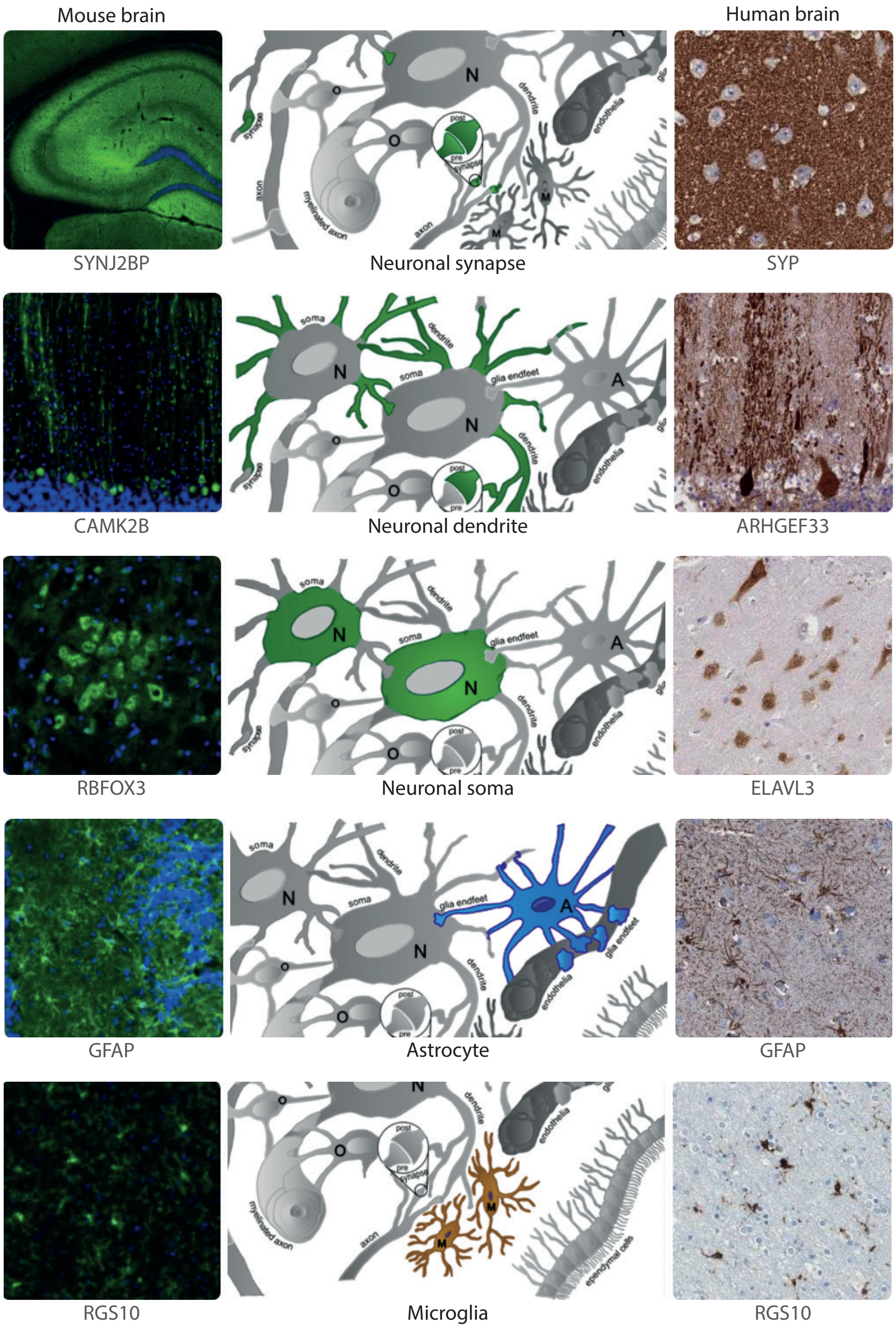


Figure legend: Fluorescent staining of mouse brain (left) and chromogenic staining of human brain (right) for various proteins, listed under each panel. The center sketch indicates the cell or structure being detected.

The atlas also contains detailed protein profiling information obtained from serial sectioning of the brain. Below, serial sections of a young mouse brain show protein localization throughout the organ. Whole mouse brain profiles of 271 different proteins can be found at www.proteinatlas.org/brain.

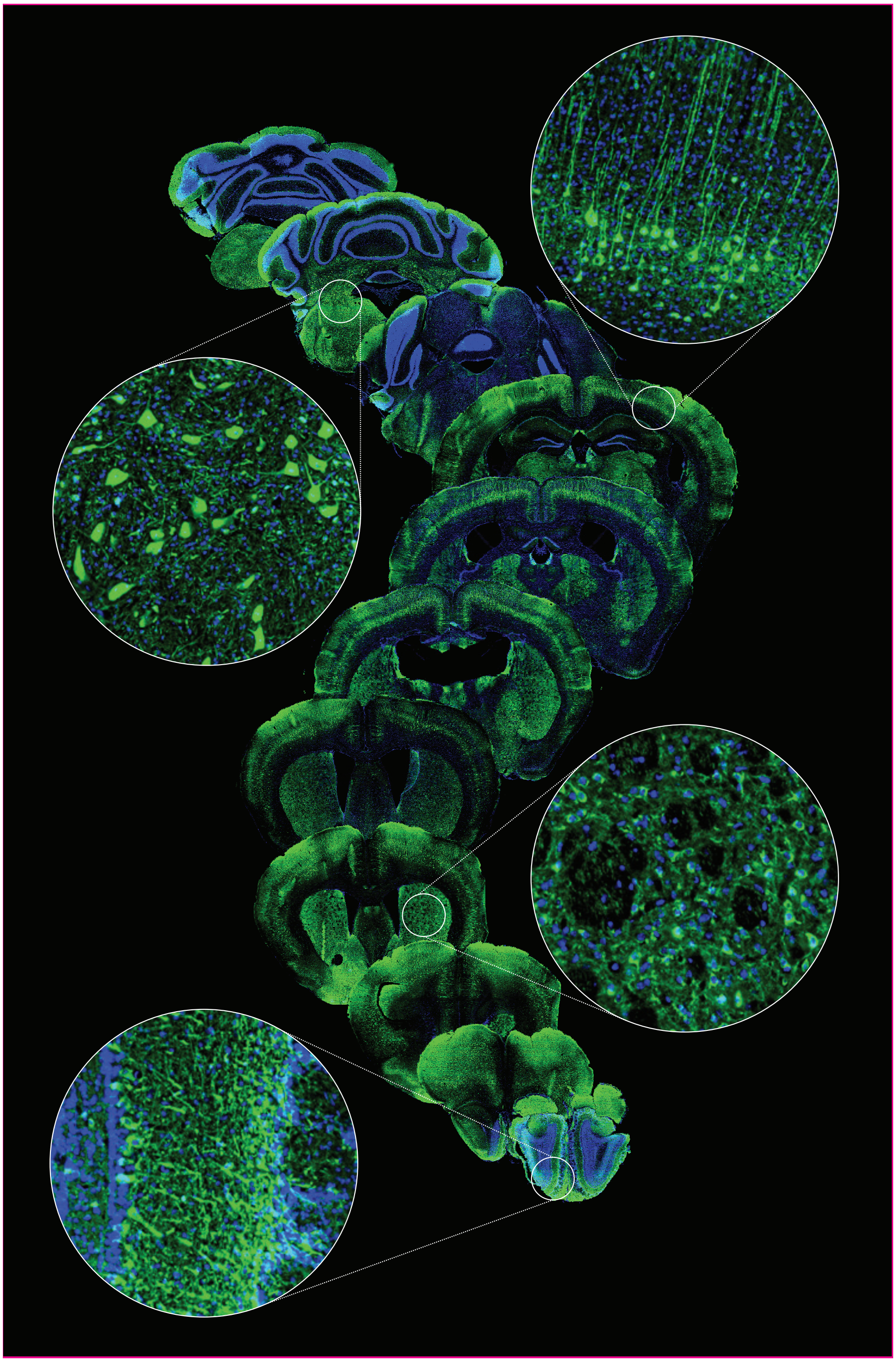
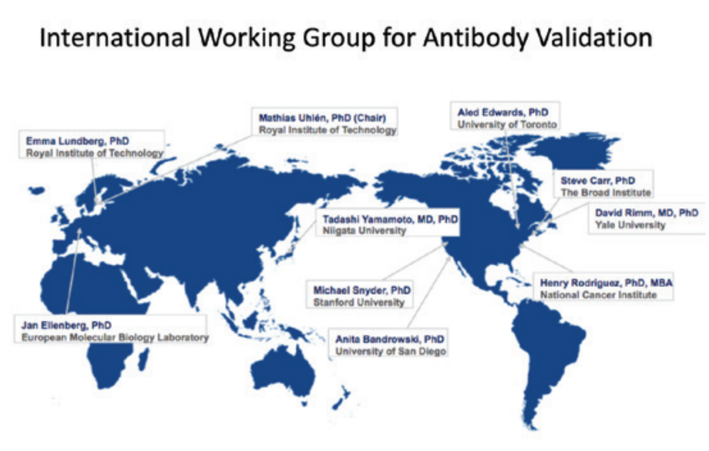


Figure legend: Serial profiling of a 2-month-old C57BL/6J mouse shows fluorescent immunolabeling of calcium/calmodulin-dependent protein kinase II β (CAMK2B).

Validation of antibodies

The HPA has spent considerable resources to validate all antibodies used in the program. More than 20,000 antibodies have passed the validation criteria of HPA, and more than half of these antibodies also meet the stringent criteria for enhanced validation proposed by the International Working Group for Antibody Validation (IWGAV) published in 2016.[1] Three main issues are critical for reliable antibody use; (1) target binding, (2) cross-reactivity, and (3) reproducibility. HPA has focused on these to provide the most accurate data possible in the profiles of published proteins.


International Working Group for Antibody Validation
Emma Lundberg, PhD
Royal Institute of Technology
Mathias Uhlén, PhD (Chair)
Royal Institute of Technology
Aled Edwards, PhD
University of Toronto
Steve Carr, PhD
The Broad Institute
David Rimm, MD, PhD
Yale University
Tadashi Yamamoto, MD, PhD
Niigata University
Michael Snyder, PhD
Stanford University
Henry Rodriguez, PhD, MBA
National Cancer Institute
Jan Ellenberg, PhD
European Molecular Biology Laboratory
Anita Bandrowski, PhD
University of San Diego

A primary consideration when using antibodies for research is whether the analysis will be performed on native or denaturated (nonnative) protein targets. When performing antibody-based protein detection, a range of sample preparation methods are used, including heat inactivation, detergents, and solvents (see table below). For most of the data in HPA, polyclonal antibodies recognizing several epitopes were used, since the primary applications have been western blotting, immunohistochemistry, and immunocytochemistry. Monoclonal antibodies recognizing single epitopes are suitable for applications in which the target protein remains in its native conformation during the analysis. In the HPA program, monoclonal antibodies have often been found to give misleading results due to unintentional denaturation of the native protein. It is important to note that all antibodies used in the HPA program have been purified by affinity chromatography using the antigen as the ligand to ensure that they are specific for the target protein and that the nonspecific antibodies in the polyclonal serum are removed.

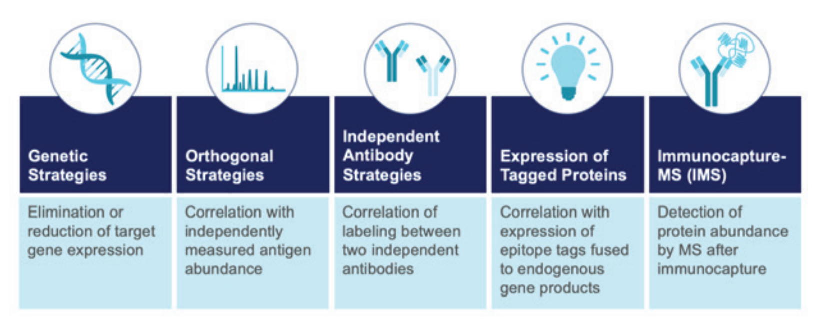
Antibody application	Sample handling	Description
Western blotting (WB)	Denaturing (SDS)	Proteins are treated with detergent (sodium dodecyl sulfate, SDS) but might be partially refolded when blotted onto filters.
Immunohisto-chemistry (IHC)	Denaturing (temperature and formalin)	Proteins are cross-linked with formalin and then often treated with high temperatures (115°C) for epitope retrieval.
Immunocyto-chemistry (ICC)	Denaturing (solvents)	Cellular proteins are normally fixed; permeabilization is often needed. Organic solvents (such as alcohols and acetone) or cross-linking reagents (such as paraformaldehyde) are typically used.
Immunoprecipitation (IP)	Native	Lysis buffers are often used to stabilize native protein conformation, inhibit protease activity, and allow the protein to be released from the cell or tissue.
Sandwich assays	Native	Blood samples (plasma and serum) are not normally pretreated with denaturing agents; tissues are usually treated with lysis buffers.
Flow cytometry	Native	Cells are not normally pretreated with denaturing agents when extracellular markers are used, but permeabilization might be necessary for intracellular markers.
Reverse-phase protein arrays	Native	Blood samples (plasma and serum) are not normally pretreated with denaturing agents; tissue samples are often treated with lysis buffers.

IWGAV was formed as a coalition of researchers from Europe, Asia, and North America (see figure) with the goal of setting guidelines for the validation of antibodies used across common research applications. The working group proposed five pillars for enhanced validation that require no prior knowledge of the target protein, stressing the importance of performing validation in an application-specific manner and using sample handling mirroring that used at the bench under standard conditions for that application.

[1] M. Uhlen *et al., Nat. Meth.* **13,** 823–827 (2016).

In the HPA program, the pillars are used for the following applications (see figure below): Western blot (WB), immunohistochemistry (IHC), and immunocytochemistry (ICC) using confocal microscopy. The various pillars are described in more detail below.

Genetic validation: Knockdown or knockout of the target protein using genetic methods, such as CRISPR or siRNA, in a suitable cell line. The staining of the antibody is evaluated before and after knockdown of the corresponding target gene.



Genetic Strategies
Elimination or reduction of target gene expression

Orthogonal Strategies
Correlation with independently measured antigen abundance

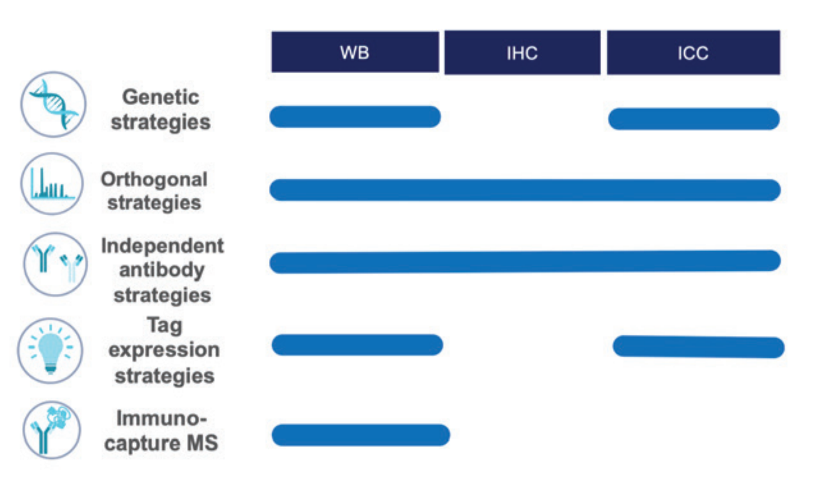
Independent Antibody Strategies
Correlation of labeling between two independent antibodies

Expression of Tagged Proteins
Correlation with expression of epitope tags fused to endogenous gene products

Immunocapture-MS (IMS)
Detection of protein abundance by MS after immunocapture

Orthogonal validation: Comparing the staining pattern with an antibody-independent method analyzing the expression level of the target protein. The levels of the target protein in the different samples determined by the two independent methods must show the same pattern.

Independent antibody validation: Comparing the staining pattern using two independent antibodies with nonoverlapping epitopes. The staining pattern generated by the two antibodies is compared in at least two tissues or cell lines, preferably expressing the target protein at different levels. The two antibodies must show similar results.



Recombinant expression validation: Overexpression of the target protein in a cell line preferably not expressing that target protein, or recombinant expression of a fluorescently tagged version of the target protein in a cell line preferably at an endogeneous level. The staining is evaluated by comparing the signal by the overexpressed or tagged version of the target protein with the unmodified or endogenous target protein.

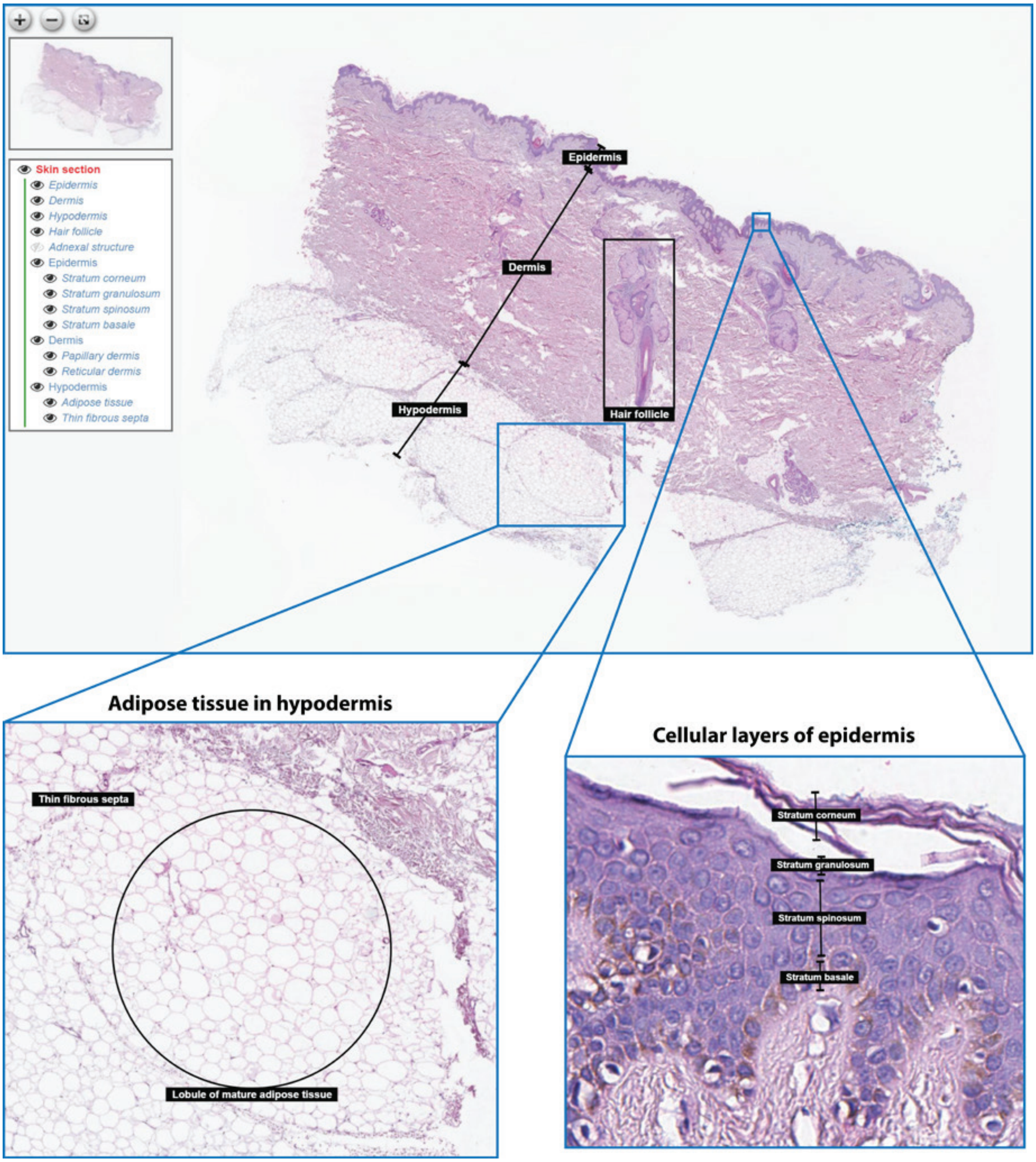
Capture mass spectrometry (MS) validation: Comparing the staining pattern and protein size of the antibody with results obtained by a capture MS method. The size detected by the antibody should be equivalent to the size of the corresponding target protein detected by capture MS.

HPA dictionaries

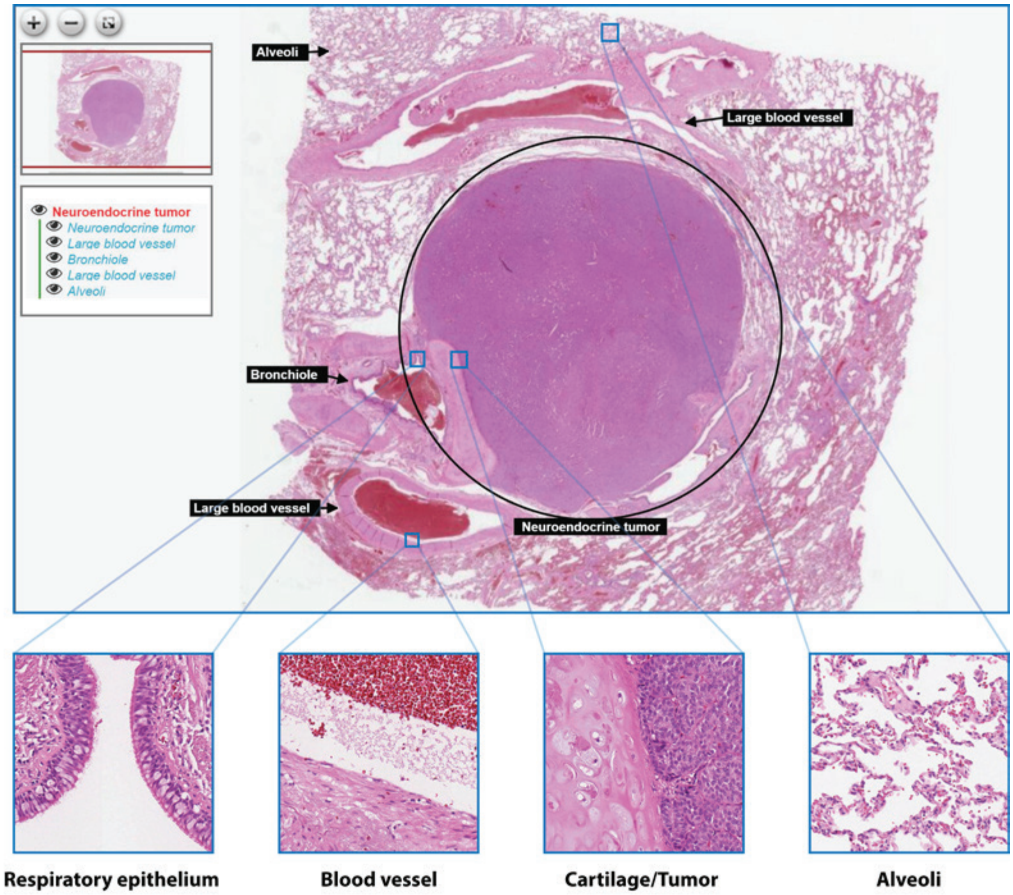
The aim of the dictionary (www.proteinatlas.org/learn/dictionary) is to facilitate the interpretation and use of the image-based data available in the HPA. It also serves as a valuable tool for training and understanding tissue histology, pathology, and cell biology. The dictionary consists of normal histological samples, pathology samples, and cell structures, and provides an unmatched resource for the exploration of tissue samples at different levels:

- **A new flexible exploration tool:** Freely explore large, high-resolution images of hematoxylin and eosin stained tissue sections corresponding to both normal and cancer tissues, at different magnifications.
- **The cell-structure dictionary:** Covers subcellular structures and is based on immunofluorescence and confocal microscopy images, using different color channels to highlight the organellar structure of the cell.
- **Clickable annotation list:** Simple navigation of annotated structures with the option to show/hide annotations.
- **Normal pig tissues:** Normal tissues from pig have been included to enable better tissue comparison between mammals.

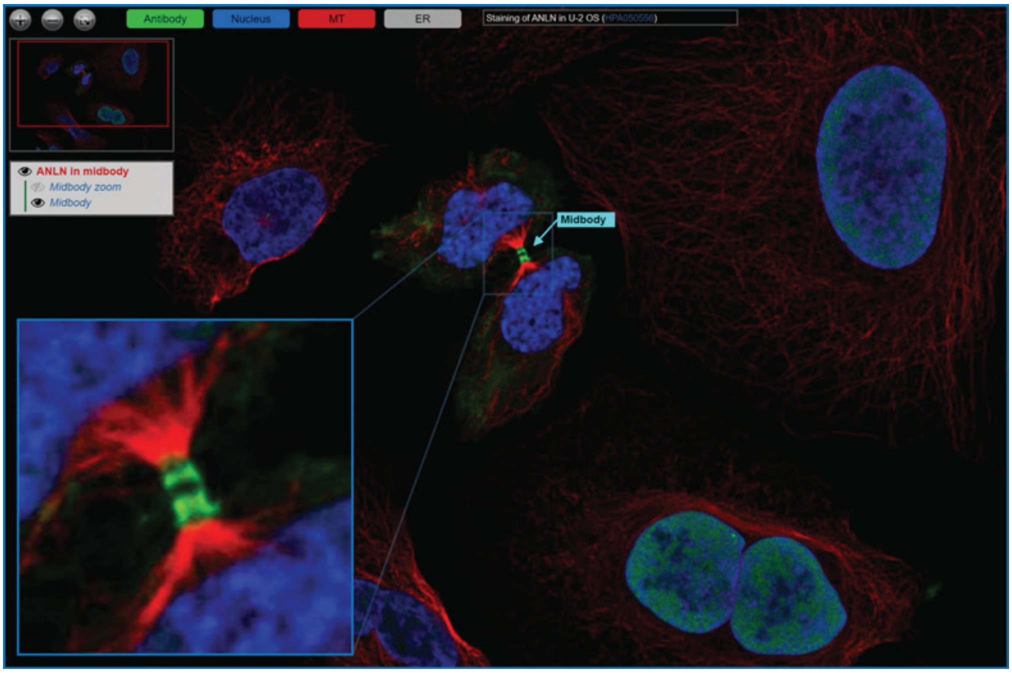
Normal tissue histology – Skin:



Pathology – Neuroendocrine tumor in lung:



Cell structures – Midbody:



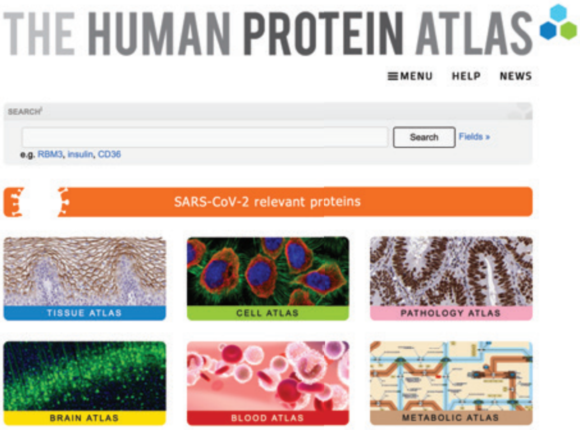
Staining of ANLN in U-2 OS (HPA050556)

Figure legend: Visualization of midbody formation during cell division using antibody-based staining of ANLN protein (green).

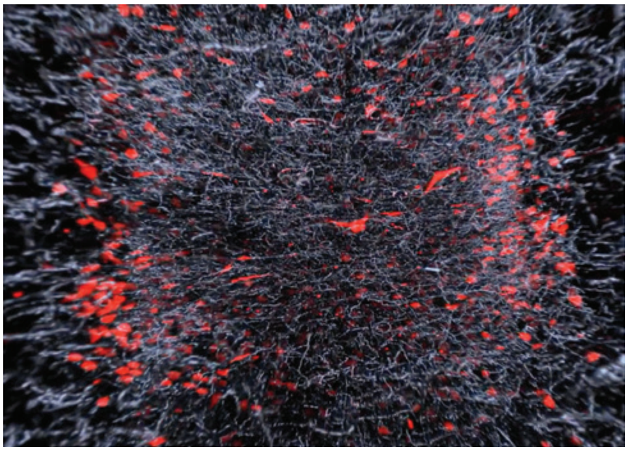
HPA tutorial videos

A number of tutorial videos have been produced to show different aspects of the HPA and its content. They can be accessed at the HPA video channel: https://www.proteinatlas.org/learn/videos. Here is information about some of these movies.

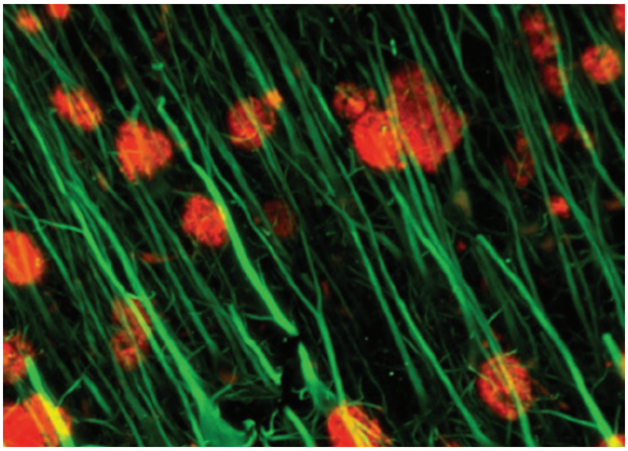
Introduction to the Human Protein Atlas



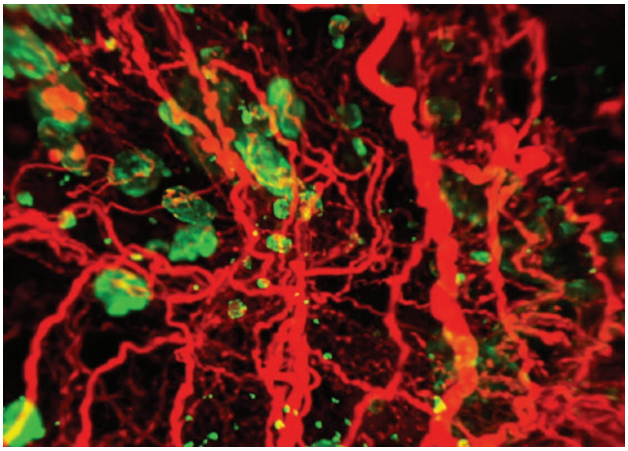
Sleep, orexin, narcolepsy



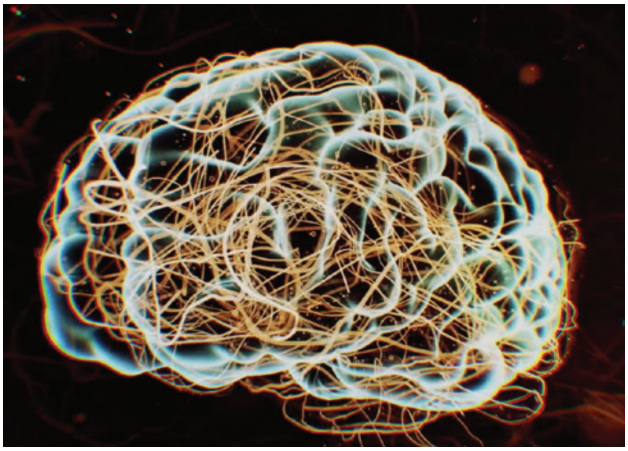
Alzheimer's disease



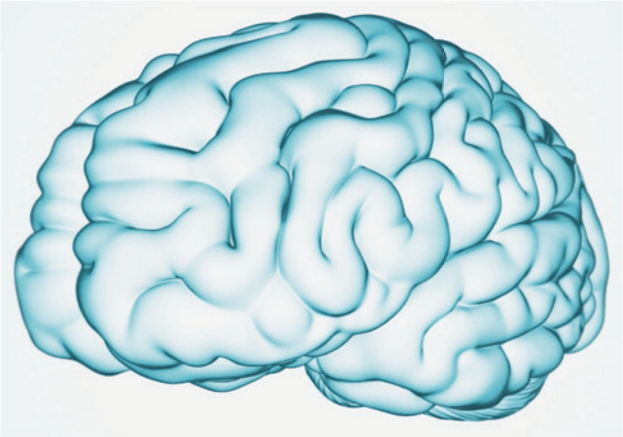
Insulin in the pancreas



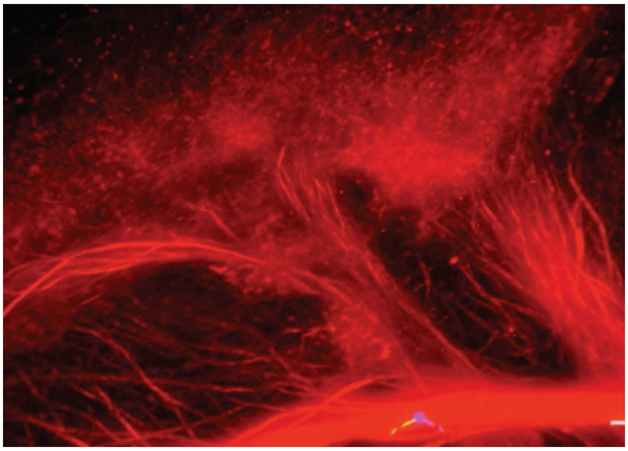
Parkinson's disease



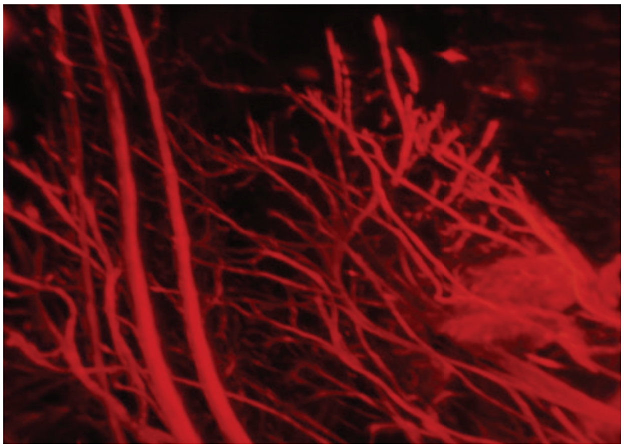
The Blood-Brain Barrier



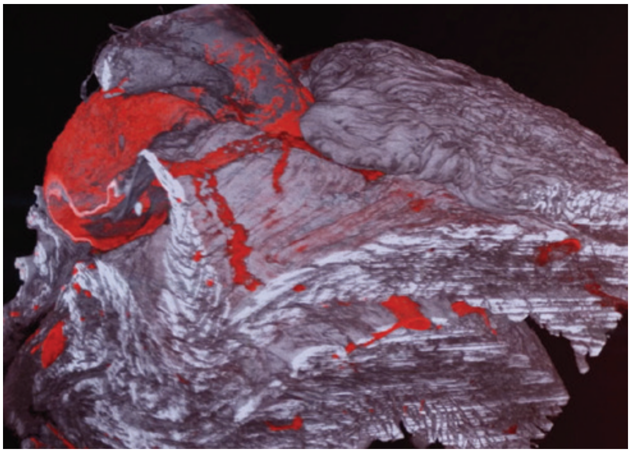
Orphan receptor GPR151 in the brain



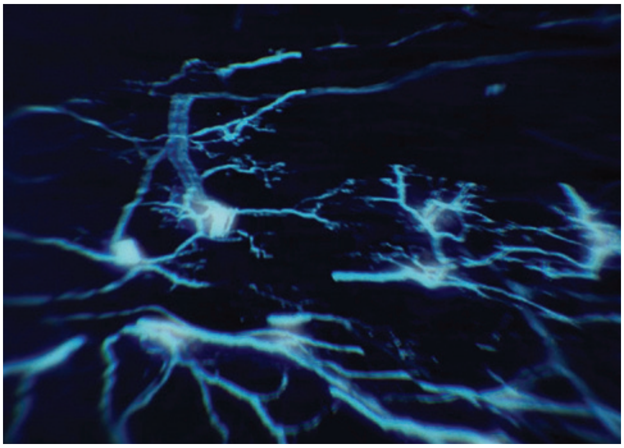
Wiring the nerves



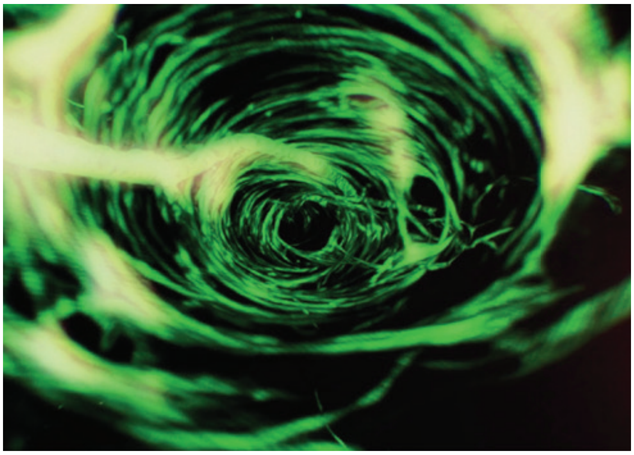
The strangled heart



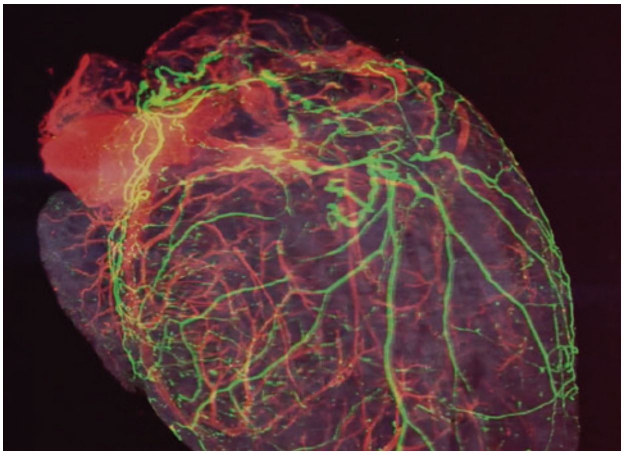
ALS and muscles



The fatty liver



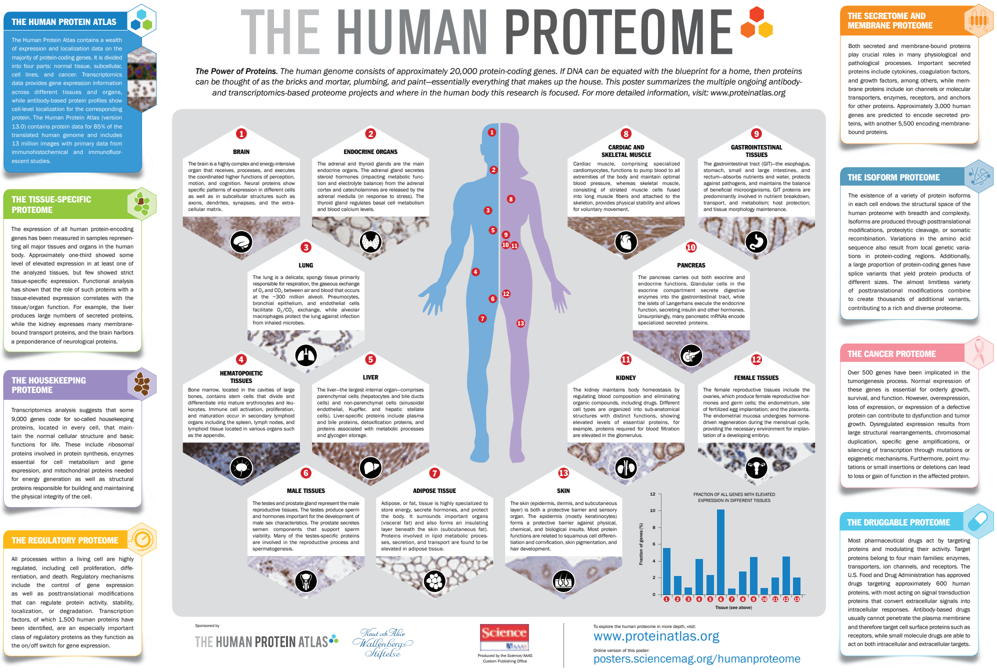
The nervous heart



HPA posters

Four posters have been produced in collaboration with *Science*/AAAS to show different aspects of the HPA and its content. You can access each poster using the URL provided.

Tissue Atlas poster (2015)



THE HUMAN PROTEOME

The Power of Proteins. *The human genome consists of approximately 20,000 protein-coding genes. If DNA can be equated with the blueprint for a home, then proteins can be thought of as the bricks and mortar, plumbing, and paint—essentially everything that makes up the house. This poster summarizes the multiple ongoing antibody- and transcriptomics-based proteome projects and where in the human body this research is focused. For more detailed information, visit: www.proteinatlas.org*

THE HUMAN PROTEIN ATLAS

The Human Protein Atlas contains a wealth of expression and localization data on the majority of protein-coding genes. It is divided into four parts: normal tissue, subcellular, cell lines, and cancer. Transcriptomics data provides gene expression information across different tissues and organs, while antibody-based protein profiles show cell-level localization for the corresponding protein. The Human Protein Atlas (version 13.0) contains protein data for 85% of the translated human genome and includes 13 million images with primary data from immunohistochemical and immunofluorescent studies.

THE TISSUE-SPECIFIC PROTEOME

The expression of all human protein-encoding genes has been measured in samples representing all major tissues and organs in the human body. Approximately one-third showed some level of elevated expression in at least one of the analyzed tissues, but few showed strict tissue-specific expression. Functional analysis has shown that the role of such proteins with a tissue-elevated expression correlates with the tissue/organ function. For example, the liver produces large numbers of secreted proteins, while the kidney expresses many membrane-bound transport proteins, and the brain harbors a preponderance of neurological proteins.

THE HOUSEKEEPING PROTEOME

Transcriptomics analysis suggests that some 9,000 genes code for so-called housekeeping proteins, located in every cell, that maintain the normal cellular structure and basic functions for life. These include ribosomal proteins involved in protein synthesis, enzymes essential for cell metabolism and gene expression, and mitochondrial proteins needed for energy generation as well as structural proteins responsible for building and maintaining the physical integrity of the cell.

THE REGULATORY PROTEOME

All processes within a living cell are highly regulated, including cell proliferation, differentiation, and death. Regulatory mechanisms include the control of gene expression as well as posttranslational modifications that can regulate protein activity, stability, localization, or degradation. Transcription factors, of which 1,500 human proteins have been identified, are an especially important class of regulatory proteins as they function as the on/off switch for gene expression.

1 BRAIN

The brain is a highly complex and energy-intensive organ that receives, processes, and executes the coordinated higher functions of perception, motion, and cognition. Neural proteins show specific patterns of expression in different cells as well as in subcellular structures such as axons, dendrites, synapses, and the extracellular matrix.

2 ENDOCRINE ORGANS

The adrenal and thyroid glands are the main endocrine organs. The adrenal gland secretes steroid hormones (impacting metabolic function and electrolyte balance) from the adrenal cortex and catecholamines are released by the adrenal medulla (in response to stress). The thyroid gland regulates basal cell metabolism and blood calcium levels.

3 LUNG

The lung is a delicate, spongy tissue primarily responsible for respiration, the gaseous exchange of O_2 and CO_2 between air and blood that occurs at the ~300 million alveoli. Pneumocytes, bronchial epithelium, and endothelial cells facilitate O_2/CO_2 exchange, while alveolar macrophages protect the lung against infection from inhaled microbes.

4 HEMATOPOIETIC TISSUES

Bone marrow, located in the cavities of large bones, contains stem cells that divide and differentiate into mature erythrocytes and leukocytes. Immune cell activation, proliferation, and maturation occur in secondary lymphoid organs including the spleen, lymph nodes, and lymphoid tissue located in various organs such as the appendix.

5 LIVER

The liver—the largest internal organ—comprises parenchymal cells (hepatocytes and bile ducts cells) and non-parenchymal cells (sinusoidal endothelial, Kupffer, and hepatic stellate cells). Liver-specific proteins include plasma and bile proteins, detoxification proteins, and proteins associated with metabolic processes and glycogen storage.

6 MALE TISSUES

The testes and prostate gland represent the male reproductive tissues. The testes produce sperm and hormones important for the development of male sex characteristics. The prostate secretes semen components that support sperm viability. Many of the testes-specific proteins are involved in the reproductive process and spermatogenesis.

7 ADIPOSE TISSUE

Adipose, or fat, tissue is highly specialized to store energy, secrete hormones, and protect the body. It surrounds important organs (visceral fat) and also forms an insulating layer beneath the skin (subcutaneous fat). Proteins involved in lipid metabolic processes, secretion, and transport are found to be elevated in adipose tissue.

8 CARDIAC AND SKELETAL MUSCLE

Cardiac muscle, comprising specialized cardiomyocytes, functions to pump blood to all extremities of the body and maintain optimal blood pressure, whereas skeletal muscle, consisting of striated muscle cells fused into long muscle fibers and attached to the skeleton, provides physical stability and allows for voluntary movement.

9 GASTROINTESTINAL TISSUES

The gastrointestinal tract (GIT)—the esophagus, stomach, small and large intestines, and rectum—absorbs nutrients and water, protects against pathogens, and maintains the balance of beneficial microorganisms. GIT proteins are predominantly involved in nutrient breakdown, transport, and metabolism; host protection; and tissue morphology maintenance.

10 PANCREAS

The pancreas carries out both exocrine and endocrine functions. Glandular cells in the exocrine compartment secrete digestive enzymes into the gastrointestinal tract, while the islets of Langerhans execute the endocrine function, secreting insulin and other hormones. Unsurprisingly, many pancreatic mRNAs encode specialized secreted proteins.

11 KIDNEY

The kidney maintains body homeostasis by regulating blood composition and eliminating organic compounds, including drugs. Different cell types are organized into sub-anatomical structures with distinct functions, showing elevated levels of essential proteins, for example, proteins required for blood filtration are elevated in the glomerulus.

12 FEMALE TISSUES

The female reproductive tissues include the ovaries, which produce female reproductive hormones and germ cells; the endometrium, site of fertilized egg implantation; and the placenta. The endometrial mucosa undergoes hormone-driven regeneration during the menstrual cycle, providing the necessary environment for implantation of a developing embryo.

13 SKIN

The skin (epidermis, dermis, and subcutaneous layer) is both a protective barrier and sensory organ. The epidermis (mostly keratinocytes) forms a protective barrier against physical, chemical, and biological insults. Most protein functions are related to squamous cell differentiation and cornification, skin pigmentation, and hair development.

THE SECRETOME AND MEMBRANE PROTEOME

Both secreted and membrane-bound proteins play crucial roles in many physiological and pathological processes. Important secreted proteins include cytokines, coagulation factors, and growth factors, among others, while membrane proteins include ion channels or molecular transporters, enzymes, receptors, and anchors for other proteins. Approximately 3,000 human genes are predicted to encode secreted proteins, with another 5,500 encoding membrane-bound proteins.

THE ISOFORM PROTEOME

The existence of a variety of protein isoforms in each cell endows the structural space of the human proteome with breadth and complexity. Isoforms are produced through posttranslational modifications, proteolytic cleavage, or somatic recombination. Variations in the amino acid sequence also result from local genetic variations in protein-coding regions. Additionally, a large proportion of protein-coding genes have splice variants that yield protein products of different sizes. The almost limitless variety of posttranslational modifications combine to create thousands of additional variants, contributing to a rich and diverse proteome.

THE CANCER PROTEOME

Over 500 genes have been implicated in the tumorigenesis process. Normal expression of these genes is essential for orderly growth, survival, and function. However, overexpression, loss of expression, or expression of a defective protein can contribute to dysfunction and tumor growth. Dysregulated expression results from large structural rearrangements, chromosomal duplication, specific gene amplifications, or silencing of transcription through mutations or epigenetic mechanisms. Furthermore, point mutations or small insertions or deletions can lead to loss or gain of function in the affected protein.

THE DRUGGABLE PROTEOME

Most pharmaceutical drugs act by targeting proteins and modulating their activity. Target proteins belong to four main families: enzymes, transporters, ion channels, and receptors. The U.S. Food and Drug Administration has approved drugs targeting approximately 600 human proteins, with most acting on signal transduction proteins that convert extracellular signals into intracellular responses. Antibody-based drugs usually cannot penetrate the plasma membrane and therefore target cell surface proteins such as receptors, while small molecule drugs are able to act on both intracellular and extracellular targets.

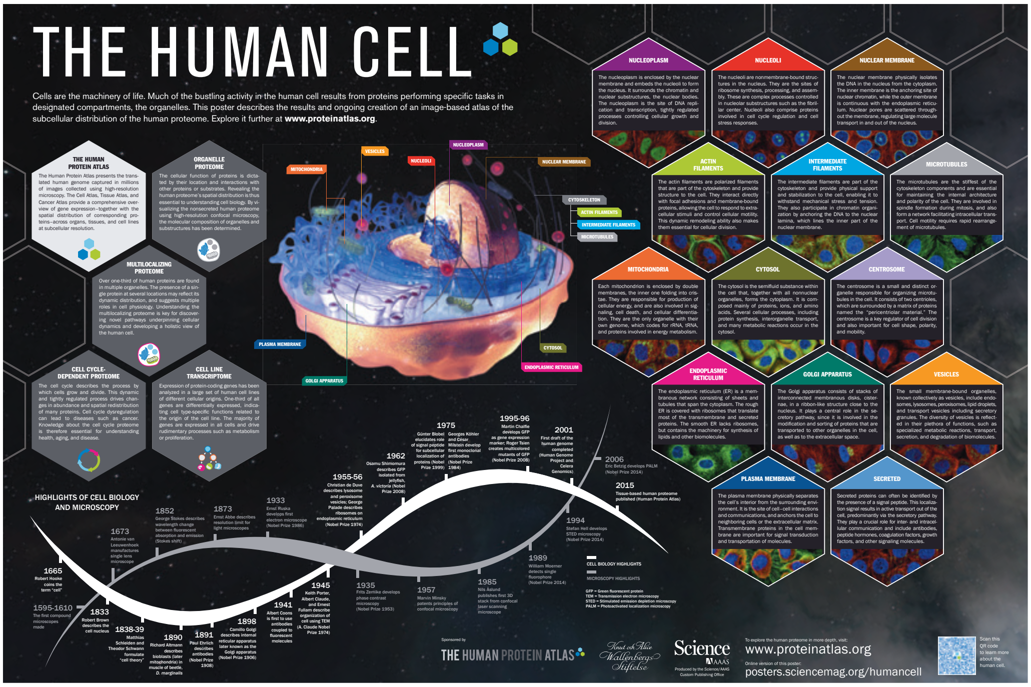
Sponsored by THE HUMAN PROTEIN ATLAS, Knut och Alice Wallenbergs Stiftelse

Science AAAS
Produced by the Science/AAAS Custom Publishing Office

To explore the human proteome in more depth, visit:
www.proteinatlas.org
Online version of this poster:
posters.sciencemag.org/humanproteome

Authors: Mathias Uhlén, Caroline Kampf, Fredrik Pontén
Link: www.proteinatlas.org/download/poster_proteome.pdf

Cell Atlas poster (2017)



THE HUMAN CELL

Cells are the machinery of life. Much of the bustling activity in the human cell results from proteins performing specific tasks in designated compartments, the organelles. This poster describes the results and ongoing creation of an image-based atlas of the subcellular distribution of the human proteome. Explore it further at **www.proteinatlas.org**.

THE HUMAN PROTEIN ATLAS

The Human Protein Atlas presents the translated human genome captured in millions of images collected using high-resolution microscopy. The Cell Atlas, Tissue Atlas, and Cancer Atlas provide a comprehensive overview of gene expression–together with the spatial distribution of corresponding proteins–across organs, tissues, and cell lines at subcellular resolution.

ORGANELLE PROTEOME

The cellular function of proteins is dictated by their location and interactions with other proteins or substrates. Revealing the human proteome's spatial distribution is thus essential to understanding cell biology. By visualizing the nonsecreted human proteome using high-resolution confocal microscopy, the molecular composition of organelles and substructures has been determined.

MULTILOCALIZING PROTEOME

Over one-third of human proteins are found in multiple organelles. The presence of a single protein at several locations may reflect its dynamic distribution, and suggests multiple roles in cell physiology. Understanding the multilocalizing proteome is key for discovering novel pathways underpinning cellular dynamics and developing a holistic view of the human cell.

CELL CYCLE-DEPENDENT PROTEOME

The cell cycle describes the process by which cells grow and divide. This dynamic and tightly regulated process drives changes in abundance and spatial redistribution of many proteins. Cell cycle dysregulation can lead to diseases such as cancer. Knowledge about the cell cycle proteome is therefore essential for understanding health, aging, and disease.

CELL LINE TRANSCRIPTOME

Expression of protein-coding genes has been analyzed in a large set of human cell lines of different cellular origins. One-third of all genes are differentially expressed, indicating cell type-specific functions related to the origin of the cell line. The majority of genes are expressed in all cells and drive rudimentary processes such as metabolism or proliferation.

NUCLEOPLASM

The nucleoplasm is enclosed by the nuclear membrane and embeds the nucleoli to form the nucleus. It surrounds the chromatin and nuclear substructures, the nuclear bodies. The nucleoplasm is the site of DNA replication and transcription, tightly regulated processes controlling cellular growth and division.

NUCLEOLI

The nucleoli are nonmembrane-bound structures in the nucleus. They are the sites of ribosome synthesis, processing, and assembly. These are complex processes controlled in nucleolar substructures such as the fibrillar center. Nucleoli also comprise proteins involved in cell cycle regulation and cell stress responses.

NUCLEAR MEMBRANE

The nuclear membrane physically isolates the DNA in the nucleus from the cytoplasm. The inner membrane is the anchoring site of nuclear chromatin, while the outer membrane is continuous with the endoplasmic reticulum. Nuclear pores are scattered throughout the membrane, regulating large molecule transport in and out of the nucleus.

ACTIN FILAMENTS

The actin filaments are polarized filaments that are part of the cytoskeleton and provide structure to the cell. They interact directly with focal adhesions and membrane-bound proteins, allowing the cell to respond to extracellular stimuli and control cellular motility. This dynamic remodeling ability also makes them essential for cellular division.

INTERMEDIATE FILAMENTS

The intermediate filaments are part of the cytoskeleton and provide physical support and stabilization to the cell, enabling it to withstand mechanical stress and tension. They also participate in chromatin organization by anchoring the DNA to the nuclear lamina, which lines the inner part of the nuclear membrane.

MICROTUBULES

The microtubules are the stiffest of the cytoskeleton components and are essential for maintaining the internal architecture and polarity of the cell. They are involved in spindle formation during mitosis, and also form a network facilitating intracellular transport. Cell motility requires rapid rearrangement of microtubules.

MITOCHONDRIA

Each mitochondrion is enclosed by double membranes, the inner one folding into cristae. They are responsible for production of cellular energy, and are also involved in signaling, cell death, and cellular differentiation. They are the only organelle with their own genome, which codes for rRNA, tRNA, and proteins involved in energy metabolism.

CYTOSOL

The cytosol is the semifluid substance within the cell that, together with all nonnuclear organelles, forms the cytoplasm. It is composed mainly of proteins, ions, and amino acids. Several cellular processes, including protein synthesis, interorganelle transport, and many metabolic reactions occur in the cytosol.

CENTROSOME

The centrosome is a small and distinct organelle responsible for organizing microtubules in the cell. It consists of two centrioles, which are surrounded by a matrix of proteins named the "pericentriolar material." The centrosome is a key regulator of cell division and also important for cell shape, polarity, and mobility.

ENDOPLASMIC RETICULUM

The endoplasmic reticulum (ER) is a membranous network consisting of sheets and tubules that span the cytoplasm. The rough ER is covered with ribosomes that translate most of the transmembrane and secreted proteins. The smooth ER lacks ribosomes, but contains the machinery for synthesis of lipids and other biomolecules.

GOLGI APPARATUS

The Golgi apparatus consists of stacks of interconnected membranous disks, cisternae, in a ribbon-like structure close to the nucleus. It plays a central role in the secretory pathway, since it is involved in the modification and sorting of proteins that are transported to other organelles in the cell, as well as to the extracellular space.

VESICLES

The small membrane-bound organelles, known collectively as vesicles, include endosomes, lysosomes, peroxisomes, lipid droplets, and transport vesicles including secretory granules. The diversity of vesicles is reflected in their plethora of functions, such as specialized metabolic reactions, transport, secretion, and degradation of biomolecules.

PLASMA MEMBRANE

The plasma membrane physically separates the cell's interior from the surrounding environment. It is the site of cell–cell interactions and communications, and anchors the cell to neighboring cells or the extracellular matrix. Transmembrane proteins in the cell membrane are important for signal transduction and transportation of molecules.

SECRETED

Secreted proteins can often be identified by the presence of a signal peptide. This localization signal results in active transport out of the cell, predominantly via the secretory pathway. They play a crucial role for inter- and intracellular communication and include antibodies, peptide hormones, coagulation factors, growth factors, and other signaling molecules.

HIGHLIGHTS OF CELL BIOLOGY AND MICROSCOPY

- 1595-1610 The first compound microscopes made
- 1665 Robert Hooke coins the term "cell"
- 1673 Antonie van Leeuwenhoek manufactures single lens microscope
- 1833 Robert Brown describes the cell nucleus
- 1838-39 Matthias Schleiden and Theodor Schwann formulate "cell theory"
- 1852 George Stokes describes wavelength change between fluorescent absorption and emission (Stokes shift)
- 1873 Ernst Abbe describes resolution limit for light microscopes
- 1890 Richard Altmann describes bioblasts (later mitochondria) in muscle of beetle, *D. marginalis*
- 1891 Paul Ehrlich describes antibodies (Nobel Prize 1908)
- 1898 Camillo Golgi describes internal reticular apparatus later known as the Golgi apparatus (Nobel Prize 1906)
- 1933 Ernst Ruska develops first electron microscope (Nobel Prize 1986)
- 1935 Frits Zernike develops phase contrast microscopy (Nobel Prize 1953)
- 1941 Albert Coons is first to use antibodies coupled to fluorescent molecules
- 1945 Keith Porter, Albert Claude, and Ernest Fullam describe organization of cell using TEM (A. Claude Nobel Prize 1974)
- 1955-56 Christian de Duve describes lysosome and peroxisome vesicles; George Palade describes ribosomes on endoplasmic reticulum (Nobel Prize 1974)
- 1957 Marvin Minsky patents principles of confocal microscopy
- 1962 Osamu Shimomura describes GFP isolated from jellyfish, *A. victoria* (Nobel Prize 2008)
- 1975 Günter Blobel elucidates role of signal peptide for subcellular localization of proteins (Nobel Prize 1999) Georges Köhler and César Milstein develop first monoclonal antibodies (Nobel Prize 1984)
- 1985 Nils Åslund publishes first 3D stack from confocal laser scanning microscope
- 1989 William Moerner detects single fluorophore (Nobel Prize 2014)
- 1994 Stefan Hell develops STED microscopy (Nobel Prize 2014)
- 1995-96 Martin Chalfie develops GFP as gene expression marker; Roger Tsien creates multicolored mutants of GFP (Nobel Prize 2008)
- 2001 First draft of the human genome completed (Human Genome Project and Celera Genomics)
- 2006 Eric Betzig develops PALM (Nobel Prize 2014)
- 2015 Tissue-based human proteome published (Human Protein Atlas)

CELL BIOLOGY HIGHLIGHTS
MICROSCOPY HIGHLIGHTS

GFP = Green fluorescent protein
TEM = Transmission electron microscopy
STED = Stimulated emission depletion microscopy
PALM = Photoactivated localization microscopy

Sponsored by THE HUMAN PROTEIN ATLAS, Knut och Alice Wallenbergs Stiftelse

Science AAAS
Produced by the Science/AAAS Custom Publishing Office

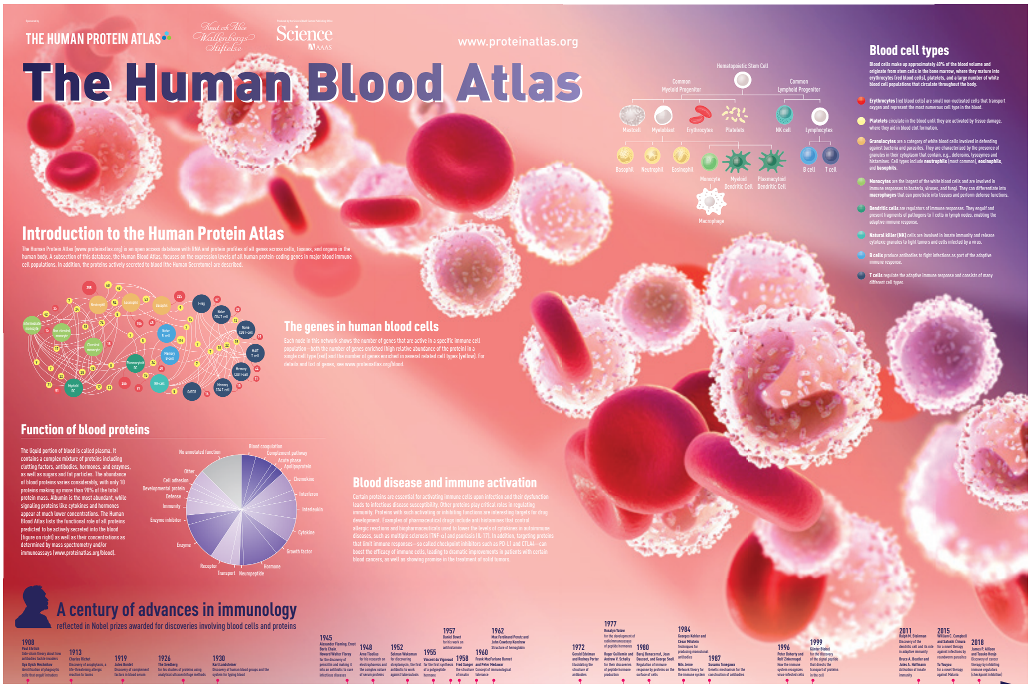
To explore the human proteome in more depth, visit:
www.proteinatlas.org
Online version of this poster:
posters.sciencemag.org/humancell

Scan this QR code to learn more about the human cell.

Authors: Mikaela Wiking, Tove Alm, Emma Lundberg
Link: www.proteinatlas.org/download/poster_cell.pdf

HPA posters

Blood Atlas poster (2019)



Sponsored by THE HUMAN PROTEIN ATLAS

Knut och Alice Wallenbergs Stiftelse

Produced by the Science/AAAS Custom Publishing Office

Science AAAS

www.proteinatlas.org

The Human Blood Atlas

Introduction to the Human Protein Atlas

The Human Protein Atlas (www.proteinatlas.org) is an open access database with RNA and protein profiles of all genes across cells, tissues, and organs in the human body. A subsection of this database, the Human Blood Atlas, focuses on the expression levels of all human protein-coding genes in major blood immune cell populations. In addition, the proteins actively secreted to blood (the Human Secretome) are described.

The genes in human blood cells

Each node in this network shows the number of genes that are active in a specific immune cell population—both the number of genes enriched (high relative abundance of the protein) in a single cell type (red) and the number of genes enriched in several related cell types (yellow). For details and list of genes, see www.proteinatlas.org/blood.

Function of blood proteins

The liquid portion of blood is called plasma. It contains a complex mixture of proteins including clotting factors, antibodies, hormones, and enzymes, as well as sugars and fat particles. The abundance of blood proteins varies considerably, with only 10 proteins making up more than 90% of the total protein mass. Albumin is the most abundant, while signaling proteins like cytokines and hormones appear at much lower concentrations. The Human Blood Atlas lists the functional role of all proteins predicted to be actively secreted into the blood (figure on right) as well as their concentrations as determined by mass spectrometry and/or immunoassays (www.proteinatlas.org/blood).

Blood disease and immune activation

Certain proteins are essential for activating immune cells upon infection and their dysfunction leads to infectious disease susceptibility. Other proteins play critical roles in regulating immunity. Proteins with such activating or inhibiting functions are interesting targets for drug development. Examples of pharmaceutical drugs include anti histamines that control allergic reactions and biopharmaceuticals used to lower the levels of cytokines in autoimmune diseases, such as multiple sclerosis (TNF-α) and psoriasis (IL-17). In addition, targeting proteins that limit immune responses—so called checkpoint inhibitors such as PD-L1 and CTLA4—can boost the efficacy of immune cells, leading to dramatic improvements in patients with certain blood cancers, as well as showing promise in the treatment of solid tumors.

Blood cell types

Blood cells make up approximately 40% of the blood volume and originate from stem cells in the bone marrow, where they mature into erythrocytes (red blood cells), platelets, and a large number of white blood cell populations that circulate throughout the body.

- **Erythrocytes** (red blood cells) are small non-nucleated cells that transport oxygen and represent the most numerous cell type in the blood.
- **Platelets** circulate in the blood until they are activated by tissue damage, where they aid in blood clot formation.
- **Granulocytes** are a category of white blood cells involved in defending against bacteria and parasites. They are characterized by the presence of granules in their cytoplasm that contain, e.g., defensins, lysozymes and histamines. Cell types include **neutrophils** (most common), **eosinophils**, and **basophils**.
- **Monocytes** are the largest of the white blood cells and are involved in immune responses to bacteria, viruses, and fungi. They can differentiate into **macrophages** that can penetrate into tissues and perform defense functions.
- **Dendritic cells** are regulators of immune responses. They engulf and present fragments of pathogens to T cells in lymph nodes, enabling the adaptive immune response.
- **Natural killer (NK)** cells are involved in innate immunity and release cytotoxic granules to fight tumors and cells infected by a virus.
- **B cells** produce antibodies to fight infections as part of the adaptive immune response.
- **T cells** regulate the adaptive immune response and consists of many different cell types.

A century of advances in immunology

reflected in Nobel prizes awarded for discoveries involving blood cells and proteins

1908 Paul Ehrlich Side-chain theory about how antibodies tackle invaders; Ilya Ilyich Mechnikov Identification of phagocytic cells that engulf intruders

1913 Charles Richet Discovery of anaphylaxis, a life-threatening allergic reaction to toxins

1919 Jules Bordet Discovery of complement factors in blood serum

1926 The Svedberg for his studies of proteins using analytical ultracentrifuge methods

1930 Karl Landsteiner Discovery of human blood groups and the system for typing blood

1945 Alexander Fleming, Ernst Boris Chain, Howard Walter Florey for the discovery of penicillin and making it into an antibiotic to cure infectious diseases

1948 Arne Tiselius for his research on electrophoresis and the complex nature of serum proteins

1952 Selman Waksman for discovering streptomycin, the first antibiotic to work against tuberculosis

1955 Vincent du Vigneaud for the first synthesis of a polypeptide hormone

1957 Daniel Bovet for his work on antihistamine

1958 Fred Sanger the structure of insulin

1960 Frank MacFarlane Burnet and Peter Medawar Concept of immunological tolerance

1962 Max Ferdinand Perutz and John Cowdery Kendrew Structure of hemoglobin

1972 Gerald Edelman and Rodney Porter Elucidating the structure of antibodies

1977 Rosalyn Yalow for the development of radioimmunoassays of peptide hormones; Roger Guillemin and Andrew V. Schally for their discoveries of peptide hormone production

1980 Baruj Benacerraf, Jean Dausset, and George Snell Regulation of immune response by proteins on the surface of cells

1984 Georges Kohler and César Milstein Techniques for producing monoclonal antibodies; Nils Jerne Network theory for the immune system

1987 Susumu Tonegawa Genetic mechanism for the construction of antibodies

1996 Peter Doherty and Rolf Zinkernagel How the immune system recognizes virus-infected cells

1999 Günter Blobel for the discovery of the signal peptide that directs the transport of proteins in the cell

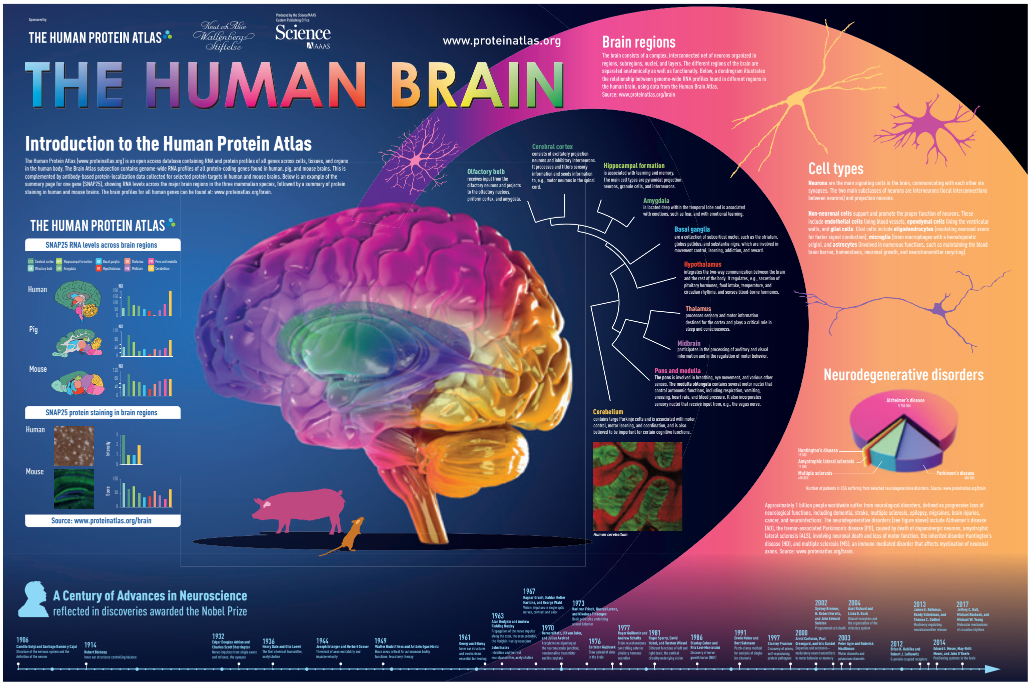
2011 Ralph M. Steinman Discovery of the dendritic cell and its role in adaptive immunity; Bruce A. Beutler and Jules A. Hoffmann Activation of innate immunity

2015 William C. Campbell and Satoshi Ōmura for a novel therapy against infections by roundworm parasites; Tu Youyou for a novel therapy against Malaria

2018 James P. Allison and Tasuku Honjo Discovery of cancer therapy by inhibiting immune regulators (checkpoint inhibition)

Authors: Åsa Sivertsson, Mathias Uhlén
Illustration/Design: Mattias Karlén
Link: www.proteinatlas.org/download/poster_blood.pdf

Brain Atlas poster (2020)



Sponsored by THE HUMAN PROTEIN ATLAS

Knut och Alice Wallenbergs Stiftelse

Produced by the Science/AAAS Custom Publishing Office

Science AAAS

www.proteinatlas.org

THE HUMAN BRAIN

Introduction to the Human Protein Atlas

The Human Protein Atlas (www.proteinatlas.org) is an open access database containing RNA and protein profiles of all genes across cells, tissues, and organs in the human body. The Brain Atlas subsection contains genome-wide RNA profiles of all protein-coding genes found in human, pig, and mouse brains. This is complemented by antibody-based protein-localization data collected for selected protein targets in human and mouse brains. Below is an example of the summary page for one gene (SNAP25), showing RNA levels across the major brain regions in the three mammalian species, followed by a summary of protein staining in human and mouse brains. The brain profiles for all human genes can be found at: www.proteinatlas.org/brain.

Brain regions

The brain consists of a complex, interconnected net of neurons organized in regions, subregions, nuclei, and layers. The different regions of the brain are separated anatomically as well as functionally. Below, a dendrogram illustrates the relationship between genome-wide RNA profiles found in different regions in the human brain, using data from the Human Brain Atlas.
Source: www.proteinatlas.org/brain

Olfactory bulb receives input from the olfactory neurons and projects to the olfactory nucleus, piriform cortex, and amygdala.

Cerebral cortex consists of excitatory projection neurons and inhibitory interneurons. It processes and filters sensory information and sends information to, e.g., motor neurons in the spinal cord.

Hippocampal formation is associated with learning and memory. The main cell types are pyramidal projection neurons, granule cells, and interneurons.

Amygdala is located deep within the temporal lobe and is associated with emotions, such as fear, and with emotional learning.

Basal ganglia are a collection of subcortical nuclei, such as the striatum, globus pallidus, and substantia nigra, which are involved in movement control, learning, addiction, and reward.

Hypothalamus integrates the two-way communication between the brain and the rest of the body. It regulates, e.g., secretion of pituitary hormones, food intake, temperature, and circadian rhythms, and senses blood-borne hormones.

Thalamus processes sensory and motor information destined for the cortex and plays a critical role in sleep and consciousness.

Midbrain participates in the processing of auditory and visual information and in the regulation of motor behavior.

Pons and medulla The pons is involved in breathing, eye movement, and various other senses. The medulla oblongata contains several motor nuclei that control autonomic functions, including respiration, vomiting, sneezing, heart rate, and blood pressure. It also incorporates sensory nuclei that receive input from, e.g., the vagus nerve.

Cerebellum contains large Purkinje cells and is associated with motor control, motor learning, and coordination, and is also believed to be important for certain cognitive functions.

Human cerebellum

Cell types

Neurons are the main signaling units in the brain, communicating with each other via synapses. The two main subclasses of neurons are interneurons (local interconnections between neurons) and projection neurons.

Non-neuronal cells support and promote the proper function of neurons. These include **endothelial cells** lining blood vessels, **ependymal cells** lining the ventricular walls, and **glial cells**. Glial cells include **oligodendrocytes** (insulating neuronal axons for faster signal conduction), **microglia** (brain macrophages with a hematopoietic origin), and **astrocytes** (involved in numerous functions, such as maintaining the blood brain barrier, homeostasis, neuronal growth, and neurotransmitter recycling).

Neurodegenerative disorders

Alzheimer's disease 5 700 000
Huntington's disease 15 000
Amyotrophic lateral sclerosis 17 000
Multiple sclerosis 490 000
Parkinson's disease 800 000

Number of patients in USA suffering from selected neurodegenerative disorders. Source: www.proteinatlas.org/brain

Approximately 1 billion people worldwide suffer from neurological disorders, defined as progressive loss of neurological functions, including dementia, stroke, multiple sclerosis, epilepsy, migraines, brain injuries, cancer, and neuroinfections. The neurodegenerative disorders (see figure above) include Alzheimer's disease (AD), the tremor-associated Parkinson's disease (PD), caused by death of dopaminergic neurons, amyotrophic lateral sclerosis (ALS), involving neuronal death and loss of motor function, the inherited disorder Huntington's disease (HD), and multiple sclerosis (MS), an immune-mediated disorder that affects myelination of neuronal axons. Source: www.proteinatlas.org/brain.

A Century of Advances in Neuroscience

reflected in discoveries awarded the Nobel Prize

1906 Camillo Golgi and Santiago Ramón y Cajal Structure of the nervous system and the definition of the neuron

1914 Robert Bárány Inner ear structures controlling balance

1932 Edgar Douglas Adrian and Charles Scott Sherrington Nerve impulses from single axons and reflexes; the synapse

1936 Henry Dale and Otto Loewi The first chemical transmitter, acetylcholine

1944 Joseph Erlanger and Herbert Gasser Threshold of axon excitability and impulse velocity

1949 Walter Rudolf Hess and António Egas Moniz Brain areas critical for autonomous bodily functions; leucotomy therapy

1961 Georg von Békésy Inner ear structures and mechanisms essential for hearing

1963 Alan Hodgkin and Andrew Fielding Huxley Propagation of the nerve impulse along the axon, the axon potential, the Hodgkin-Huxley equations; John Eccles Inhibition and the first neurotransmitter, acetylcholine

1967 Ragnar Granit, Haldan Keffer Hartline, and George Wald Vision: impulses in single optic nerves, contrast and color

1970 Bernard Katz, Ulf von Euler, and Julius Axelrod Acetylcholine signaling at the neuromuscular junction; noradrenaline transmitter and its reuptake

1973 Karl von Frisch, Konrad Lorenz, and Nikolaas Tinbergen Basic principles underlying animal behavior

1976 Carleton Gajdusek Slow spread of virus in the brain

1977 Roger Guillemin and Andrew Schally Brain neurohormones controlling anterior pituitary hormone secretion

1981 Roger Sperry, David Hubel, and Torsten Wiesel Different functions of left and right brain; the cortical circuitry underlying vision

1986 Stanley Cohen and Rita Levi-Montalcini Discovery of nerve growth factor (NGF)

1991 Erwin Neher and Bert Sakmann Patch-clamp method for analysis of single-ion channels

1997 Stanley Prusiner Discovery of prions, self-reproducing protein pathogens

2000 Arvid Carlsson, Paul Greengard, and Eric Kandel Dopamine and serotonin—modulatory neurotransmitters in motor behavior or memory

2002 Sydney Brenner, H. Robert Horvitz, and John Edward Sulston Programmed cell death

2003 Peter Agre and Roderick MacKinnon Water channels and potassium channels

2004 Axel Richard and Linda B. Buck Odorant receptors and the organization of the olfactory system

2012 Brian K. Kobilka and Robert J. Lefkowitz G-protein coupled receptors

2013 James E. Rothman, Randy Schekman, and Thomas C. Südhof Machinery regulating neurotransmitter release

2014 Edvard I. Moser, May-Britt Moser, and John O'Keefe Positioning systems in the brain

2017 Jeffrey C. Hall, Michael Rosbash, and Michael W. Young Molecular mechanisms of circadian rhythms

Authors: Evelina Sjöstedt, Jan Mulder, Tomas Hökfelt, Mathias Uhlén
Illustration/Design: Mattias Karlén
Link: www.proteinatlas.org/download/poster_brain.pdf

RESEARCH ARTICLE SUMMARY

PROTEOMICS

Tissue-based map of the human proteome

Mathias Uhlén,* Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, IngMarie Olsson, Karolina Edlund, Emma Lundberg, Sanjay Navani, Cristina Al-Khalili Szigyarto, Jacob Odeberg, Dijana Djureinovic, Jenny Ottosson Takanen, Sophia Hober, Tove Alm, Per-Henrik Edqvist, Holger Berling, Hanna Tegel, Jan Mulder, Johan Rockberg, Peter Nilsson, Jochen M. Schwenk, Marica Hamsten, Kalle von Feilitzen, Mattias Forsberg, Lukas Persson, Fredric Johansson, Martin Zwahlen, Gunnar von Heijne, Jens Nielsen, Fredrik Pontén

INTRODUCTION: Resolving the molecular details of proteome variation in the different tissues and organs of the human body would greatly increase our knowledge of human biology and disease. Here, we present a map of the human tissue proteome based on quantitative transcriptomics on a tissue and organ level combined with protein profiling using microarray-based immunohistochemistry to achieve spatial localization of proteins down to the single-cell level. We provide a global analysis of the secreted and membrane proteins, as well as an analysis of the expression profiles for all proteins targeted by pharmaceutical drugs and proteins implicated in cancer.

RATIONALE: We have used an integrative omics approach to study the spatial human proteome. Samples representing all major tissues and organs (n = 44) in the human body have been analyzed based on 24,028 antibodies corresponding to 16,975 protein-encoding genes, complemented with RNA-sequencing data for 32 of the tissues. The antibodies have been used to produce more than 13 million tissue-based immunohistochemistry images, each annotated by pathologists for all sampled tissues. To facilitate integration with other biological resources, all data are available for download and cross-referencing.

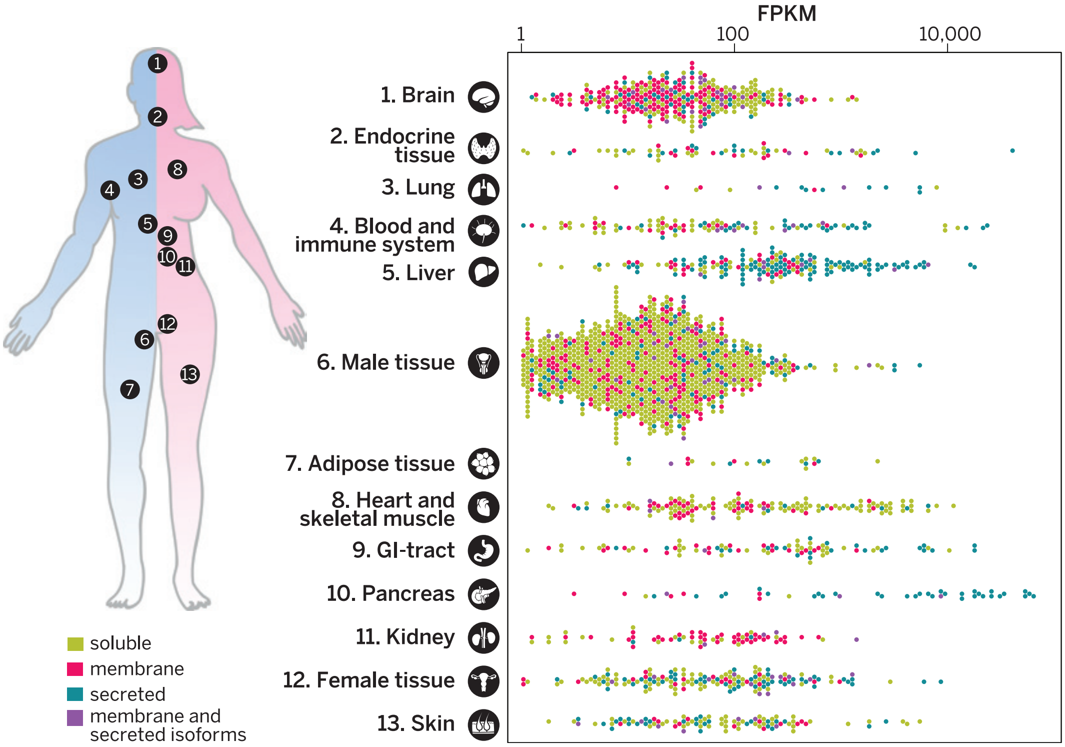
RESULTS: We report a genome-wide analysis of the tissue specificity of RNA and protein expression covering more than 90% of the putative protein-coding genes, complemented with analyses of various subproteomes, such as predicted secreted proteins (n = 3171) and membrane-bound proteins (n = 5570). The analysis shows that almost half of the genes are expressed in all analyzed tissues, which suggests that the gene products are needed in all cells to maintain "housekeeping" functions such as cell growth, energy generation, and basic metabolism. Furthermore, there is enrichment in metabolism among these genes, as 60% of all metabolic enzymes are expressed in all analyzed tissues. The largest number of tissue-enriched genes is found in the testis, followed by the brain and the liver. Analysis of the 618 proteins targeted by clinically approved drugs unexpectedly showed that 30% are expressed in all analyzed tissues. An analysis of metabolic activity based on genome-scale metabolic models (GEMS) revealed liver as the most metabolically active tissue, followed by adipose tissue and skeletal muscle.

CONCLUSIONS: A freely available interactive resource is presented as part of the Human Protein Atlas portal (www.proteinatlas.org), offering the possibility to explore the tissue-elevated proteomes in tissues and organs and to analyze tissue profiles for specific protein classes. Comprehensive lists of proteins expressed at elevated levels in the different tissues have been compiled to provide a spatial context with localization of the proteins in the subcompartments of each tissue and organ down to the single-cell level. ■

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.1260419

The list of author affiliations is available in the full article online.
*Corresponding author. E-mail: mathias.uhlen@scilifelab.se
Cite this article as M. Uhlén *et al.*, *Science* **347**, 1260419 (2015). DOI: 10.1126/science.1260419



The human tissue–enriched proteins. All tissue-enriched proteins are shown for 13 representative tissues or groups of tissues, stratified according to their predicted subcellular localization. Enriched proteins are mainly intracellular in testis, mainly membrane bound in brain and kidney, and mainly secreted in pancreas and liver.

Originally published 23 January 2015

RESEARCH ARTICLE

PROTEOMICS

Tissue-based map of the human proteome

Mathias Uhlén,[1,2,3]* Linn Fagerberg,[1] Björn M. Hallström,[1,2] Cecilia Lindskog,[4] Per Oksvold,[1] Adil Mardinoglu,[5] Åsa Sivertsson,[1] Caroline Kampf,[4] Evelina Sjöstedt,[1,4] Anna Asplund,[4] IngMarie Olsson,[4] Karolina Edlund,[6] Emma Lundberg,[1] Sanjay Navani,[7] Cristina Al-Khalili Szigyarto,[2] Jacob Odeberg,[1] Dijana Djureinovic,[4] Jenny Ottosson Takanen,[2] Sophia Hober,[2] Tove Alm,[1] Per-Henrik Edqvist,[4] Holger Berling,[2] Hanna Tegel,[2] Jan Mulder,[8] Johan Rockberg,[2] Peter Nilsson,[1] Jochen M. Schwenk,[1] Marica Hamsten,[2] Kalle von Feilitzen,[1] Mattias Forsberg,[1] Lukas Persson,[1] Fredric Johansson,[1] Martin Zwahlen,[1] Gunnar von Heijne,[9] Jens Nielsen,[3,5] Fredrik Pontén[4]

Resolving the molecular details of proteome variation in the different tissues and organs of the human body will greatly increase our knowledge of human biology and disease. Here, we present a map of the human tissue proteome based on an integrated omics approach that involves quantitative transcriptomics at the tissue and organ level, combined with tissue microarray–based immunohistochemistry, to achieve spatial localization of proteins down to the single-cell level. Our tissue-based analysis detected more than 90% of the putative protein-coding genes. We used this approach to explore the human secretome, the membrane proteome, the druggable proteome, the cancer proteome, and the metabolic functions in 32 different tissues and organs. All the data are integrated in an interactive Web-based database that allows exploration of individual proteins, as well as navigation of global expression patterns, in all major tissues and organs in the human body.

There is much interest in annotating all human genes at the level of DNA (*1*, *2*), RNA (*3*, *4*), and proteins (*5*, *6*), with the ultimate goal of defining structure, function, localization, expression, and interactions of all proteins. This has resulted in large-scale projects, such as ENCODE (*7*) and the Human Proteome Project (*8*), aimed to integrate results from many research groups and technical platforms to reach a detailed understanding of each of the ~20,000 human protein-coding genes predicted from the human genome and their corresponding protein isoforms. Recently, drafts of the human proteome based on proteogenomics efforts have been described (*9*, *10*), focusing on recent advances in mass spectrometry that allow comprehensive analyses using both isotope-labeled analysis systems (*11*) and deep proteomics methods (*12*) or genome-wide targeted proteomics efforts (*13*).

A complement to these efforts is the Human Protein Atlas program (*14*), which is exploring the human proteome using genecentric and genome-wide antibody-based profiling on tissue microarrays. This allows for spatial pathology-based annotation of protein expression, in combination with deep-sequencing transcriptomics of the same tissue types. The strategy is based on the quantitative assessment of transcript expression in complex tissue homogenates, involving a mixture of cell types combined with the precise localization of the corresponding proteins down to the single-cell level, using immunohistochemistry. Recently, we performed a transcriptomics study of 27 different tissues using this approach (*15*), followed by subsequent in-depth studies of the global proteome in a number of these tissues and organs, such as liver (*16*), testis (*17*), and the gastrointestinal (GI) tract (*18*). Here, we have used this approach and extended the analysis to 32 tissue types, representing all major tissues and organs in the human body, to create a genome-wide map of the human tissue–based proteome, with a focus on the analysis of the tissue-elevated proteins and all secreted and membrane proteins. Particular emphasis has been placed on analyses of proteins targeted by pharmaceutical drugs (*19*) and proteins implicated in cancer (*20*). We used the data to generate comprehensive metabolic maps for all 32 tissue types in order to identify differences in metabolism between tissues. In addition, new transcriptomics data from 36 human cell lines allowed us to compare the proteomes between cell lines and normal cells derived from the same tissue types. Finally, the protein isoforms generated by differential splicing between different tissues were studied with a focus on splice variants with predicted differential subcellular localization. All data are presented in an interactive database (www.proteinatlas.org).

Results

Classification of all human protein-coding genes

Samples representing all major tissues and organs (n = 44) in the human body were analyzed (Fig. 1A) by using 20,456 antibodies generated "in-house," as well as 3572 antibodies provided by external suppliers. The antibodies have been used to produce more than 13 million tissue-based immunohistochemistry images, with each image annotated on the single-cell level for all sampled tissues by pathologists. The analysis was complemented with RNA sequencing (RNAseq) data for 32 out of the 44 tissue types. We investigated global expression profiles using hierarchical clustering based on the correlation between 122 biological replicates from the 32 organs and tissues (Fig. 1B and fig. S1). The results reveal testis and brain as outliers and a clear connectivity between the samples from the GI tract (stomach, duodenum, small intestine, colon, and rectum), the hematopoietic tissues (bone marrow, lymph node, spleen, tonsil, and appendix) and the two striated muscle samples (cardiac and skeletal muscle). A principal component analysis (fig. S2A) confirms a close resemblance between cardiac and skeletal muscle but also suggests similarities in global expression between pancreas and salivary gland, as well as differences between the primary lymphoid tissue (bone marrow) and the secondary lymphoid tissues, such as tonsil and spleen.

The transcriptomics study allowed us to refine the classification performed earlier (*15*) of all the 20,344 putative protein-coding genes with RNA-seq data into categories based on their expression across all 32 tissue types (Fig. 1C, Table 1, and tables S1 to S4). Indirectly, this also provides an estimate of the relative protein levels corresponding to each gene, because proteogenomics analyses have shown that the translation rate, in most cases, is constant for a specific transcript across different human cells and tissues at both a cellular level (*21*) and a tissue level (*9*). Although it is still a matter of scientific debate (*22*) whether protein degradation rates could, in some cases, vary for an individual protein in different tissues, an overall concurrence between mRNA and protein levels for a given gene product across various tissues is generally expected (*9*, *21*). A large fraction (44%) of the protein-coding genes were detected in all analyzed tissues, and these ubiquitously expressed genes include known "housekeeping" genes encoding mitochondrial proteins and proteins involved in overall cell structure, translation, transcription, and replication. Of all the protein coding genes, 34% showed an elevated expression in at least one of the analyzed tissues,

[1]Science for Life Laboratory, KTH—Royal Institute of Technology, SE-171 21 Stockholm, Sweden. [2]Department of Proteomics, KTH—Royal Institute of Technology, SE-106 91 Stockholm, Sweden. [3]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2970 Hørsholm, Denmark. [4]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden· [5]Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. [6]Leibniz Research Centre for Working Environment and Human Factors (IfADo) at Dortmund TU, D-44139 Dortmund, Germany. [7]Lab Surgpath, Mumbai, India. [8]Science for Life Laboratory, Department of Neuroscience, Karolinska Institute, SE-171 77 Stockholm, Sweden. [9]Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden.
*Corresponding author. E-mail: mathias.uhlen@scilifelab.se

Originally published 23 January 2015

and these were further subdivided into (i) enriched genes with mRNA levels in one tissue type at least five times the maximum levels of all other analyzed tissues, (ii) group-enriched genes with enriched expression in a small number of tissues, and (iii) enhanced genes with only a moderately elevated expression. The use of the word "tissue-specific" has been avoided because this definition depends on arbitrary cut-off levels, and many proteins described in the literature as "tissue-specific" are here shown to be expressed in several tissues. This is exemplified by albumin, which we, as expected, identified as enriched in liver but also found at high levels, albeit much lower than for liver, in kidney and pancreas.

Evidence for the human protein-coding genes

We have determined the number of genes for which evidence is available at a protein level by combining our antibody-based data with the manual annotation of literature by the UniProt consortium (*5*) and the results from the recent mass spectrometry–based proteogenomics analyses (*9*, *10*, *12*). The analysis shows that there are 17,132 protein-coding genes with proteins identified from at least one of the three efforts and 13,841 genes with experimental evidence from at least two of the efforts (Fig. 1D). Furthermore, there is evidence, at the RNA level, for 2546 additional genes based on either our data or annotations by UniProt. Although proteins not yet detected by one of the three methods should be further investigated to establish them as true human proteins, it is noteworthy that out of the 20,356 putative protein-coding genes (in Ensembl release 75) there are only 677 genes (3.3%) for which there is no experimental evidence (table S5). Many of these genes were removed in the

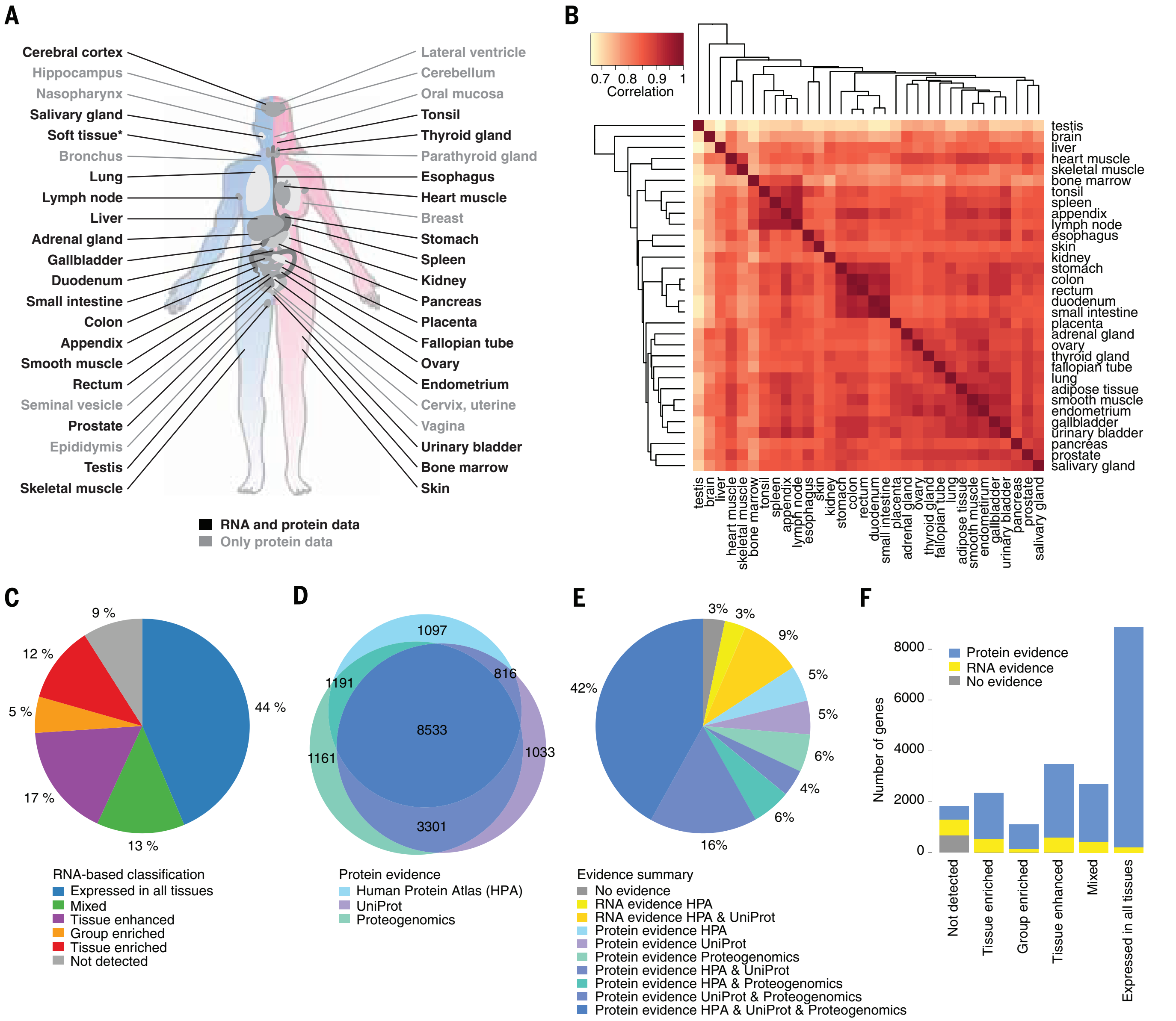


Fig. 1. Classification and protein evidence of the human protein-coding genes. (**A**) The tissues analyzed in this study, including tissues studied both by RNAseq and antibody-based profiling and those analyzed only by antibody-based profiling. For details see table S1. (**B**) Heat map showing the pairwise correlation between all 32 tissues based on transcript expression levels of 20,344 genes. The average FPKM values for each gene and tissue are used in the analysis. For correlation results of all individual samples, see fig. S1. (**C**) The number of genes classified in each expression category according to the definition stated in Table 1. (**D**) Venn diagram showing the overlap between protein evidence on the basis of three sources: Human Protein Atlas, UniProt, and Proteogenomics. (**E**) The distribution of genes classified as having protein evidence, evidence only at the transcript level, and genes without any experimental evidence. (**F**) The number of genes with protein evidence, RNA evidence, and no evidence stratified according to their transcriptomics-based classification into six categories.

later update of Ensembl (release 76) (fig. S2B), and others have been suggested to be noncoding genes on the basis of the lack of correlation in gene family age and cross-species conservation studies. Thus, it is possible that most of these "missing genes" will be removed from the list of protein-coding genes in the future. These genes and the genes with evidence only at the RNA level are obvious targets for more in-depth functional protein studies. A summary of the supporting data is shown in Fig. 1E. Few (2%) of the ubiquitously expressed genes lack protein evidence (Fig. 1F); however, protein evidence is lacking for 18% of the genes identified here by RNA analysis as elevated (tissue enriched, group enriched, or enhanced). Examples of genes with no previous evidence on the protein level according to UniProt, but now confirmed using antibody-based profiling and proteogenomics (*9*, *10*), are chromosome 2 open reading frame 57 (C2orf57), shown here with an enriched expression in testis localized to the sperm (Fig. 2A), and chromosome 8 open reading frame 47 (C8orf47), with expression in a subset of endocrine islet cells and ductal cells of the exocrine pancreas (Fig. 2B).

The tissue-elevated proteome

A network plot shows the number of tissue-enriched genes for each tissue type, as well as the number of genes enriched in different groups of tissues and organs (fig. S4). An analysis of selected tissues and organs (Fig. 2O) reveals a large number of elevated genes in male tissue, brain,

Table 1. Classification of all human protein-coding genes based on transcript expression levels in 32 tissues.

Category	Description	No. of genes	Fraction of genes (%)
Tissue enriched	mRNA levels in a particular tissue at least five times those in all other tissues	2,355	12
Group enriched	mRNA levels at least five times those in a group of 2–7 tissues	1,109	5
Tissue enhanced	mRNA levels in a particular tissue at least five times average levels in all tissues	3,478	17
Expressed in all	Detected in all tissues (FPKM > 1)	8,874	44
Mixed	Detected in fewer than 32 tissues but not elevated in any tissue	2,696	13
Not detected	FPKM < 1 in all tissues	1,832	9
Total	Total number of genes analyzed with RNAseq	20,344	100
Total elevated	Total number of tissue-enriched, group-enriched, and tissue-enhanced genes	6,942	34

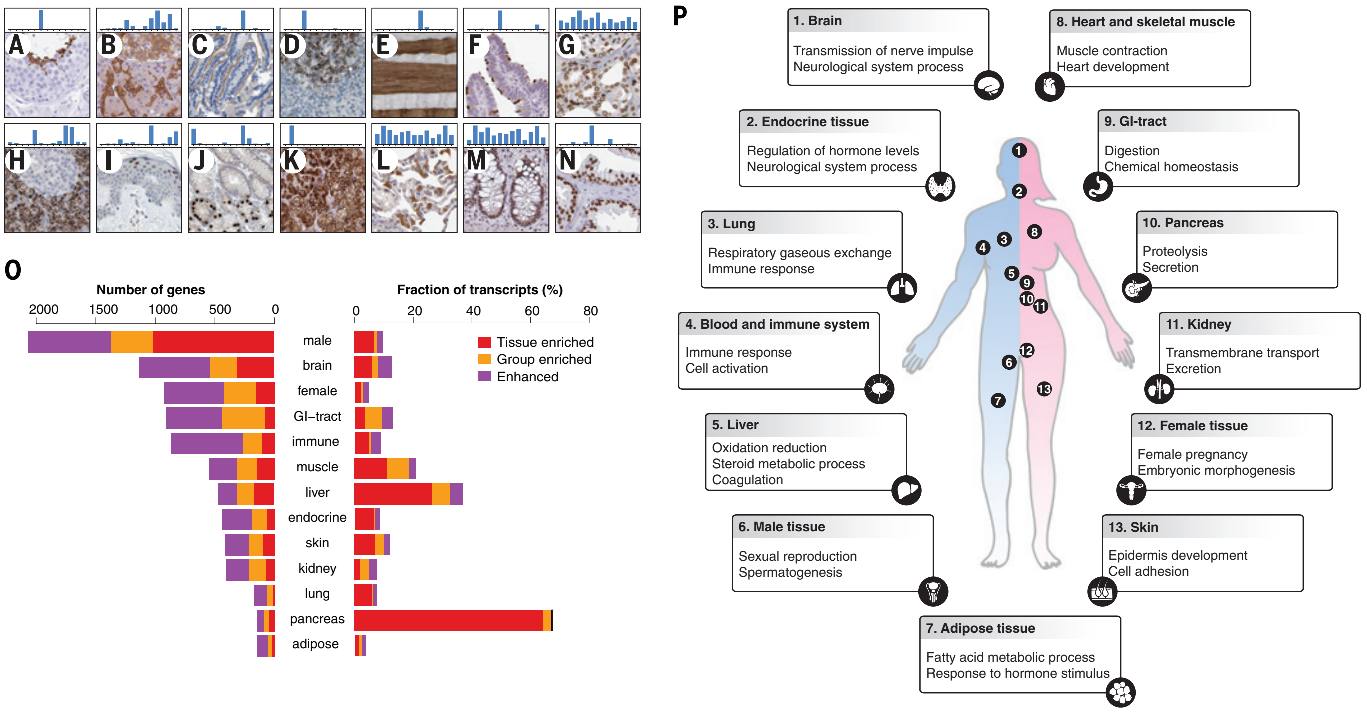


Fig. 2. Tissue microarray–based protein expression, and analysis of tissue-elevated genes in the different organ systems. (**A** to **N**) Tissue expression and localization for a selection of human proteins. Larger images corresponding to (A) to (N) of the figure are shown in fig. S3. The levels of the corresponding mRNA (FPKM) are displayed as bars for each of the 13 organ systems analyzed (from left: brain, endocrine tissue, lung, blood and immune system, liver, male tissue, adipose tissue, heart and skeletal muscle, GI tract, pancreas, kidney, female tissue, and skin). Examples include testis with C2orf57 expression in sperm (A), pancreas with cytoplasmic C8orf47 expression in both a subset of endocrine cells and ductal cells (B), duodenum with CDHR2 expression in microvilli (C), lymph node with cytoplasmic FCRLA expression in germinal center cells (D), skeletal muscle with cytoplasmic MYL3 expression in slow muscle fibers (E), fallopian tube with ROPN1L expression in cilia (F), kidney with SUN2 expression in all nuclear membranes (G), pancreas with GATM expression in mitochondria throughout the exocrine compartment (H), skin with GRHL1 expression in nuclei of the upper epidermal layer (I), stomach with nuclear PAX6 expression in endocrine cells (J), adrenal gland with cytoplasmic expression of CYP11B1 in cortical cells (K), lung with cytoplasmic COMTexpression in a subset of pneumocytes and macrophages (L), colon with nuclear ATF1 expression in glandular cells (M), and prostate with nuclear FOXA1 expression in glandular cells (N). (**O**) The number of elevated genes in the 13 organ systems, as described in (P), and the fraction of all transcripts (FPKM) encoded by these elevated genes for each of these organ systems. (**P**) An analysis of major GO terms for each tissue on the basis of the tissue-elevated genes in 13 selected tissues or groups of tissues, as described in supplementary methods. For more details of the GO analysis, see table S6.

and liver and relatively few in lung, pancreas, and fat (adipose tissue). The transcriptomics analysis also allowed us to determine the fraction of elevated transcripts in each tissue (Fig. 2O). For most tissues, only ~10% of the transcripts are encoded by tissue-elevated genes, with the exception of pancreas and liver, where elevated genes encode 70% and 35% of the transcripts, respectively.

A functional Gene Ontology (GO) analysis for 13 tissues or groups of tissues is summarized in Fig. 2P (see table S6 for details), and the terms identified are consistent with the function of the respective tissues. The largest number of enriched genes is found in the testis (n = 999), with many of the corresponding testis-specific proteins involved in the reproductive process and spermatogenesis. It is not unlikely that many of these genes will show a shared expression with oocytes in the female ovaries, which are difficult to analyze because of the different kinetics of germ cell development, including first rounds of meiosis at the embryonic stages during female life. The tissue with the second largest number of enriched genes is the brain (n = 318). The number of genes with expression restricted to neuronal tissue is relatively small, but it is likely that more enriched genes would be added to the list if additional regions, such as the various specialized regions of the brain, were sampled. Genes elevated in liver encode secreted plasma and bile proteins, detoxification proteins, and proteins associated with metabolic processes and glycogen storage, whereas genes elevated in adipose tissue encode proteins involved in lipid metabolic processes, secretion, and transport. Genes elevated in skin encode proteins associated with functions related to the barrier function (squamous cell differentiation and cornification), skin pigmentation, and hair development. In the GI tract, elevated genes predominantly encode proteins involved in nutrient breakdown, transport, and metabolism; host protection; and tissue morphology maintenance.

As expected, many of the genes enriched in groups of tissues are common for the GI tract and the hematopoietic tissues, respectively, as exemplified on the protein level by cadherin-related family member 2 (CDHR2), expressed in the microvilli of duodenum and small intestine (Fig. 2C), and Fc receptor-like A (FCRLA), expressed in lymph node, tonsil, appendix, and spleen (Fig. 2D). A large number of group-enriched genes involved in contraction are observed in striated (cardiac and skeletal) muscle, as exemplified by the fiber type–specific expression of myosin light chain 3 (MYL3) (Fig. 2E), whereas many genes shared between testis and the fallopian tube, as well as testis and lung, are involved in cell motility, as exemplified by rophilin-associated tail protein–like (ROPN1L), which is expressed in

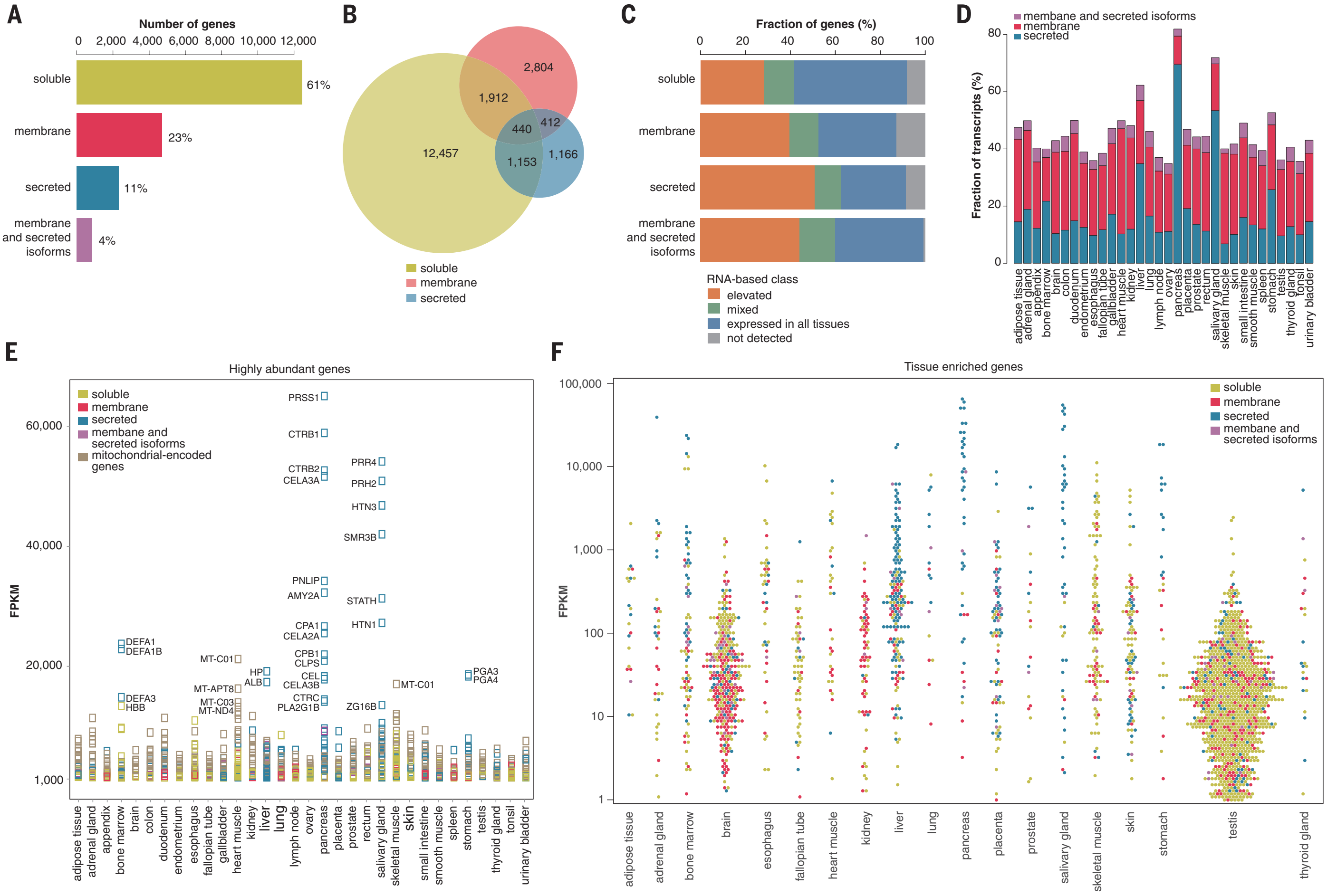


Fig. 3. Prediction and analysis of the human secreted and membrane-spanning proteins. (**A**) The number and fraction of all human genes (n = 20,356) classified into the categories soluble, membrane-spanning, and secreted, as well as genes with isoforms belonging to two or all three categories. (**B**) Venn diagram showing the number of genes in each of the three main subcellular location categories: membrane, secreted, and soluble. The overlap between the categories gives the number of genes with isoforms belonging to two or all three categories. (**C**) The fraction of genes in the various protein expression classes for the soluble, secreted, and membrane-spanning proteins, as well as genes with both secreted and membrane-spanning isoforms. (**D**) The fraction of transcripts based on FPKM values from each of the three secreted or membrane-spanning categories across the 32 analyzed tissues. (**E**) The 370 most-abundant genes (FPKM > 1000) in the different tissues, stratified according to their predicted localization on the basis of (C), as well as an additional category of the 13 genes encoded by the mitochondrial genome. The gene names for a selection of the most abundant genes are shown. (**F**) The transcript levels (FPKM) on a $\log_{10}$ scale for all genes identified as tissue-enriched are shown for a few selected tissues, with each gene stratified according to predicted localization.

sperm (testis), ciliated cells in respiratory epithelia (lung), and ciliated cells in the fallopian tube (Fig. 2F).

The human secretome and membrane proteome

Both secreted and membrane-bound proteins play crucial roles in many physiological and pathological processes. Important secreted proteins include cytokines, coagulation factors, hormones, and growth factors, whereas membrane proteins include ion channels or molecular transporters, enzymes, receptors, and anchors for other proteins. Here, we performed a whole-proteome scan to predict the complete set of human secreted proteins ("secretome") using three methods for signal-peptide prediction: SignalP4.0 (*23*), Phobius (*24*), and SPOCTOPUS (*25*). In addition, the human membrane proteome was predicted using seven membrane–protein topology prediction methods as described (*21*), which resulted in a majority decision–based method (MDM). For each protein-coding gene, all protein isoforms were annotated for predicted localization: secreted, membrane spanning, or soluble (intracellular proteins without a predicted signal peptide or membrane-spanning region) (table S1). Some of the proteins predicted to be membrane-spanning are intracellular, e.g., in the Golgi or mitochondrial membranes, and some of the proteins predicted to be secreted could potentially be retained in a compartment belonging to the secretory pathway, such as the endoplasmic reticulum (ER), or remain attached to the outer face of the cell membrane by a GPI anchor. About 3000 human genes are predicted to encode secreted proteins, with another 5500 encoding membrane-bound proteins (Fig. 3A). In the interactive database (www.proteinatlas.org), many of the secreted proteins are detected at the RNA level in tissues, but no protein expression is observed in the antibody-based analysis in the same tissue—most likely because the steady-state levels of proteins in the cell during the secretion process are too low to be detected.

A large fraction (72%) of human genes encode multiple splice variants with different protein sequences. In Fig. 3B, all genes have been classified according to the presence of protein isoforms that are intracellular, membrane-spanning, and/or secreted. Note that two-thirds of the genes encoding secreted proteins have at least one splice variant with alternative localization. All protein isoforms (n = 94,856) with their predicted localization based on the three signal-peptide–prediction methods, as well as the number of predicted transmembrane segments, are listed in table S7. An analysis across the 32 tissues (Fig. 3C) supports earlier suggestions (*21*, *26*) that a larger fraction of tissue-enriched proteins are secreted or membrane-spanning proteins than are intracellular proteins.

Furthermore, we investigated the fraction of the transcriptome that codes for each class of proteins across the 32 tissues (Fig. 3D and fig. S4). In most cases, the secreted proteins account for between 10 and 20% of the transcripts. In contrast, more than 70% of the transcripts from the pancreas and ~60% from the salivary gland encode secreted proteins. This demonstrates the extreme specialization of these two tissues for production of secreted proteins into the duodenum

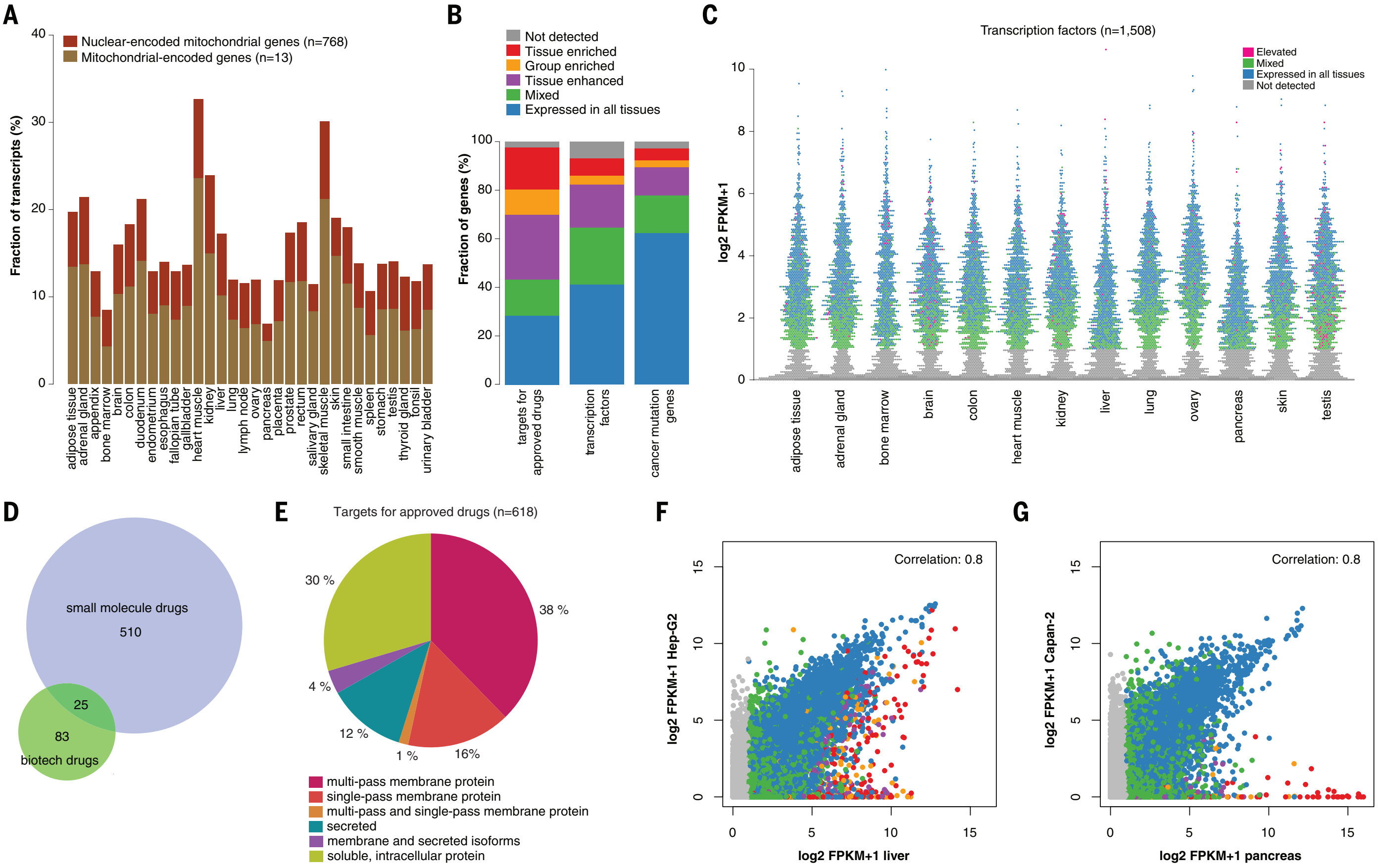


Fig. 4. The human transcriptome in different tissues and organs. (**A**) The fraction of transcripts encoded by mitochondrial genes for each of the different tissues and organs, subdivided by genes encoded by the mitochondrial genome and chromosomes, respectively. (**B**) The fraction of genes classified according to tissue expression pattern and analyzed for all targets of approved drugs (n = 618), all transcription factors (n = 1508), and proteins implicated in cancer (n = 525). (**C**) The transcript levels (FPKM) for all genes encoding transcription factors in some selected tissues, color-coded according to their global expression category. (**D**) The number of pharmaceutical drugs approved by FDA, according to Drugbank (*19*), that are chemical (small-molecule) or biotech drugs. (**E**) The number of pharmaceutical drugs approved by FDA (*19*) stratified according to the predicted localization of the target protein. (**F**) Pairwise comparison showing all genes expressed in liver tissue and the liver cell line Hep-G2, color-coded according to protein expression category as shown in (B). (**G**) Pairwise comparison showing all genes expressed in pancreas tissue and the pancreas cell line Capan-2, color-coded according to protein expression category as shown in (B).

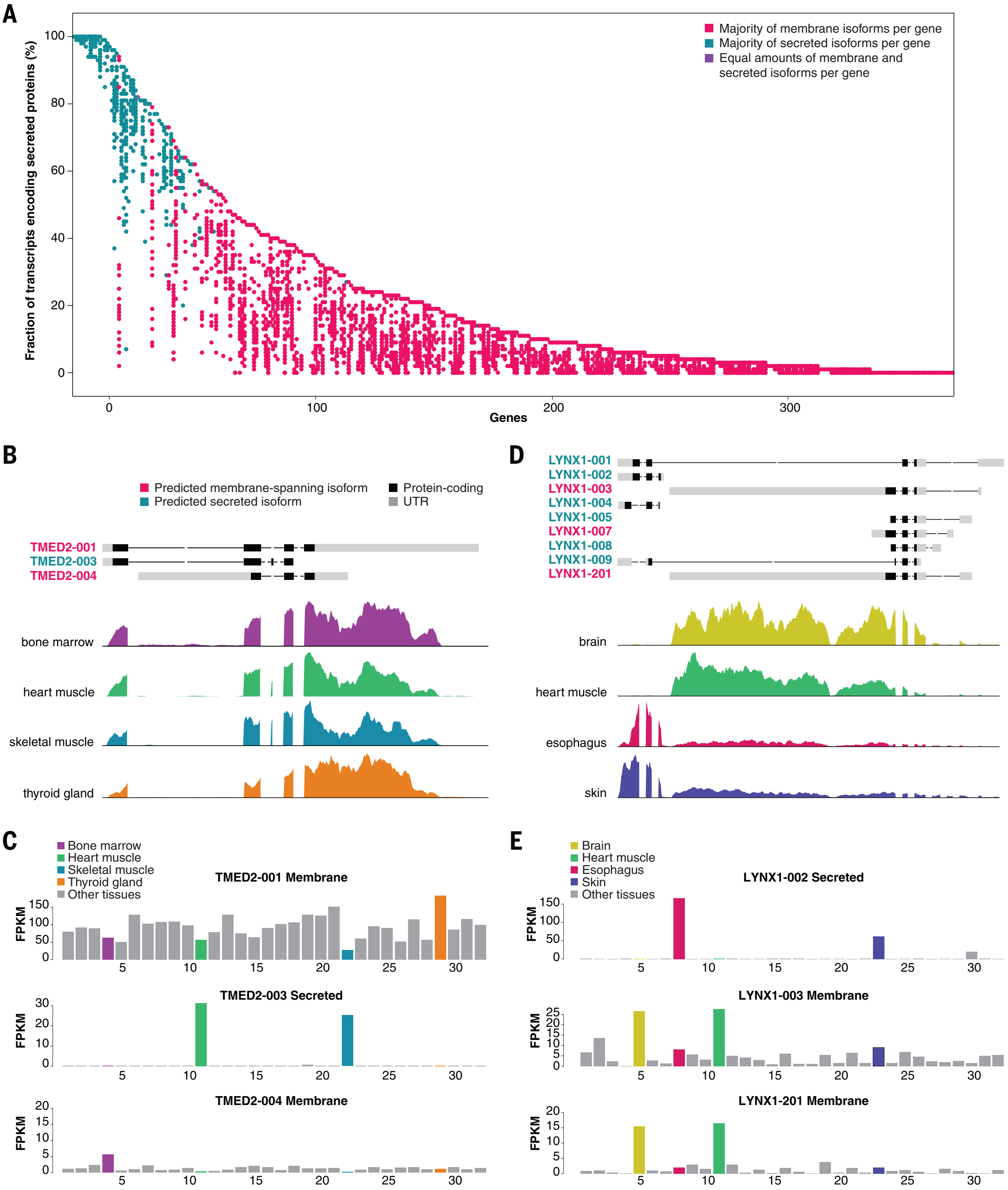


Fig. 5. Differential splicing analysis of transcripts. (**A**) Dot plot of genes with multiple isoforms, where at least one isoform is classified as membrane-spanning and another classified as secreted. The *x* axis shows 366 genes expressed at >5 FPKM in one or more tissues; the *y* axis shows the sum of FPKM values for all secreted isoforms divided by the total sum of FPKM values for each tissue expressed at >5 FPKM. For each gene, the number of tissues where the secreted transcripts are more abundant than the membrane-spanning transcripts is calculated to define a majority fraction-type as membrane (red), secreted (blue), or equal number for both categories (purple). Each tissue is represented by a circle, and the color is the same across all tissues for the same gene. (**B**) Example of differential splicing for the gene *TMED2*, with two isoforms predicted as membrane-spanning and one isoform predicted as secreted. The exon-intron structure (with pure intronic sites removed), as well as the location of the untranslated regions (UTR) of three splice variants of *TMED2*, are shown on top. Normalized read coverage plots for cardiac muscle, skeletal muscle, thyroid gland, and bone marrow highlight the differential use of exons in the selected tissues. (**C**) Transcript abundance (FPKM values) plotted across all 32 tissues for each isoform. The predicted membrane-spanning transcript (top) is expressed in all tissues, with thyroid gland as the most abundant tissue; a secreted isoform (middle) is only detected in cardiac and skeletal muscle; and a second membrane-spanning isoform (bottom) is expressed at very low levels, with bone marrow as most abundant. (**D**) Examples of differential splicing for the gene *LYNX1*, with three isoforms predicted as membrane-spanning and six isoforms predicted as secreted from the visualization used in (B). (**E**) Transcript abundance (FPKM values) for three isoforms of *LYNX1* detected at >5 FPKM. The secreted isoform (top) is expressed at high levels in esophagus and skin, whereas the two membrane-spanning isoforms (middle and bottom) are most abundant in brain and cardiac muscle.

and oral cavity, respectively. About 40% of the transcripts in liver encode secreted proteins. Other tissues with relatively high levels of transcripts encoding secreted proteins include gallbladder, bone marrow, placenta, and different parts of the GI tract, such as stomach, duodenum, and small intestine.

The most abundant genes, normalized as fragments per kilobase of exon per million fragments mapped (FPKM) with a value >1000, in the diferent tissues are shown in Fig. 3E, and the prediction of the localization of the corresponding proteins reveals that many (53%) are secreted proteins. Among the predicted membrane-spanning proteins, 13 proteins encoded in the mitochondrial genome are the most highly expressed. In Fig. 3F, tissue-enriched genes are shown stratified according to their predicted subcellular localization. Many of the tissue-enriched genes in testis are intracellular, whereas a large number of the tissue-enriched genes in brain and kidney are membrane-bound. In contrast, in many other tissues, such as pancreas, salivary gland, liver, stomach, and bone marrow, most tissue-enriched genes are secreted (fig. S5).

The housekeeping proteome

Transcriptomics analysis shows that close to 9000 genes (table S1) are expressed in all analyzed tissues, which suggests that the gene products are needed in all cells to maintain basic cellular structure and function. These housekeeping proteins include ribosomal proteins involved in protein synthesis, enzymes essential for cell metabolism and gene expression, and mitochondrial proteins needed for energy generation, as well as proteins responsible for the structural integrity of the cell. Most of these proteins are expressed at similar levels throughout the human body, as exemplified in kidney by the expression of the nuclear membrane protein SUN2 present in all cells (Fig. 2G), whereas a few proteins show great variability in expression levels—for example, the mitochondrial protein glycin amino transferase (GATM), with high expression in exocrine pancreas (Fig. 2H), kidney, and liver but relatively low expression levels in all other tissues. An interesting class of proteins is encoded by mitochondrial genes, and in Fig. 4A, the transcriptional load of these genes is shown across different tissues. The highest fractions of transcripts encoding mitochondrial proteins are found in cardiac muscle (32% of all transcripts) and skeletal muscle (28%), which demonstrates the importance of energy metabolism for striated muscle tissue.

The regulatory proteome

Transcription factors, of which ~1,500 have been identified in humans ([27]), comprise an important class of regulatory proteins as they function as on/off switches for gene expression. The fraction of transcription factor genes classified according to tissue specificity is shown in Fig. 4B, which suggests a tissue distribution similar to that of the complete proteome, with as many as 41% of the genes expressed in all tissues and only 29% identified as elevated (enriched, group enriched, or enhanced). Many of the more-abundantly expressed transcription factors are found in all tissues (Fig. 4C). However, there are examples of abundant transcription factors that belong to the tissue-elevated categories, such as (i) grainyhead-like 1 (GRHL1) with enhanced expression in esophagus and skin (squamous epithelia) and selective localization to the uppermost nucleated epidermal keratinocytes (Fig. 2I) and (ii) paired box 6 (PAX6) involved in eye and brain development and differentiation of pancreatic islet cells, with group-enriched expression in brain, pancreas, and stomach, selectively localized to a subset of glandular cells in the stomach mucosa (Fig. 2J) and to islet cells in the pancreas. The tissue-enriched transcription factors identified here (table S8) will enable new insights into the regulatory pattern of the different tissues.

The druggable proteome

Most pharmaceutical drugs act by targeting proteins and modulating their activity. Target proteins belong to four main families: enzymes, transporters, ion channels, and receptors. The U.S. Food and Drug Administration (FDA) has approved drugs targeting human proteins from 618 genes, according to Drugbank ([19]), with most drugs acting on signal transduction proteins that convert extracellular signals into intracellular responses. Antibody-based drugs are usually unable to penetrate the plasma membrane, and therefore, they target cell surface proteins, such

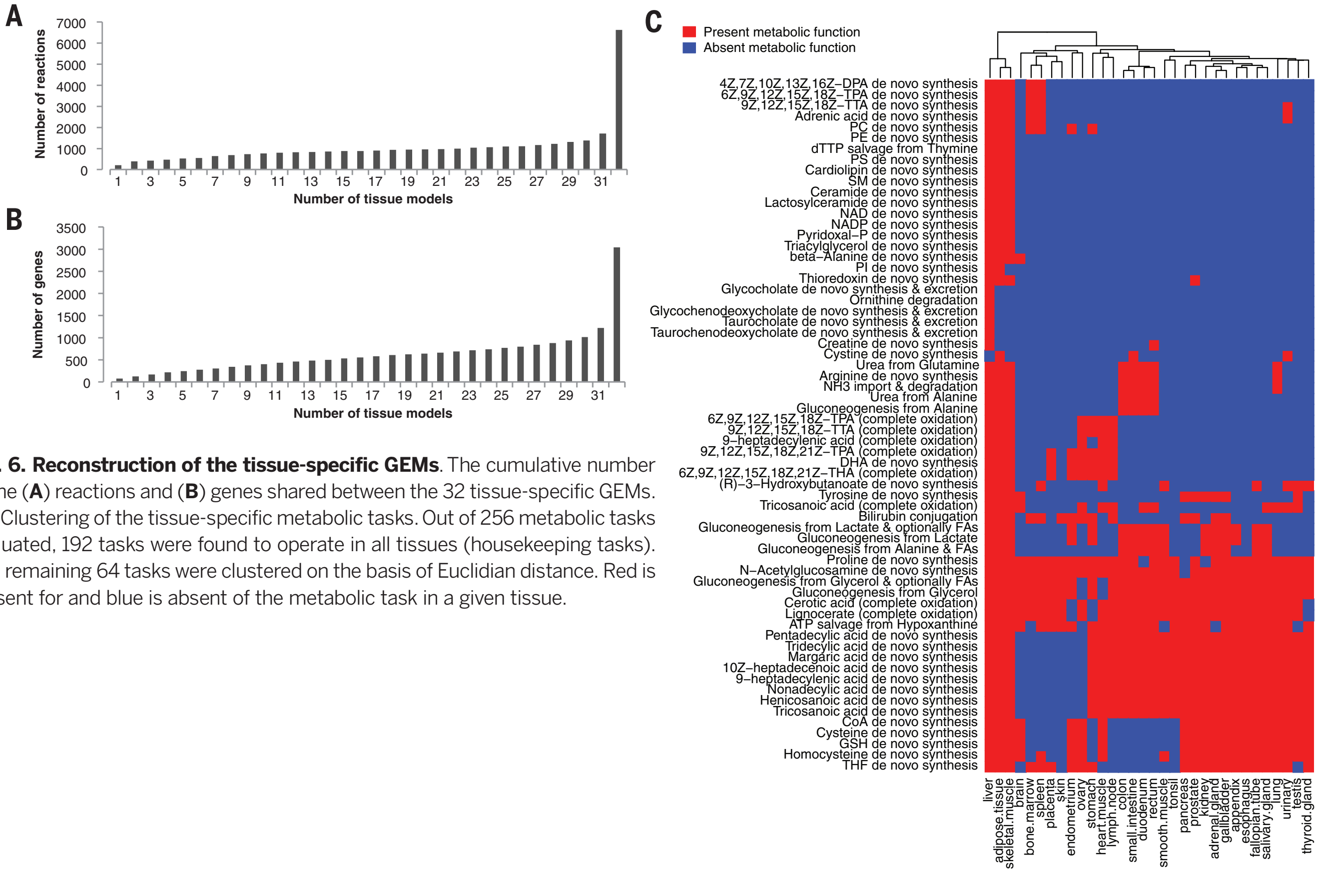


Fig. 6. Reconstruction of the tissue-specific GEMs. The cumulative number of the (**A**) reactions and (**B**) genes shared between the 32 tissue-specific GEMs. (**C**) Clustering of the tissue-specific metabolic tasks. Out of 256 metabolic tasks evaluated, 192 tasks were found to operate in all tissues (housekeeping tasks). The remaining 64 tasks were clustered on the basis of Euclidian distance. Red is present for and blue is absent of the metabolic task in a given tissue.

as receptors, whereas small-molecule drugs can diffuse into cells and act also on intracellular targets. An analysis of the proteins encoded from the 618 genes shows that 535 proteins are targeted by small chemical molecules, whereas 108 proteins are targeted by biotech drugs (Fig. 4D). The predicted subcellular localization (Fig. 4E) shows that 59% of the targets are predicted membrane proteins and that 16% are secreted, including those with both secreted and membrane-bound isoforms. The genes corresponding to these drug targets were classified according to tissue specificity, and the results (Fig. 4B and table S9) show a bias for tissue-elevated proteins (enriched, group enriched, or enhanced), although as many as 30% of the approved drugs target proteins expressed in all analyzed tissues. One example of a target with enriched expression is cytochrome P450 11B1 (CYP11B1), which is involved in the conversion of progesterone to cortisol in the adrenal gland (Fig. 2K), whereas a ubiquitously expressed protein is the catechol-*O*-methyltransferase (COMT), which is associated with degradation of neurotransmittors and is important in the metabolism of drugs used in treatment of Parkinson's disease. COMT displays cytoplasmic expression in all analyzed tissues, including lung (Fig. 2L). The ubiquitous expression may have implications for treatments using these proteins as drug targets.

The cancer proteome

Genes implicated in cancer are often essential for orderly growth, survival, and basic cell functions in normal cells and tissues, whereas overexpression, loss of expression, or expression of a mutated protein contributes to dysfunction and tumor growth. The number of genes implicated in cancer is dependent on definitions; however, 259 genes have been shown to be mutated across 21 tumor types ([28]); 290 genes have been reported as cancer driver genes across 12 tumor types ([29]); and 525 genes have been implicated in malignant transformation, according to a catalog of somatic mutations in cancer (COSMIC) ([20]). Expression analysis based on our transcriptomics data shows that a majority (60%) of these last-mentioned genes (Fig. 4B and table S10) is expressed in all tissues, with only a fraction of genes expressed in a tissue- or group-enriched manner. Examples are the activating transcription factor 1 (ATF1) (Fig. 2M), a protein expressed in all tissues with known translocations in sarcomas, and the forkhead box A1 (FOXA1) (Fig. 2N), a protein with enhanced expression where somatic mutations in a subset of prostate cancers have been reported ([30]). The lack of tissue specificity for many of these genes is not surprising because many of the corresponding proteins are involved in normal growth regulation and cell cycle control, but it also emphasizes the possible adverse effects of treatment with drugs targeting proteins expressed in all tissues.

Tissue versus cell lines

Human biology and diseases are often explored using cell lines as model systems. We compared the body-wide expression in human tissues with expression in cancer cell lines derived from corresponding tissue types. The transcriptomes for 11 cell lines were described earlier ([31]), whereas the transcriptomes for an additional 36 cell lines were generated as part of this study (see table S11). Genome-wide expression patterns comparing normal tissues with corresponding human cell lines are shown in fig. S6, as exemplified by the liver cancer–derived cell line Hep-G2 (Fig. 4F), and the pancreas cancer–derived cell line Capan-2 (Fig. 4G). Many of the tissue-enriched genes identified in normal tissues are down-regulated or completely "turned off" in the corresponding cell lines, and in contrast, the housekeeping proteins are expressed at the same level in both tissues and corresponding cell lines. These results support earlier studies ([32]) suggesting that cell lines are "dedifferentiated," with shared characteristics and lack of tissue-specific features due to down-regulation of tissue-enriched genes. This implies that conclusions from cell line studies should only be conferred on the corresponding tissue with caution.

The isoform proteome

Protein isoforms endow the structural space of the human proteome with breadth and complexity ([33]). Isoforms are produced through alternative splicing, posttranslational modifications, proteolytic cleavage, somatic recombination, or genetic variations in protein-coding regions. We explored genes encoding isoforms with different predicted localization (secreted or membrane spanning) (table S12). A large number of these genes (n = 366) are displayed together with the fraction of all transcripts (mRNA molecules) in Fig. 5A, with splice variants that yield secreted proteins. Most of the genes (67%) have more than 80% of the transcripts encoding only one of the two localizations across all 32 tissues, but there are some proteins for which the majority of the transcripts encode a secreted form in one tissue, whereas the majority of the transcripts encode a membrane protein in another tissue. As an example, the expression levels for different isoforms of the poorly understood transmembrane emp24 domain–trafficking protein 2 (TMED2) are shown in Fig. 5, B and C. Cardiac muscle has a tissue-specific expression of the secreted form, whereas the membrane-bound form is detected in all other tissue types, although at variable levels. Similarly, the protein Ly6 or neurotoxin 1 (LYNX1) shows a selective expression of the secreted isoform in the esophagus and the skin, whereas the membrane-bound form is found in other tissue types and is most abundantly expressed in the brain and the cardiac muscle (Fig. 5, D and E). The different localizations of the isoforms are consistent with the predicted functions of the different isoforms. In most cases, one of the isoforms dominates across all tissues, which is also consistent with earlier studies ([34]). These are starting points to explore the relation between tissue-specific expression and function.

Tissue-based map of human metabolism

Genome-scale metabolic models (GEMs) provide not only the best representation of the metabolic capabilities of cell and/or tissue types but also quantitative descriptions of the genotype-phenotype relationship ([35]). Using the RNAseq data, we reconstructed tissue-specific GEMs for 32 different tissues using the generic metabolic model, HMR2 ([36]), and generated a map of the complete human metabolism. All models were generated such that they can carry out 56 metabolic tasks identified to be present in all human cell types ([37]). The numbers of the reactions, metabolites, and genes incorporated into each tissue-specific GEM are presented (table S13), and the models are provided in SBML format at the Human Metabolic Atlas portal ([38]). In order to confirm that none of the models have futile cycles, we ensured that high-energy compounds cannot be generated from low-energy compounds using metabolic tasks including rephosphorylation of adenosine triphosphate or the generation of a proton gradient over the membranes (table S14).

A total of 6627 reactions, 3040 genes, and 4847 metabolites were present in at least one of the tissue models, and 4912 reactions, 1822 genes, and 3984 metabolites were present in all models. This shows that about 75% of all metabolic reactions in the human body are operating in all key tissues, which clearly illustrates the central role metabolism is playing for basic cellular function. At a gene level, the consensus expression in all tissues is, however, less (i.e., about 60%), which shows that, even though different tissues have the same metabolic reactions, it is different isoforms of the enzymes that are responsible for catalyzing these reactions. Our analysis is the first genome-wide illustration of this wide variation in enzyme usage for catalyzing the same reaction between human tissues.

We found that only 207 of the reactions (Fig. 6A) and 74 of the genes (Fig. 6B) were unique to any of the tissues, and notable differences between the genes (fig. S7) and reactions (fig. S8) based on pairwise comparisons of the various tissues were observed. Between 57 and 632 genes differed in these comparisons of the tissue models, representing 9 to 21% of the genes shared in all models. Bone marrow has the lowest number of genes and reactions, whereas liver has a large number of genes and reactions not present in any other tissue. Many of the metabolic reactions in liver involve specialized lipid metabolism, e.g., de novo synthesis and secretion of bile acids including glycocholate, taurocholate, glycochenodeoxycholate, and taurochenodeoxycholate, but there are also other metabolic functions specific to liver such as ornithine degradation. To further investigate the metabolic capability of each tissue-specific GEM, we defined 256 metabolic tasks (table S15) that are known to occur in humans. The analysis shows that 192 of these metabolic tasks can be performed in all analyzed tissues, whereas the remaining 64 metabolic tasks were performed by some GEMs and clustering of these 64 metabolic tasks is shown in Fig. 6C (see also table S16). The analysis demonstrates liver as the most metabolically active tissue, followed by adipose and skeletal muscle. For all the remaining tissues, there are variations in the metabolic activities, but with clustering of activities in tissues with similar function and morphology, e.g., stomach, duodenum, and small intestine.

Discussion

Here, we present a tissue-based map of the human proteome from analyses of 32 tissues and 47 cell lines, with gene expression data on both the RNA and protein level and with supplementary analyses on the protein level for an additional 12 tissues. An interactive resource is presented as part of the Human Protein Atlas portal (www.proteinatlas.org). This allows exploration of the tissue-elevated proteomes in these tissues and organs and analysis of tissue profiles for specific protein classes, including proteins involved in housekeeping functions in the human body, such as cell growth, energy generation, and metabolic pathways; groups of proteins involved in diseases; and proteins targeted by pharmaceutical drugs. Comprehensive lists of genes expressed at elevated levels in these tissues have been compiled, with quantitative expression profiles provided by the deep-sequencing transcriptomics complemented with immunohistochemistry. This provides localization of the proteins in the subcompartments of each tissue and organ down to the single-cell level. To facilitate integration with other biological resources, all data are available for download and through collaborations cross-linked with efforts such as UniProt (*5*), NextProt (*6*), ProteomicsDB (*9*), Metabolic Atlas (*38*), and the pan-European ELIXIR project (*39*). An important short-term objective is to facilitate international efforts (*5*, *7*, *8*, *40*) to explore the "missing proteins," with the aim to provide a finite list of human protein-coding genes and to generate firm protein evidence and expression characteristics for all of these genes. In addition, the primary data here can be used to expand the analysis of the isoform proteome to better understand the role of this diverse proteome for the functional biology of humans.

REFERENCES AND NOTES

1. P. Flicek *et al*., Ensembl 2013. *Nucleic Acids Res.* **41** (Database), D48–D55 (2013). doi: 10.1093/nar/gks1236; pmid: 23203987
2. K. D. Pruitt, T. Tatusova, G. R. Brown, D. R. Maglott, NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **40** (Database), D130–D135 (2012). doi: 10.1093/nar/gkr1079; pmid: 22121212
3. H. Kawaji *et al*., Update of the FANTOM web resource: From mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.* **39** (Database), D856–D860 (2011). doi: 10.1093/nar/gkq1112; pmid: 21075797
4. A. Brazma *et al*., ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003). doi: 10.1093/nar/gkg091; pmid: 12519949
5. M. Magrane, U. Consortium, UniProt Knowledgebase: A hub of integrated protein data. *Database (Oxford)* **2011**, bar009 (2011). doi: 10.1093/database/bar009; pmid: 21447597
6. P. Gaudet *et al*., neXtProt: Organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **12**, 293–298 (2013). doi: 10.1021/pr300830v; pmid: 23205526
7. I. Dunham *et al*., An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
8. Y. K. Paik *et al*., The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **30**, 221–223 (2012). doi: 10.1038/nbt.2152; pmid: 22398612
9. M. Wilhelm *et al*., Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).pmid: 24870543
10. M. S. Kim *et al*., A draft map of the human proteome. *Nature* **509**, 575–581 (2014). doi: 10.1038/nature13302; pmid: 24870542
11. M. Mann, Functional and quantitative proteomics using SILAC. *Nat. Rev. Molec. Cell Biol.* **7**, 952–958 (2006). doi: 10.1038/nrm2067; pmid: 17139335
12. I. Ezkurdia *et al*., Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014). doi: 10.1093/hmg/ddu309; pmid: 24939910
13. V. Lange, P. Picotti, B. Domon, R. Aebersold, Selected reaction monitoring for quantitative proteomics: A tutorial. *Mol. Syst. Biol.* **4**, 222 (2008). doi: 10.1038/msb.2008.61; pmid: 18854821
14. M. Uhlen *et al*., Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010). doi: 10.1038/nbt1210-1248; pmid: 21139605
15. L. Fagerberg *et al*., Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014). doi: 10.1074/mcp.M113.035600; pmid: 24309898
16. C. Kampf *et al*., The human liver-specific proteome defined by transcriptomics and antibody-based profiling. *FASEB J.* **28**, 2901–2914 (2014). doi: 10.1096/fj.14-250555; pmid: 24648543
17. D. Djureinovic *et al*., The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Mol. Hum. Reprod.* **20**, 476–488 (2014). doi: 10.1093/molehr/gau018; pmid: 24598113
18. G. Gremel *et al*., The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling. *J. Gastroenterol.* **50**, 46–57 (2014). doi: 10.1007/s00535-014-0958-7; pmid: 24789573
19. V. Law *et al*., DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **42** (D1), D1091–D1097 (2014). doi: 10.1093/nar/gkt1068; pmid: 24203711
20. COSMIC catalogue of somatic mutations in cancer (2014); http://cancer.sanger.ac.uk/cancergenome/projects/census.
21. E. Lundberg *et al*., Defining the transcriptome and proteome in three functionally different human cell lines. *Mol. Syst. Biol.* **6**, 450 (2010). doi: 10.1038/msb.2010.106; pmid: 21179022
22. Y. Taniguchi *et al*., Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010). doi: 10.1126/science.1188308; pmid: 20671182
23. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011). doi: 10.1038/nmeth.1701; pmid: 21959131
24. L. Käll, A. Krogh, E. L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35** (Web server), W429–W432 (2007). doi: 10.1093/nar/gkm256; pmid: 17483518
25. H. Viklund, A. Bernsel, M. Skwark, A. Elofsson, SPOCTOPUS: A combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**, 2928–2929 (2008). doi: 10.1093/bioinformatics/btn550; pmid: 18945683
26. D. Ramsköld, E. T. Wang, C. B. Burge, R. Sandberg, An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLOS Comput. Biol.* **5**, e1000598 (2009). doi: 10.1371/journal.pcbi.1000598; pmid: 20011106
27. E. Wingender, T. Schoeps, J. Dönitz, TFClass: An expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **41** (D1), D165–D170 (2013).
28. D. Tamborero *et al*., Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013). pmid: 24084849
29. D. M. Muzny *et al*., Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012). doi: 10.1038/nature11252; pmid: 22810696
30. C. S. Grasso *et al*., The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012). doi: 10.1038/nature11125; pmid: 22722839
31. F. Danielsson *et al*., RNA deep sequencing as a tool for selection of cell lines for systematic subcellular localization of all human proteins. *J. Proteome Res.* **12**, 299–307 (2013). doi: 10.1021/pr3009308; pmid: 23227862
32. M. Schnabel *et al*., Dedifferentiation-associated changes in morphology and gene expression in primary human articular chondrocytes in cell culture. *Osteoarthritis Cartilage* **10**, 62–70 (2002). doi: 10.1053/joca.2001.0482; pmid: 11795984
33. E. Birney *et al*., Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007). doi: 10.1038/nature05874; pmid: 17571346
34. M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, A. Brazma, Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14**, R70 (2013). doi: 10.1186/gb-2013-14-7-r70; pmid: 23815980
35. A. Mardinoglu, J. Nielsen, Systems medicine and metabolic modelling. *J. Intern. Med.* **271**, 142–154 (2012). doi: 10.1111/j.1365-2796.2011.02493.x; pmid: 22142312
36. A. Mardinoglu *et al*., Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 3083 (2014). doi: 10.1038/ncomms4083; pmid: 24419221
37. R. Agren *et al*., Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.* **10**, 721 (2014). doi: 10.1002/msb.145122; pmid: 24646661
38. Human Metabolic Atlas (2014); www.metabolicatlas.org/.
39. L. C. Crosswell, J. M. Thornton, ELIXIR: A distributed infrastructure for European biological data. *Trends Biotechnol.* **30**, 241–242 (2012). doi: 10.1016/j.tibtech.2012.02.002; pmid: 22417641
40. J. F. Rual *et al*., Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).

ACKNOWLEDGMENTS

We acknowledge the entire staff of the Human Protein Atlas program; the Science for Life Laboratory; and the pathology team in Mumbai, India, for valuable contributions. We thank the Department of Pathology at the Uppsala Akademiska Hospital, Uppsala, Sweden, and Uppsala Biobank for kindly providing clinical diagnostics and specimens used in this study. We also acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure (NGI), and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure. Funding was provided by the Knut and Alice Wallenberg Foundation. The authors declare that they have no conflict of interest. Correspondence and requests for materials should be addressed to M.U. The mRNA levels of all genes in each tissue sample (*n* = 122) are available in table S18. The supplementary Excel tables are available in the supplementary material and at www.proteinatlas.org/about/publicationdata. The raw sequencing data are available at ArrayExpress (www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2836/) and BioProject (NIH) (www.ncbi.nlm.nih.gov/bioproject/PRJNA183192). All Protein Atlas (protein) data are available in structured XML format and can be downloaded from www.proteinatlas.org/about/download.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/347/6220/1260419/suppl/DC1
Materials and Methods
Figs. S1 to S8
Tables S1 to S18
References (*41*–*61*)

25 August 2014; accepted 5 December 2014
10.1126/science.1260419

RESEARCH ARTICLE SUMMARY

PROTEOMICS

A subcellular map of the human proteome

Peter J. Thul,* Lovisa Åkesson,* Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M. Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M. Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M. Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P. Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle von Feilitzen, Kathryn S. Lilley, Mathias Uhlén,† Emma Lundberg†

INTRODUCTION: A complete view of human biology can only be achieved by studying the molecular components of its smallest functional unit, the cell. Cells are internally organized into compartments called organelles. The spatial partitioning provided by organelles creates an enclosed environment or surface for chemical reactions tailored to fulfill specific functions. These functions are tightly linked to a specific set of proteins. Therefore, resolving the subcellular location of the human proteome provides information about the function of the organelle and its underlying cellular mechanisms. We present a subcellular map of the human proteome, called the Cell Atlas, to facilitate functional exploration of individual proteins and their role in human biology and disease.

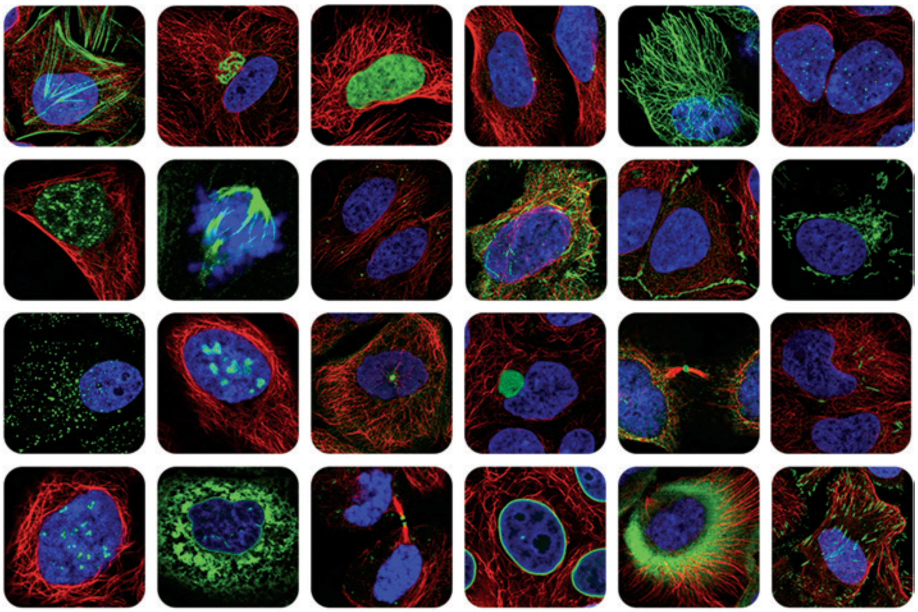
RATIONALE: Immunofluorescence (IF) microscopy was used to systematically resolve the spatial distribution of human proteins in cultivated cell lines and map them to cellular compartments and substructures with single-cell resolution. This approach allowed definition of the precise location of a majority of the human proteins in their cellular context and exploration of single-cell variations in protein expression patterns. The proteome-wide information about protein spatial distribution was validated with an orthogonal proteomics method, and the results were integrated into existing network models of protein-protein interactions for increased accuracy.

RESULTS: We report a high-resolution characterization of the spatial subcellular distribution of the human proteome based on more than 80,000 confocal IF images. A total of 12,003 proteins targeted by 13,993 antibodies were classified into one or several of 30 cellular compartments and substructures, altogether defining the proteomes of 13 major organelles. The organelles with the largest proteomes were the nucleus and its substructures (6245 proteins), such as bodies and speckles, and the cytosol (4279 proteins). However, smaller organelles such as the midbody, rods and rings, and nucleoli also showed a larger diversity than previously recognized. Intriguingly, about half of all proteins were localized to multiple compartments, showing that there is a shared pool of proteins even among functionally unrelated organelles. Single-cell analysis revealed 1855 proteins with variation in their expression pattern, either in terms of expression levels or spatial distribution. Last, the spatial information was used to refine biological networks. Our location-pruned network that restricts protein interaction to the same organelle improved the accuracy of the human interactome model. The analysis also included transcriptomics data for all putative protein-coding genes (19,628) in 56 human cell lines of various origins. On average, cell lines expressed 11,490 genes, with half of them (6295) being expressed across all samples, suggesting a "housekeeping" role.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aal3321

CONCLUSION: The cellular proteome is compartmentalized and spatiotemporally regulated to a high degree. The high-resolution subcellular map of the human proteome that we provide describes this cellular complexity, with many multilocalizing proteins and single-cell variations. The map is presented as an interactive database called the Cell Atlas, part of the Human Protein Atlas (www.proteinatlas.org). The Cell Atlas constitutes a key resource for a holistic understanding of the human cell and its complex underlying molecular machinery, as well as a major step toward modeling the human cell. ■



Creation of an image-based map of the human subcellular proteome. The subcellular locations of 12,003 proteins were determined by IF microscopy in cell lines of various origins. High-resolution IF images such as those shown above enabled mapping of proteins to distinct subcellular structures. This resulted in the definition of the proteomes of 13 major cellular organelles, revealing multilocalizing proteins, as well as expression variability on a single-cell level.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: mathias.uhlen@scilifelab.se (M.U.); emma.lundberg@scilifelab.se (E.L.)

Cite this article as P. J. Thul *et al.*, *Science* **356**, eaal3321 (2017). DOI: 10.1126/science.aal3321

Originally published 26 May 2017

RESEARCH ARTICLE

PROTEOMICS

A subcellular map of the human proteome

Peter J. Thul,[1]* Lovisa Åkesson,[1]* Mikaela Wiking,[1] Diana Mahdessian,[1] Aikaterini Geladaki,[2,3] Hammou Ait Blal,[1] Tove Alm,[1] Anna Asplund,[4] Lars Björk,[1] Lisa M. Breckels,[2,5] Anna Bäckström,[1] Frida Danielsson,[1] Linn Fagerberg,[1] Jenny Fall,[1] Laurent Gatto,[2,5] Christian Gnann,[1] Sophia Hober,[6] Martin Hjelmare,[1] Fredric Johansson,[1] Sunjae Lee,[1] Cecilia Lindskog,[4] Jan Mulder,[7] Claire M. Mulvey,[2] Peter Nilsson,[1] Per Oksvold,[1] Johan Rockberg,[6] Rutger Schutten,[1] Jochen M. Schwenk,[1] Åsa Sivertsson,[1] Evelina Sjöstedt,[4] Marie Skogs,[1] Charlotte Stadler,[1] Devin P. Sullivan,[1] Hanna Tegel,[6] Casper Winsnes,[1] Cheng Zhang,[1] Martin Zwahlen,[1] Adil Mardinoglu,[1] Fredrik Pontén,[4] Kalle von Feilitzen,[1] Kathryn S. Lilley,[2] Mathias Uhlén,[1]† Emma Lundberg[1]†

Resolving the spatial distribution of the human proteome at a subcellular level can greatly increase our understanding of human biology and disease. Here we present a comprehensive image-based map of subcellular protein distribution, the Cell Atlas, built by integrating transcriptomics and antibody-based immunofluorescence microscopy with validation by mass spectrometry. Mapping the in situ localization of 12,003 human proteins at a single-cell level to 30 subcellular structures enabled the definition of the proteomes of 13 major organelles. Exploration of the proteomes revealed single-cell variations in abundance or spatial distribution and localization of about half of the proteins to multiple compartments. This subcellular map can be used to refine existing protein-protein interaction networks and provides an important resource to deconvolute the highly complex architecture of the human cell.

Spatial partitioning of biological functions is a phenomenon that is fundamental to life. In humans, this spatial partitioning constitutes a hierarchy of specialized systems ranging across scales—from organs to specialized cells to subcellular structures, down to macromolecular complexes. At the cellular level, proteins function at specific times and subcellular locations, such as organelles. These locations provide a specific chemical environment and set of interaction partners that are necessary to fulfill the protein's function. Mislocalization of proteins can be associated with cellular dysfunction and disease (*1*, *2*). Thus, knowledge of the spatial distribution of proteins at a subcellular level is essential for understanding protein function, interactions, and cellular mechanisms.

Several approaches for systematic analysis of protein localizations have been described. Quantitative mass spectrometric readouts allow identification of proteins with similar distribution profiles across fractionation gradients (*3*–*7*) or proteins labeled by proximity-dependent enzymatic reactions in cells (*8*–*11*). In contrast, imaging-based approaches using tagged proteins (*12*–*14*) or affinity reagents (*15*, *16*) enable exploration of the subcellular distribution of proteins in situ in single cells and can also effectively identify cell-to-cell variability and multi-organelle distribution. Complementary to these experimental methods, a number of in silico approaches have been used to predict subcellular localization in eukaryotic cells [e.g., (*17*, *18*)]. The manually curated UniProt database (*19*) is an important resource for protein localization that collects subcellular data from literature and external databases for a large number of species. Despite these efforts, experimental data on subcellular localization are lacking for the majority of human proteins. To address this need, pilot studies have been initiated to probe human proteins by means of immunofluorescence (IF) and high-resolution confocal microscopy (*15*, *20*, *21*) and mass spectrometry (*7*). To date, maps of the subcellular proteome of murine stem cells (*6*), HeLa cells (*7*), and rat liver (*22*) are the best-characterized data sets for mammals.

Here we report the establishment of the Cell Atlas—a comprehensive, proteome-wide knowledge resource for subcellular localization in human cells—within the framework of the Human Protein Atlas (HPA) (*23*, *24*). By integration of transcriptomics data and an antibody-based image-profiling approach, we provide experimental localization data for 12,003 proteins, using a panel of 22 human cell lines and 13,993 antibodies. The spatial distribution of these proteins is resolved to 30 cellular structures and substructures, altogether representing 13 major organelles. Particular emphases were on defining the organelle proteomes and describing multilocalizing proteins and proteins displaying single-cell variability. We expect the availability of localization information for the human proteome to complement other systematic efforts on the DNA (*25*, *26*), RNA (*27*, *28*), and proteome (*19*, *29*) levels and aid in the molecular understanding of the human cell and its interactions.

Cell lines and transcriptomics analysis

The aim in creating the Cell Atlas was to define the proteomes of organelles and subcellular compartments by IF imaging (Fig. 1). To select suitable cell lines for the effort, transcriptomics analysis using RNA sequencing (RNA-seq) was performed on 56 human cell lines from various origins representing different germ layers and tissues (table S1). A hierarchical clustering analysis based on RNA-seq data (Fig. 2A) showed that cell lines of similar origin or phenotype clustered together, indicating a common pattern of gene expression. Prominent clusters included myeloid cell lines, lymphoid cell lines, endothelial cells, and cells immortalized by introduction of telomerase. Twenty-two cell lines were selected for IF imaging—together expressing 84% of all protein-coding genes (16,504 of 19,628) predicted by Ensembl [version 83.38 (*26*)]—based on a transcripts-per-million (TPM) cutoff of ≥1 (table S2). Interestingly, by applying TPM values, the average number of expressed genes in the sequenced cell lines was 11,490 (table S2), and the range spanned from 10,136 in Daudi cells (B lymphoblast) to 12,816 in SCLC-21H cells (small cell lung carcinoma). This is notably less than the previously measured average of ~14,000 transcripts obtained using FPKM values (fragments per kilobase of transcript per million mapped reads) as a normalization method. However, the TPM-based number corresponds more accurately to the number of proteins actually detected in this and other proteomic studies (*30*, *31*).

A classification of the RNA expression levels according to the principle previously described in (*24*) was performed to define genes expressed in all cell lines and those expressed in a cell line–restricted manner (fig. S1). About one-third (6295) of the protein-coding genes were expressed in all cell lines, suggesting a "housekeeping" role, whereas 45% showed a more variable expression. Eleven percent (2090) were not detected in any of the analyzed cell lines. Of these genes, 1225 were detected in tissues, suggesting that they code for proteins restricted to a smaller number of specialized cell types or representing specific developmental stages (table S3). Functional annotations from Gene Ontology (GO) support

[1]Science for Life Laboratory, School of Biotechnology, KTH Royal Institute of Technology, SE-171 21 Stockholm, Sweden. [2]Cambridge Centre for Proteomics, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK. [3]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. [4]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, SE-751 85 Uppsala, Sweden. [5]Computational Proteomics Unit, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QR, UK. [6]Department of Proteomics, School of Biotechnology, KTH Royal Institute of Technology, SE-106 91 Stockholm, Sweden. [7]Science for Life Laboratory, Department of Neuroscience, Karolinska Institute, SE-171 77 Stockholm, Sweden.
*These authors contributed equally to this work.
†**Corresponding author. Email: mathias.uhlen@scilifelab.se (M.U.); emma.lundberg@scilifelab.se (E.L.)**

Originally published 26 May 2017

this hypothesis, showing enrichment for tissue-restricted proteins, such as receptors in the sensory cells or reproduction-related proteins (table S4).

Creation of a subcellular map

As an integrated part of the HPA effort (23), antibodies have been generated, affinity-purified using the antigen, and validated by protein microarray analysis to ensure specific and selective binding to the intended target antigen (32). These antibodies cover the majority of all predicted human protein-coding genes. A systematic workflow for subcellular localization of proteins was established that uses IF and high-resolution confocal microscopy, as described in fig. S2 (15, 16). Altogether, 13,993 antibodies (13,073 antibodies generated by the HPA project, complemented with 920 commercially available antibodies) were selected to be included in the Cell Atlas after reliability analysis. Every antibody was used for immunostaining of the bone osteosarcoma–derived U-2 OS cell line and two additional cell lines from the panel showing a high expression of the target gene. In addition to the antibody of interest, reference markers outlining the nucleus, microtubules, and endoplasmic reticulum (ER) were included in each sample (fig. S3). For all proteins, the spatial expression patterns observed in the confocal images were assigned to one or more of 30 cellular organelles and substructures (Fig. 1 and table S5) and classified by a location-specific reliability score, as outlined below. The images and primary data are presented in the Cell Atlas in a gene-centric manner, including the classification of all images and a description of the validation and reliability of the antibodies and identified locations. Furthermore, the images were annotated by a citizen science approach through the Project Discovery platform within EVE Online, a massive multiplayer online game; more than 180,000 players across the world have contributed more than 7 million minutes of active participation to date (33). In total, the Cell Atlas (in version 16.1 of the HPA) contains 82,152 high-resolution annotated images covering 61% of all human protein-coding genes and 73% of the genes expressed in the IF cell line panel. The complete localization data set containing the results for all proteins in the Cell Atlas, as well as all successful stainings obtained in the different cell lines, are given in tables S6 and S7, respectively.

Validation of data in the Cell Atlas

Recently, there have been many articles questioning the quality and use of antibodies in research [e.g., (34, 35)]. Because off-target antibody binding can cause false-positive results, efforts have gone into manually annotating all antibodies regarding their reliability and quality of the staining. In the Cell Atlas, we provide a reliability score for every annotated location and protein on a four-tiered scale: “validated,” “supported,” “approved,” and “uncertain.” Locations obtained the score “validated” if the antibody was validated according to one of the validation “pillars” proposed by an international working group (36) as suitable for IF: (i) genetic methods using

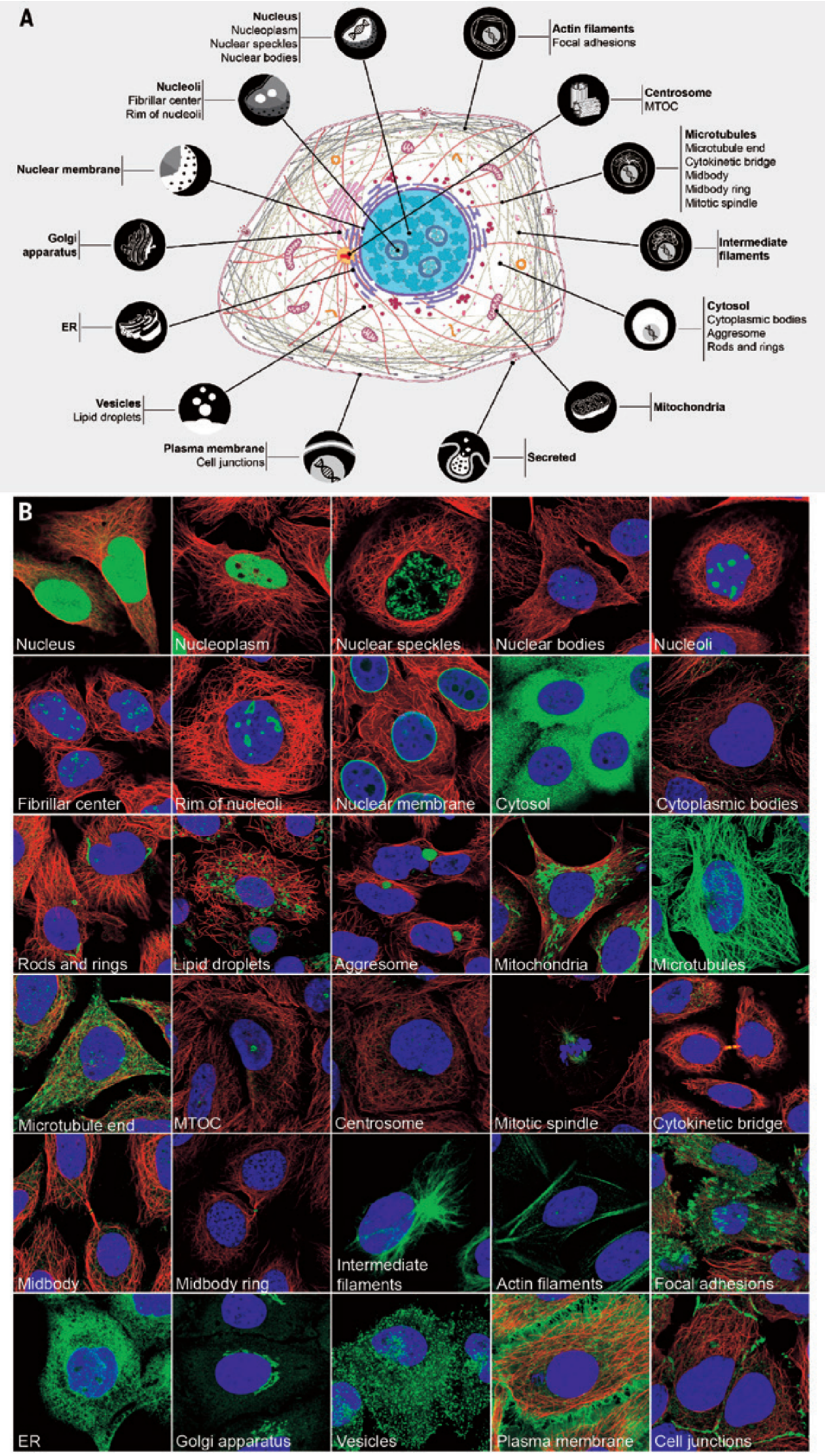


Fig. 1. Subcellular locations in the Cell Atlas. (**A**) Schematic overview of the cell. Thirteen subcellular proteomes, as well as a proteome of secreted proteins, were defined in the Cell Atlas by determining the localization of proteins to 30 subcellular structures. (**B**) Subcellular structures annotated in the Cell Atlas by immunofluorescence (IF) microscopy. Examples of proteins (green) localizing to each annotated structure in the representative set of human cell lines used in the Cell Atlas. Microtubules are marked with an antibody against tubulin (red); the nucleus is counterstained with DAPI (blue). The side of an image is 64 μm. Information about cell lines, antibodies, and proteins is given in table S6.

short interfering RNA (siRNA) silencing (37) or CRISPR-Cas9 knockout, (ii) expression of a fluorescent protein–tagged protein at endogenous levels (38), or (iii) independent antibodies targeting different epitopes (see fig. S4 for examples). The second tier, "supported" locations, is defined by agreement with external experimental data from the UniProt database. An "approved" location score indicates a lack of external experimental information about the protein location. Last, an "uncertain" location is contradictory to complementary information, such as literature or transcriptomics data. "Uncertain" locations are only shown when it cannot be ruled out that the data are correct. In fig. S5, the distributions of scores for all proteins are shown. Forty-three percent of the protein locations are in the top two tiers, representing a high degree of certainty in the results, and half of the proteins are in the "approved" category. Although these proteins have no external evidence to support their location, the antibodies passed our quality tests and showed a consistent IF staining. Nevertheless, the likelihood of false-positive results may be higher and should be taken into consideration when looking at individual proteins, whereas the effect on global proteomic analyses is negligible (fig. S6).

The human organelle proteomes

The spatial information provided by the IF images enabled the development of a subcellular map. The distribution of 12,003 proteins into 30 cellular compartments and substructures is shown in Fig. 2B and detailed in table S8. We were able to describe the proteomes for 13 major organelles. In addition, we defined a secretome containing proteins secreted through the classical pathway by combining three bioinformatic methods for signal peptide recognition with seven prediction methods for transmembrane regions (24), which indicated that 2918 proteins are secreted (table S9). Most proteins in the Cell Atlas were found in the nucleoplasm and its substructures (6245). The number of nuclear proteins considerably exceeds previously reported numbers. Although false nuclear localizations can be observed because of cross-reactivity of antibodies (21), the fraction of nuclear locations are similar in the higher- and lower-reliability tiers. The second largest number of proteins was identified in the cytosol (4279), followed by vesicles (1806), including transport vesicles and small membrane-bound organelles such as endosomes or peroxisomes. The nucleoli, including their fibrillar center, contained 1270 different proteins, which is a more diverse proteome than that of the mitochondria or Golgi apparatus, although nucleoli are more restricted in their known function. In total, we acquired subcellular experimental evidence for 5662 proteins (47%) lacking an experimentally determined GO term for a cellular compartment. Furthermore, we refined or confirmed subcellular locations for 6341 (53%) proteins already classified by experimentally determined GO terms (fig. S7).

We further investigated the enrichment of RNA classification categories for the defined organelle proteomes. Figure 2C shows that proteins located in the mitochondria, nucleus, nucleoli, and ER are more often expressed in all cell lines, which emphasizes their housekeeping role and important function for cellular survival. In contrast, proteins with RNA expression patterns categorized as "enriched" (expression in a cell line at least five times as high as in all other cell lines) and "enhanced" (expression in one or more lines five times as high as the mean expression across all cell lines) are more commonly secreted or located in the plasma membrane, vesicles, and cytoskeleton, which indicates that these compartments play important roles in intercellular communication and adaptation to the surrounding microenvironment. An analogous pattern was seen in the RNA class distribution across 59 human tissues (fig. S8), indicating general similarities in organelle organization between cell lines and tissues.

The goal of proteomic studies lies in the large-scale localization of previously uncharacterized proteins to achieve a complete picture of organelle function. IF images are particularly advantageous in the identification of protein constituents of compartments that are challenging to purify or have distinct substructures. For example, specialized domains within a compartment, such as cell junctions in the plasma membrane, are easily visible in IF—for example, in the case of the uncharacterized protein C4orf19 (Fig. 2D). Other compartments, such as the cytokinetic bridge, correspond to a rare cellular event and are thus challenging for proteomic studies. However, with our high-resolution images, we were able not only to identify 88 proteins located in the cytokinetic bridge (Fig. 2E), but also to analyze the underlying components midbody (36 proteins; Fig. 2F) and midbody ring (12 proteins; Fig. 2G). The detection of well-known constituents such as CHMP1B in the midbody, as well as less well-characterized proteins such as APC2 in the midbody ring or CCSAP in the cytokinetic bridge, provides an enhanced understanding of the final step of cell division. In nucleoli, we identified proteins such as MKI67 that are localized in the rim around the nucleolus and reorganize to line the condensed chromosomes during mitosis (Fig. 2H). As described below, additional tailored assays to complement the Cell Atlas further increase the available information about subcellular locations. The largely uncharacterized dynamic structure termed rods and rings (RR) previously had only three known members, including IMPDH1 and IMPDH2 (Fig. 2I) (39). We discovered and confirmed 21 RR candidates by actively inducing RR formation with the compound ribavirin (39). The assignment of additional proteins to the RR sheds new light on this structure and provides opportunities for better understanding its origin, composition, and function. In the nucleus, the PML body (marked by SP100; Fig. 2J) was a prominent substructure. This location can be further explored for selected proteins, because the Cell Atlas contains additional images generated by superresolution microscopy, allowing a distinction between proteins localizing to the surface (SP100; Fig. 2K, lower image) versus to the core (ZBTB8A; Fig. 2K, upper image) of the PML body.

Validation with other proteome-wide data sets

To evaluate the overall validity of our data, we assessed its agreement with functional protein information from independent proteome-wide databases. First, we performed a GO "biological process" term analysis of the proteome of each organelle. The significantly enriched terms were all related to known key processes of the respective organelle (table S10). Second, we analyzed the location enrichment of a set of proteins by a hypergeometric statistical test. In this manner, we could demonstrate that the nuclear receptors according to nucleaRDB (40) and their co-regulators as defined by the Nuclear Receptor Signaling Atlas (41) were enriched in the nucleus (Fig. 3A and fig. S9) and that the group of predicted secreted proteins were enriched in the organelles of the secretory pathway (Golgi apparatus, vesicles, and ER) (Fig. 3A). Third, enrichment tests with the mammalian complex database CORUM (42) showed similar results (Fig. 3A and fig. S9). Known complexes were significantly enriched in the respective organelle, with the exception of the cytoskeleton.

Validation by mass spectrometry

Proteome databases contain information about the subcellular localizations of already characterized proteins; however, our data set contains a large portion of proteins with a previously uncharacterized location. Therefore, we used an independent approach to reliably validate our annotations. The Cell Atlas data were compared with a high-resolution spatial protein map generated by a mass spectrometry–based method called hyperLOPIT (hyperplexed localization of organelle proteins by isotope tagging). HyperLOPIT aims to resolve all subcellular compartments in a single experiment by combining biochemical cell fractionation with quantitative mass spectrometry and robust multivariate statistical analysis (3, 6). This enables global identification and quantification of proteins and assignment to their respective subcellular compartments (43). The technique does not rely on absolute organelle purification but is based on the measurement of the distribution of cellular proteins across multiple density gradient fractions. Protein localization is assigned by comparing the distributions of proteins of unknown subcellular location with those of unambiguous organelle markers.

The hyperLOPIT approach was applied to create a subcellular map of the U-2 OS cell line. Spatial distribution profiles of 5020 proteins were determined, and a support vector machine was used to classify 1971 proteins to 12 discrete subcellular compartments, which were customized to match with the annotations in the Cell Atlas (Fig. 3B). Localization information for a total of 3626 proteins was available in both the Cell Atlas (U-2 OS only; table S11) and hyperLOPIT results (table S12). Of these, 1426 proteins were unambiguously

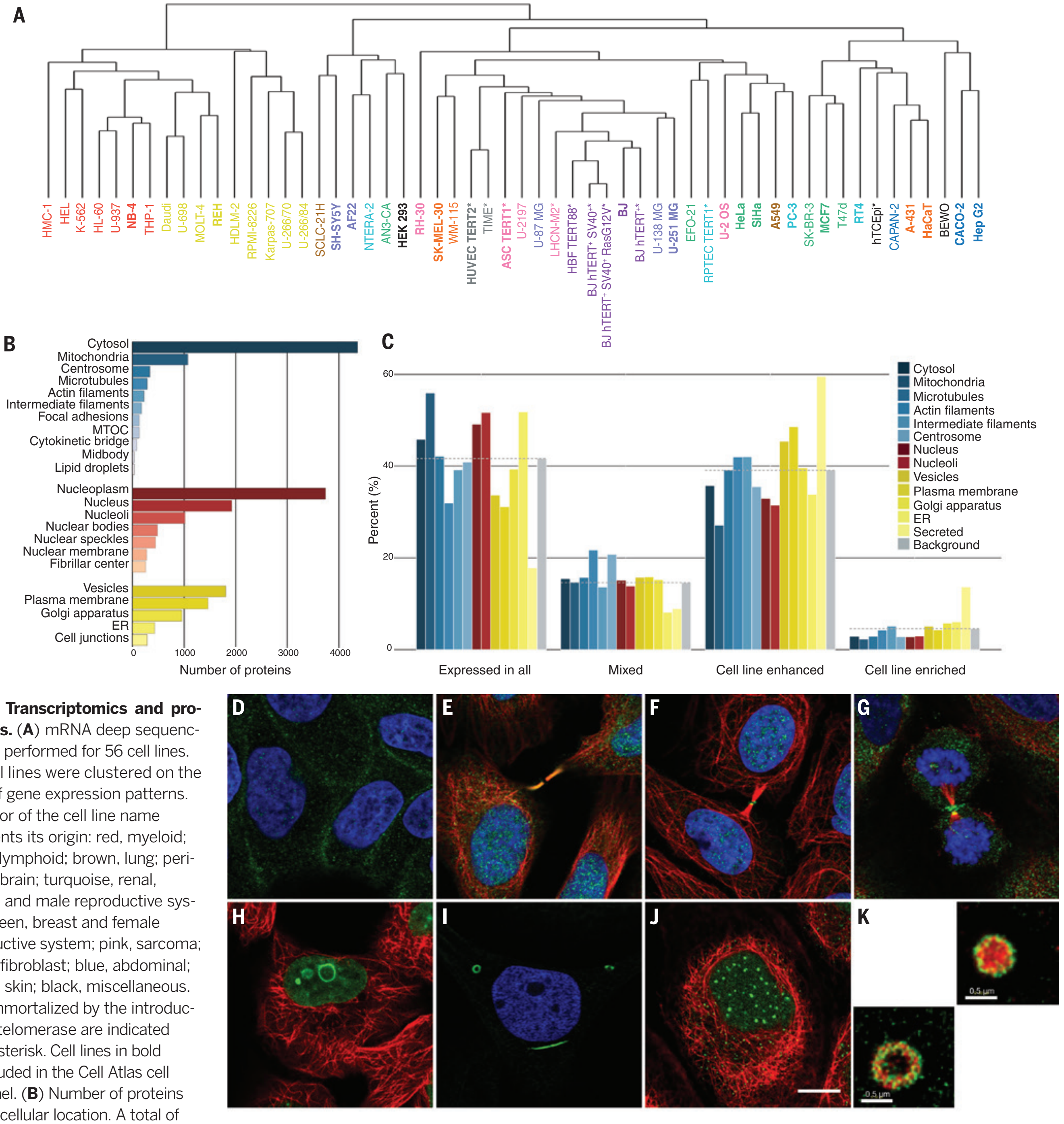


Fig. 2. Transcriptomics and proteomics. (**A**) mRNA deep sequencing was performed for 56 cell lines. The cell lines were clustered on the basis of gene expression patterns. The color of the cell line name represents its origin: red, myeloid; yellow, lymphoid; brown, lung; periwinkle, brain; turquoise, renal, urinary, and male reproductive system; green, breast and female reproductive system; pink, sarcoma; purple, fibroblast; blue, abdominal; orange, skin; black, miscellaneous. Cells immortalized by the introduction of telomerase are indicated by an asterisk. Cell lines in bold are included in the Cell Atlas cell line panel. (**B**) Number of proteins per subcellular location. A total of 12,003 proteins were localized to one or more subcellular compartments in this study. Locations are sorted and color-coded according to the number of proteins and the meta-compartments in which they occur [cytoplasm (cytosol and embedded organelles; shades of blue), nucleus (nuclear and nucleolar structures; shades of red), and secretory pathway (ER, Golgi apparatus, vesicles, and plasma membrane; shades of yellow)]. Some locations are merged: aggresomes and RR to cytosol, microtubule ends and mitotic spindle to microtubules, and midbody ring to midbody. (**C**) RNA classification categories per major organelle (nucleus and nuclear membrane are merged) compared with the background of genes in the Cell Atlas. Genes with a TPM value of ≥1 were considered as expressed and classified either as expressed in all cell lines, enriched (expression in one cell line at least fivefold as high as in all other cell lines), enhanced (average TPM level fivefold as high in one or more cell lines as the mean TPM of all cell lines), or mixed (expressed, but not in one of the other categories). (**D**) C4orf19 (detected by antibody HPA043458 in RT4 cells) localized to cell junctions, a subdomain of the plasma membrane. (**E** to **G**) Protein localization at the final stage of cytokinesis: (E) CCSAP to the cytokinetic bridge (detected by HPA028402 in U-2 OS cells) (E), CHMP1B to the midbody (detected by HPA061997 in SiHa cells) (F), and APC2 to the midbody ring (detected by HPA078002 in U-2 OS cells) (G). (**H**) MKI67 (detected by CAB000058 in U-251 MG cells) localized to the rim of nucleoli. (**I**) Previously uncharacterized protein C21orf59 (detected by CAB034170 in U-2 OS cells) localized to RR, whose formation was induced by ribavirin. (**J** and **K**) Conventional IF images in the Cell Atlas (J) and superresolution images acquired by stimulated emission depletion microscopy (K) of a PML body (a type of nuclear body) show the surface of the body marked by PML (red) and the shell protein SP100 (HPA016707, green) or the core protein ZBTB8A (HPA031768, green). Scale bars, 10 µm in (J) [applicable to (D) to (I)] and 0.05 µm in (K).

classified to a single location by hyperLOPIT. Within this group, 799 were also assigned a single location in the Cell Atlas, whereas the remaining 627 proteins had Cell Atlas annotations for more than one location.

Two comparisons between the data sets were performed: First, a comparison of proteins shown to be present in only one location in the Cell Atlas data ("unique match," table S13), and second, a comparison of all available proteins—including those shown to reside in more than one subcellular class in the Cell Atlas—with one unambiguous assignment in the hyperLOPIT data set ("partial match," table S13). Of the 799 proteins assigned by the Cell Atlas to a single location we found 76% agreement (unique match) with hyperLOPIT subcellular assignments. For the 1426 proteins common between the two data sets, 82% agreement (partial match) was observed between subcellular assignments. However, the overall agreement differed between the four reliability tiers of the Cell Atlas and was only 46% for the "approved" tier, which makes up 51% of the Cell Atlas data set (table S13). At the organelle level (table S13), the agreement ranged from 91 and 92% for the ER and mitochondria, respectively, to 60% for vesicles. This lower overlap is expected, because vesicles, as defined in the Cell Atlas, group together several organelles and structures that could be analyzed separately using hyperLOPIT. It is clear from the principal components analysis (PCA) shown in Fig. 3C that many Cell Atlas "vesicular" proteins reside in the unclassified intermediate area of the hyperLOPIT data set. Vesicles are highly dynamic structures that are generated in, and traffic between, different parts of the cell, and hence the steady-state location of their protein constituents is likely to involve multiple locations, which in the hyperLOPIT data would result in no single, unique classification. The hyperLOPIT workflow involves fractionation of chromatin-associated proteins and nucleoplasm and nucleolus, and this additional fractionation manifests itself as discrete protein correlation patterns. Interrogation of the hyperLOPIT data with Cell Atlas nuclear assignments revealed a nucleolar-like subcluster in the hyperLOPIT data; this demonstrates the power of combining data obtained using orthogonal methods (Fig. 3D).

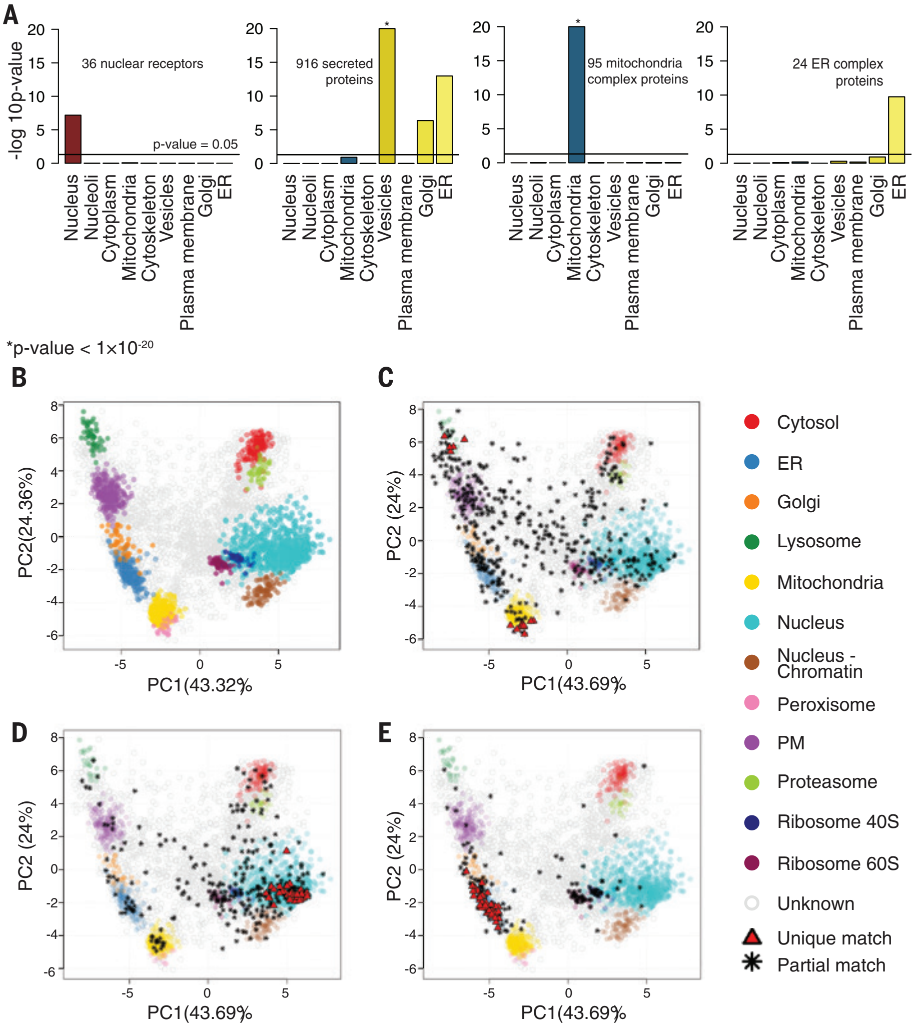


Fig. 3. Validation by proteome-wide databases and hyperLOPIT. (**A**) Location enrichment analyses of different protein sets. Hypergeometric tests were performed to evaluate subcellular locations ($P < 0.05$). Nuclear receptors were enriched in the nucleus meta-compartment. Predicted secreted proteins were enriched in organelles of the secretory pathway: ER, Golgi apparatus, and vesicles. Members of known complexes according to the CORUM database were enriched in the respective organelles—for instance, mitochondria and ER. Color-coding is as in Fig. 2. (**B**) A PCA representation of the human U-2 OS cell hyperLOPIT data (5020 proteins common across two hyperLOPIT replicates). One point represents one protein, and proteins cluster according to their density gradient distribution. Colored circles correspond to subcellular compartments that have been classified by a support vector machine. For the statistical comparison to the Cell Atlas, hyperLOPIT subcellular annotations were matched with their equivalent Cell Atlas definition. (**C** to **E**) PCA plots of the U-2 OS human data set for (C) vesicles, (D) nucleoli, and (E) the ER. Proteins occurring in both the Cell Atlas and hyperLOPIT data sets are visualized (3626 proteins). Black stars represent partial matches (a single assignment in the hyperLOPIT data, more than one in the HPA data set), and red triangles represent unique matches (a single assignment in both the HPA and hyperLOPIT data sets). PM, plasma membrane.

In the hyperLOPIT data set, 60% of the proteins identified fell into the "unclassified" category. This unclassified category may represent several dynamic scenarios, such as proteins localized to unannotated subcellular structures or multilocalizing proteins. A separate analysis was conducted for the 1755 proteins that were labeled by hyperLOPIT as "unclassified" but that contained subcellular information in the Cell Atlas (fig. S10). Interestingly, the majority of the hyperLOPIT-unclassified proteins were found in the HPA classes "nucleoplasm," "vesicles," "nucleoplasm and cytosol," and "plasma membrane and cytosol," reflecting the highly dynamic localization of the majority of cellular proteins.

To show the complementary nature of the Cell Atlas and hyperLOPIT for predicting subcellular location, we applied a transfer learning method ([44]) to integrate the two data sources. Transfer learning allows one to meaningfully integrate heterogeneous data. By combining labeled marker proteins common to both data sets, a significant increase in classifier accuracy was obtained (fig. S11) relative to that obtained using the Cell Atlas alone ($P < 2 \times 10^{-16}$). This highlights the strength of integrating the two approaches for the optimal classification of proteins to organelles.

Proteins localized to multiple compartments

In a pilot for this study ([15]), we concluded that many of the studied proteins are not restricted to a single organelle but rather localized to one or more additional locations. This observation is supported by the hyperLOPIT data described above and by data for yeast, in which 54.3% of the proteins were assigned to multiple localizations ([14]).

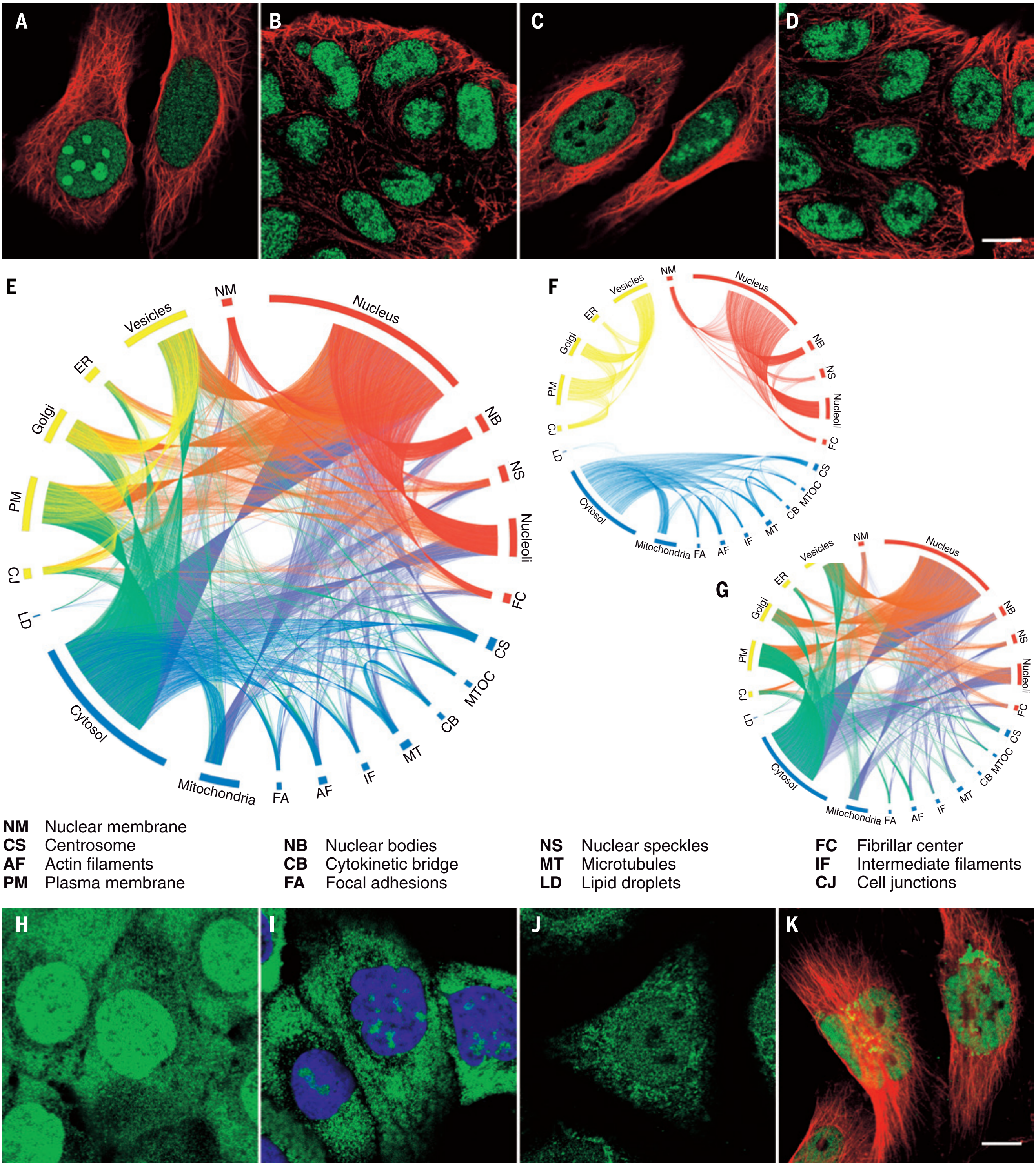


Fig. 4. Multilocalizing proteins in the human proteome. (**A** to **D**) ZNF554 is an example of a protein with a cell line–dependent subcellular localization. Two antibodies, HPA060247 [left, (A) and (B)] and HPA063358 [right, (C) and (D)], binding different epitopes detected ZNF554 in both the nucleoplasm and nucleoli in U-2 OS cells, whereas it was only detected in the nucleoplasm in RT4 and SH-SY5Y (not shown). The nucleolar expression was detected in just a fraction of the U-2 OS cells and thus additionally showed a single-cell variation. Scale bar, 10 μm. (**E** to **G**) Circular plots with the identified proteins of each compartment presented and sorted by meta-compartments. Multilocalizing proteins appearing more than once in the plots are connected by a line. Color-coding is as in Fig. 2, with secondary colors representing multilocalization across meta-compartments. The plots show (E) connections among all meta-compartments and proteins, (F) connections only within a meta-compartment, and (G) connections only across meta-compartments. (**H** to **K**) Examples of dual localizations: (H) UBE2L3 in nucleus and cytosol (detected by HPA062415 in A-431 cells), (I) 60S ribosomal protein L19 in nucleoli and cytosol (detected by HPA043014 in U-2 OS cells), (J) MTIF in nucleus and mitochondria (detected by HPA039791 in U-2 OS cells), and (K) CCAR1 in Golgi apparatus and nucleoplasm (detected by HPA007856 in U-251 MG cells). Scale bar, 10 μm.

One of the strengths of imaging-based spatial protein analysis is the ability to localize a protein in situ and simultaneously visualize protein distribution among multiple cellular structures, thus identifying multilocalizing proteins (MLPs). Here we have classified the main and additional locations for each protein on the basis of a clear difference either in the signal strength or in the occurrence across the tested cell lines. More than 50% (6163) of the proteins were detected at more than one location, of which 27% (1649) were detected at three or more locations (table S8). ER and mitochondria mainly contained specifically located proteins, whereas the proteomes of the plasma membrane and the nuclear substructures contained mainly MLPs, consistent with the hyperLOPIT data (Fig. 3E and fig. S12). This finding is consistent with the known biological function of the organelles. Whereas the proteome of the mitochondria is more self-contained, the nucleus, plasma membrane, and cytosol contain many proteins that operate across organelles to regulate metabolic reactions or gene expression or to transmit information from the surrounding environment. Also observed were MLPs that varied in their cell-to-cell spatial distribution, as well as MLPs such as ZNF554 that showed a cell line–dependent location, with different localization in the three cell lines tested (Fig. 4, A to D). In total, 3546 MLPs showed a cell line–dependent localization (table S14).

To investigate whether MLPs are organized in superordinate structures, we grouped the individual organelles and substructures into three meta-compartments—nucleus (nuclear and nucleolar structures), cytoplasm (cytosol, mitochondria, and the different types of cytoskeleton), and the secretory pathway (ER, Golgi apparatus, vesicles, and plasma membrane)—and searched for distinct patterns within and across these meta-compartments by aligning the proteins on a circular plot (Fig. 4, E to G). Within the cytoplasm meta-compartment, most MLPs appeared between the cytosol and the cytoskeletal structures and other organelles embedded in it (Fig. 4F). Similarly, most MLPs in the nucleus could be identified as a combination of nucleoplasm and the fine structures within, such as nucleoli or nuclear bodies, and likely reflect dynamic translocations of proteins between these proximal compartments (Fig. 4F). The MLPs in the secretory pathway exhibit a sequential pattern, likely reflecting the directional protein trafficking (Fig. 4F). This analysis was repeated with stratification according to the reliability of locations to control for the effect of data quality on our results (fig. S6). The patterns of multilocalization were highly similar regardless of the data set used.

Frequent patterns of multilocalization across meta-compartments included cytosol and nucleus, cytosol and nucleoli, and mitochondria and nucleoli (Fig. 4G). Enrichment analysis of GO "biological process" terms of these proteins (table S15) revealed that MLPs of the nucleus and the cytosol are related to transcription and cell cycle regulation, such as UBE2L3 (Fig. 4H); MLPs of the cytosol and nucleoli are enriched for ribosomal proteins, such as 60*S* ribosomal protein L19, which can be also found on the ER (Fig. 4I); and proteins found in both the mitochondria and nucleus are related to protein translation and cellular respiration, such as MTIF3 (Fig. 4J) and NDUFA9, respectively. Intriguingly, the meta-compartments secretory pathway and nucleus shared a very high number of MLPs, despite not being in direct physical contact with each other. These MLPs are characterized by their involvement in the regulation of transcription or cell cycle–dependent processes—for example, CCAR1 (Fig. 4K). This indicated that the proteomes of

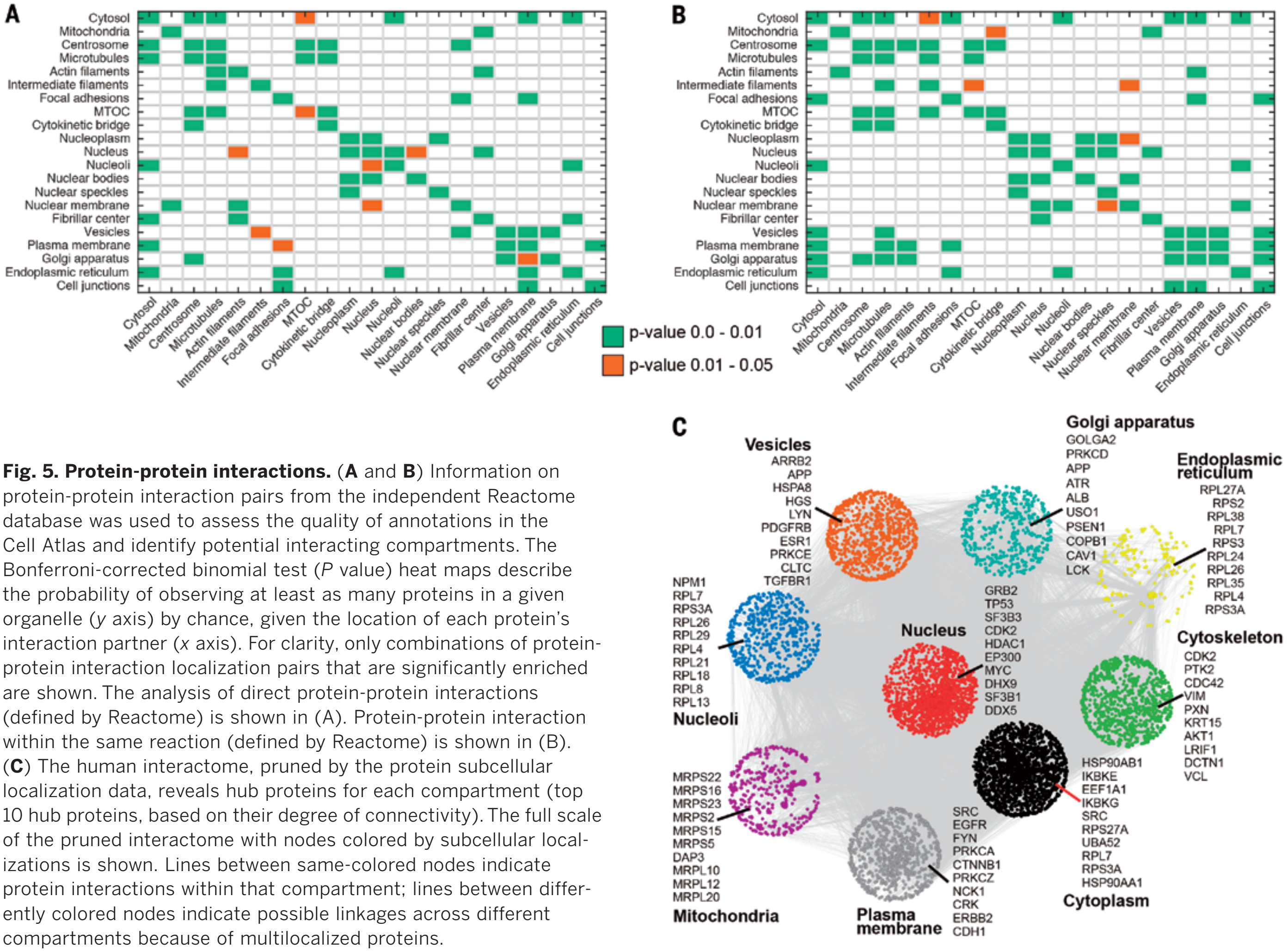


Fig. 5. Protein-protein interactions. (**A** and **B**) Information on protein-protein interaction pairs from the independent Reactome database was used to assess the quality of annotations in the Cell Atlas and identify potential interacting compartments. The Bonferroni-corrected binomial test (*P* value) heat maps describe the probability of observing at least as many proteins in a given organelle (*y* axis) by chance, given the location of each protein's interaction partner (*x* axis). For clarity, only combinations of protein-protein interaction localization pairs that are significantly enriched are shown. The analysis of direct protein-protein interactions (defined by Reactome) is shown in (A). Protein-protein interaction within the same reaction (defined by Reactome) is shown in (B). (**C**) The human interactome, pruned by the protein subcellular localization data, reveals hub proteins for each compartment (top 10 hub proteins, based on their degree of connectivity). The full scale of the pruned interactome with nodes colored by subcellular localizations is shown. Lines between same-colored nodes indicate protein interactions within that compartment; lines between differently colored nodes indicate possible linkages across different compartments because of multilocalized proteins.

the ER, Golgi apparatus, and vesicles are more functionally versatile and should not be reduced to their role in protein secretion. In fact, the MLPs create a range of interactions between functionally distant organelles and include them in a network of regulatory processes, which are primarily associated with the nucleus. This may be an indication of the complex network of events surrounding how the cell conveys signals from the exterior to the nucleus.

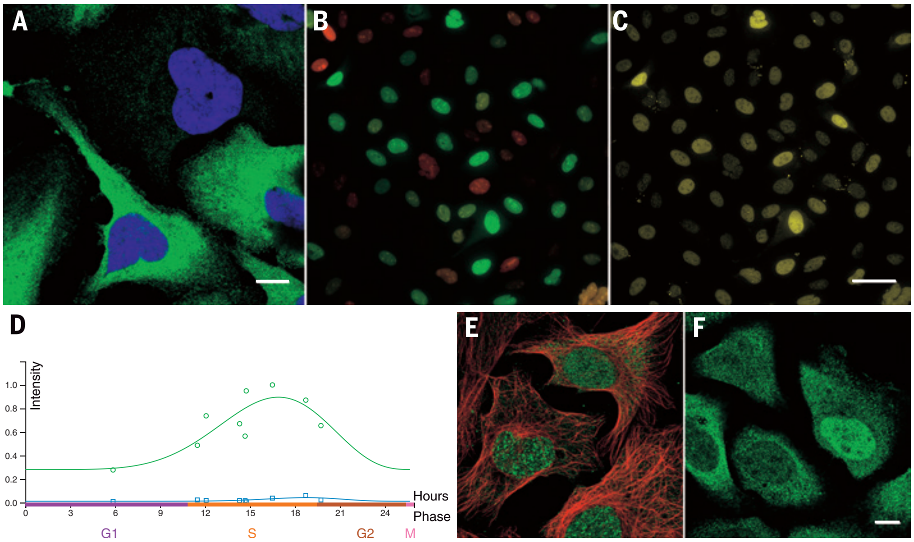


Fig. 6. Single-cell variation in protein expression. (**A**) CRYAB (detected by CAB002053 in U-2 OS cells) showed a single-cell variation in the cytosolic signal strength. (**B** and **C**) U-2 OS FUCCI cells expressed the cell cycle regulators CDT1 (red) during the G_1 phase and geminin (green) during the S and G_2 phases. An antibody targeting ANLN (yellow) stained only cells in the S and G_2 phases, marked by the green fluorescence. (**D**) Pattern of expression of ANLN across the cell cycle in U-2 OS cells by pseudo-temporal analysis using a time-regressive computational model. (**E**) The protein abundance of PCNA (detected by HPA030522 in U-2 OS cells) at nuclear bodies varied during the cell cycle. (**F**) PSMC6 (detected by HPA042823 in U-2 OS cells) changed its spatial distribution from nucleoplasm to cytosol during the cell cycle, based on data from U-2 OS FUCCI cells. Scale bars, 10 μm in (A) and (F) [applies to (E)] and 50 μm in (C) [applies to (B)].

Spatial information refines biological networks

The biological function of an organelle is not only defined by the presence or absence of proteins, but also by its underlying chain of reactions, which in turn are often conducted by protein-protein interactions. We used the spatial information of the Cell Atlas to examine the relationship between protein interaction partners. For every annotated structure in the Cell Atlas, we investigated the subcellular locations for the direct protein interaction partners, according to the Reactome database (*45*). Figure 5A shows a heat map of the probability that proteins in one cellular compartment interact directly with proteins in the same or other compartments. Within this stringent constraint, the majority of the significant enrichments ($P < 0.05$) for an interaction pair were found within the same organelle. This compartmental enrichment was even observed for small structures such as nuclear bodies and nucleoli fibrillar centers. The exception was the microtubule-organizing center (MTOC), which showed significant enrichment for interactors found in the centrosome and microtubules. For some structures, proximal structures were also found to be significantly enriched. Proteins in the plasma membrane, for example, showed increased probability of directly interacting with proteins in the plasma membrane, cell junctions, Golgi apparatus, vesicles, focal adhesions, and cytosol. These results support the quality of the locations annotated in the Cell Atlas, given that direct protein-protein interactions occur in the same or connected compartments. To explore how cellular signaling expands across cellular compartments through reaction pathways, the same analysis was performed for the organelle proteomes, looking at protein interactions within reaction pathways defined by Reactome (Fig. 5B). In this analysis, the meta-compartments became more prominent, especially in terms of interactions between the organelles of the secretory pathway and signaling between compartments. Unexpected cross-talk between compartments included apparent interactions between the cytokinetic bridge and nuclear bodies.

We examined whether existing protein-protein interaction networks would benefit from a more comprehensive annotation of a protein's subcellular location, given that it constrains the possible number of interaction partners. The localization data was integrated, as spatial boundaries, into the human interactome that was recently used to systemically uncover the molecular background of human diseases (*46*). The interactome included annotations for 79,020 interactions of 7827 proteins. By taking the subcellular main location into account, the number decreased to 51,885 (65.7%) interactions of 6985 proteins that were found to be likely to occur in vivo (fig. S13). However, a substantial number of protein interactions were found when additional locations were included, raising the total to 62,352 (78.9%) interactions of 7494 proteins (fig. S13). This further supports the important functional role of MLPs. With this new location-pruned interaction data set, we generated a map to identify the most connective proteins, also called hub proteins, of each compartment (Fig. 5C). The hub proteins of each compartment were mostly different from hubs of the original, nonannotated interactome (table S16); hence, our data set led to the identification of previously unrecognized driver genes within the network. The localization-annotated interactome is available in table S17.

Single-cell variations in protein expression

Protein profiling by IF microscopy allows analysis of expression patterns on a single-cell level to reveal variations in a protein's expression across the analyzed cells. In the Cell Atlas, we labeled proteins with an observed single-cell variation (SCV), such as the nucleolar localization of ZNF554 (Fig. 4, A and C). SCV can be observed either in protein expression levels (IF signal intensity) or in the spatial distribution pattern. Of the 12,003 detected proteins, 1855 (15%) showed a SCV (table S18). Further studies are needed to reveal whether SCV is due to dynamic protein regulation or stochastic events. The majority of these proteins showed a variation in protein expression levels (1671)—for example, CRYAB (Fig. 6A)—whereas 222 proteins showed a variation in spatial distribution (38 proteins fall into both categories). The organelles with the most SCV proteins were the cytosol (394), nucleoplasm (381), nucleoli (230), and mitochondria (206) (table S8)—organelles that also contain most known cell cycle–dependent proteins.

In addition to being related to the subcellular structures that only appear during cell division (mitotic spindle, cytokinetic bridge, midbody, and midbody ring), it is plausible to expect a majority of these SCVs to also be related to the cell cycle, because the cells in the images were growing under asynchronous conditions. To confirm this, we used two approaches for a subset of the proteins. First, we stained selected proteins with an observed SCV in the U-2 OS FUCCI [fluorescence ubiquitination cell cycle indicator (*47*)] cell line, which allows monitoring of the cell cycle; by this method, we verified a cell cycle–dependent expression of 64 proteins, including, for example, ANLN (Fig. 6, B and C; see the list of proteins in table S19). The second approach used a computational model to infer the cell cycle position on the basis of features of the microtubule and nucleus reference markers. In this manner, the cell cycle position of the cells in the images could be determined in a continuous model, and a pseudotemporal reconstruction allowed the pattern of cell cycle dependency to be modeled. Figure 6D shows such a plot for ANLN, which is expressed in cells in the S and G_2 phases, according to both FUCCI colocalization and the pseudo-temporal computational modeling. Like for SCV, cell cycle–dependent variation could be observed either in a change of the intensity—for example, in the case of PCNA (Fig. 6E)—or in a change of the localization, illustrated by the translocation of PSMC6 from nucleoplasm to cytosol (Fig. 6F).

Discussion

Here we present the most comprehensive map of the subcellular distribution of the human proteome, generated by high-resolution IF images on a single-cell level. The results are presented in an interactive resource, the Cell Atlas, as part of the Human Protein Atlas (www.proteinatlas.org). This allows exploration of the organelle proteomes, their substructures, single-location and multilocalizing proteins, and proteins exhibiting single-cell variations in expression or cell cycle–dependent expression. These defined categories can furthermore be explored in terms of gene expression patterns across a multitude of cell lines on the basis of transcriptome data. To facilitate integration with other biological resources, all data are available for download from the Human Protein Atlas and through collaborations with efforts such as UniProt (19), NextProt (29), GO (48), and the pan-European ELIXIR project (49).

Spatial partitioning of biological reactions by compartmentalization is an important cellular mechanism for allowing multiple cellular reactions to occur in parallel while avoiding cross-talk. Intriguingly, we identified more than 50% of the analyzed proteins as localizing to more than one compartment at the same time. The fact that proteins are localized at multiple sites increases the complexity of the cell from a systems perspective. On one level, it can function as a spatial confinement to control the timing of the molecular function in the designated compartment. On another level, multilocalizing proteins are more prone to have diverse protein-protein interactions because of an increased number of potential interaction partners. This is of particular relevance for network analyses and the identification of key hub proteins that play a crucial role in linking complexes to smaller subnetworks, leading to a cellular-wide network. Moreover, proteins that localize to more than one compartment may have context-specific functions, increasing the functionality of the proteome. The fact that proteins "moonlight" in different parts of the cell is now well accepted (50, 51). The high percentage of proteins in multiple locations, as indicated by the complementary IF and hyperLOPIT data sets, may be an indicator of the scale on which moonlighting occurs. The more complex a system is, the greater the number of parts that must be sustained in their proper place, and the lesser the tolerance for errors; therefore, a high degree of regulation and control is required. To understand cellular function, and particularly in the context of health and disease, detailed knowledge about the cellular system is needed. We demonstrated that current network models benefit from integration of the Cell Atlas localization data as spatial boundaries to remove false-positive interactions.

The proteome of a single cell is compartmentalized and spatiotemporally regulated to a high degree. Protein expression and localization change over time and enable the cell to react to intrinsic or extrinsic factors. Although only presenting a snapshot of the current state of a few cells, our single-cell analysis gives insight into this dynamic process. The high-resolution map of the subcellular localization of 12,003 human proteins provided by the Cell Atlas is a key resource for a comprehensive understanding of the human cell and its complex underlying molecular machinery, as well as a major step toward modeling the human cell.

Material and methods

Tissue culture cell line cultivation

All cell lines were cultivated at 37°C in a 5% CO_2 humidified environment in the following growth media: Roswell Park Memorial Institute medium (A-431, REH, RH-30, SiHa, SK-MEL-30; Sigma-Aldrich); Dulbecco's Modified Eagle Medium (A549, BJ, HaCaT, HeLa, NTERA, SH-S5Y5; Sigma-Aldrich); Eagle's Minimal Essential Medium (CACO-2, HEK293, HepG2, MCF-7, U-251 MG; Sigma-Aldrich); McCoy's 5A modified (RT-4, U-2 OS; Sigma-Aldrich). Media were always supplemented with 10% fetal bovine serum (FBS, Sigma-Aldrich); additional cell line-specific supplements were: 1% non-essential amino acids (CACO-2, HeLa, HRK293, HepG2, MCF-7), 1% L-glutamine (CACO-2, HaCaT, HepG2, MCF-7, NTERA, RT-4, U-2 OS), 5% horse serum (NTERA). No antibiotics were used.

AF22 cells were kindly provided by A. Falk. They were cultivated in DMEM/F12 supplemented with N-2 (Cat#17502048, Thermo Fisher) and Pen/Strep (Sigma-Aldrich), with freshly added B-27 (1:1000, Cat#12587010, Thermo Fisher), EGF (10 ng/ml, AF-100-15, PeproTech) and FGF2 (10 ng/ml, 100-18B, PeproTech), flask and plates were coated in two steps with poly-N-ornithine (Sigma-Aldrich) and laminin (Sigma-Aldrich). Telomerase-immortalized cell line HUVEC/TERT2 (Cat# MHT-006-2) and ASC/TERT1 (Cat# MHS-001) were a kind gift by Evercyte GmbH, Vienna, Austria, and were cultured in EndoUp2 and AdipoUp, respectively. U2-OS FUCCI cells were developed and kindly provided by A. Miyawaki (47). The cells were cultivated in McCoy's 5A modified medium supplemented with 1% L-glutamine and 10% FBS. HeLa-Kyoto cell lines stably expressing an enhanced green fluorescent protein (EGFP)–tagged protein encoded on Bacterial Artificial Chromosome (BAC) were a kind gift from A. Hyman, Max Planck Institute Dresden, Germany, and were cultivated as described in Skogs *et al.* (38, 46). CRISPR-Cas9 knockout and GFP-expressing cells were a kind gift by Horizon Discovery, Cambridge, UK. Their designed HAP1 cell lines were cultivated in IMDM (Iscove's Modified Dulbecco's Medium, Sigma-Aldrich) media supplemented with 10% FBS and 1% Pen/Strep. All cells were harvested at 60 to 70% confluency by trypsinization (Trypsin-EDTA solution from Sigma-Aldrich) for splitting or preparing in glass bottom plates.

Antibodies

All antibodies generated and validated within the HPA project were rabbit polyclonal antibodies. They were designed to bind specifically to as many isoforms of the target protein as possible. The antigens consisted of recombinant protein epitope signature tags (PrEST) with a typical length between 50 and 100 amino acids (52). The resulting antibodies were affinity purified using the antigen as affinity ligand (32). All antibodies used were first approved for sensitivity and lack of cross-reactivity to other proteins, on arrays consisting of glass slides with spotted PrEST fragments. Commercial antibodies were provided by the suppliers and used according to the supplier's recommendations.

Sample preparation for indirect immunofluorescence

A standardized protocol optimized for proteome-wide immunofluorescence localization studies was used, which has previously been described in detail by Stadler *et al.* (16). Briefly, cells were seeded in 96-well glass bottom plates (Whatman, Cat# 7716-2370, GE Healthcare, UK, and Greiner Sensoplate Plus, Cat# 655892, Greiner Bio-One, Germany) coated with fibronectin (VWR, Sigma-Aldrich) and grown to a confluency of 60 to 70% (log-phase growth). PBS-washed cells were fixed in 4% paraformaldehyde (PFA) in growth media supplemented with 10% FBS for 15 min, followed by permeabilization with 0.1% Triton X-100 in PBS for 3×5 min. After a washing step with PBS, cells were incubated with the primary antibody overnight at 4°C. Rabbit polyclonal HPA antibodies were diluted to 2 to 4 μg/ml in blocking buffer (PBS with 4% FBS) containing 1 μg/ml mouse anti-tubulin (Abcam, ab7291, RRID:AB_2241126, Cambridge, UK), and 1 μg/mL chicken anti-calreticulin (Abcam, ab14234, RRID:AB_2228460) or rat anti-KDEL antibody (MAC 256) (Abcam, ab50601, RRID:AB_880636), respectively. On the next day after 4×10 min washes with PBS, the cells were incubated for 90 min at room temperature with the following secondary antibodies (all from ThermoFisher Scientific) diluted to 1 μg/ml in blocking buffer: goat anti-rabbit AlexaFluor 488 (A11034, RRID:AB_2576217), goat anti-mouse AlexaFluor 555 (A21424, RRID:AB_2535845), and goat anti-chicken AlexaFlour 647 (A-21449, RRID:AB_2535866), or goat anti-rat AlexaFluor 647 (A21247, RRID:AB_1056356), respectively. Cells were subsequently counterstained with 4′,6-diamidino-2-phenylindole (DAPI) for 10 min. After washing with PBS, the wells were completely filled with 78% glycerol in PBS and sealed.

Fluorescence image acquisition

Fluorescent images were acquired with a Leica SP5 confocal microscope (DM6000CS) equipped with a 63× HCX PL APO 1.40 oil CS objective (Leica Microsystems, Mannheim, Germany). The settings for each image were as follows: Pinhole

1 Airy unit, 16-bit acquisition, and a pixel size of 0.08 μm. The detector gain measuring the signal of each antibody was adjusted to a maximum of 800 V to avoid strong background noise. The majority of the images were acquired manually from at least two representative field-of-views (FOVs). For proteins displaying single cell variations in their expression pattern, at least six different FOVs were acquired. A small part of the plates were imaged automatically using the MatrixScreener M3 in LAS AF software (Leica Microsystem, Mannheim, Germany). Here, *z*-stacks at six FOVs were acquired and afterward two images were manually selected for display in the Cell Atlas. All images on the Cell Atlas are unprocessed with a small compression due to conversion from TIFF to JPEG file format.

IF image annotation

The subcellular location of each protein was manually determined based on the signal pattern and relation to the markers for nucleus (DAPI), microtubules, and endoplasmic reticulum. The annotated locations were as follows: actin filaments, aggresome, cell junctions, centrosome, cytokinetic bridge, cytoplasmic bodies, cytosol, endoplasmic reticulum, focal adhesions, Golgi apparatus, intermediate filaments, lipid droplets, microtubule organizing center (MTOC), microtubules, microtubule ends, midbody, midbody ring, mitochondria, mitotic spindle, nuclear bodies, nuclear membrane, nuclear speckles, nucleolar fibrillar center, nucleolar rim, nucleoli, nucleoplasm, nucleus, plasma membrane, rods and rings, and vesicles. If more than one location was detected, they were defined as main or additional location depending on the relative signal strength between the location and the most common location when including all cell lines. Variation between single cells were annotated either as a variation in the intensity or spatial distribution based on a visual inspection. The staining was not annotated if considered negative or unspecific.

Prediction of the human secretome

For the prediction of the human secretome, the analysis was performed as previously described (*24*). Briefly, a majority decision approach was used based on results from three methods for the prediction of signal peptides (SP): SignalP4.0 (*53*), Phobius (*54*), and SPOCTOPUS (*55*). SignalP4.0 is solely focused on the prediction of SPs whereas the two latter combine the prediction of transmembrane (TM) segments and SPs. In addition, results from the prediction of the human membrane proteome (*56*) were included to classify proteins with a predicted SP as well as one or more TM regions as membrane-spanning. The resulting list of potentially secreted proteins consists of all proteins with a predicted signal peptide by two out of three methods and not including a predicted TM region.

Classification of location reliability

Detected locations were classified based on the reliability of the antibodies and their respective stainings. A score was used for the classification, which incorporated several factors: reproducibility of the antibody staining in different cell lines (also taken in account when the signal strength correlates with RNA expression); reproducibility of the staining using antibodies binding to different epitopes on the target protein; validation data for the specificity of the antibody (knockdown by siRNA or CRISPR-Cas9 knockout mutants, matching signal with fluorescent-tagged protein); experimental evidence for location described in literature. There were also soft factors such as antibody validation by non–IF-related methods such as Western blot or immunohistochemistry. The final score led either to the failing of the antibody (~50% of all tested antibodies failed) or to the assignment into one of the following four classes: (i) "validated," if at least one antibody is validated—for example, two independent antibodies show the same localization, that was also observed in experiments outside the HPA or it was supported by, e.g., siRNA silencing; (ii) "supported," if there is external experimental data for the location; (iii) "approved," if the localization of the protein has not been previously described and was detected by only one antibody without additional validation; and (iv) "uncertain," if the antibody staining is contradictory to experimental data or no expression is detected on the RNA level.

RNA sequencing

Cell lines were selected for IF imaging based on RNA expression of genes (*57*). RNA was extracted from the cells using the RNeasy kit (Qiagen), generating high-quality total RNA (i.e., RIN > 8) that was used as input material for library construction with Illumina TruSeq Stranded mRNA reagents. Duplicate samples were sequenced on the Illumina HiSeq2500 platform. Raw sequences were mapped to the human reference genome GrCh38 and further quantified using the Kallisto software (*58*) to generate normalized transcript per million (TPM) values. TPM values for genes were generated by summing up TPM values for the corresponding transcripts generated by Kallisto. Genes with a TPM value ≥1 were considered expressed.

Location enrichment of protein sets by hypergeometric test

Enrichment of a group of proteins in subcellular locations was examined by hypergeometric tests. In each subcellular location enrichment test, only proteins with subcellular location annotated were considered. Predicted secreted proteins were collected from the HPA (*24*), nuclear receptors from nucleaRDB (*40*), nuclear receptor co-regulators from nuclear receptor signaling atlas (*41*), and subcellular location-specific protein complexes from CORUM (*42*). In CORUM database, nuclear complex proteins were taken from a term "nucleus" in the database; nucleoli complex proteins from "nucleolus"; cytoskeleton complex proteins from "actin cytoskeleton," "microtubule cytoskeleton," and "centrosome" complexes; mitochondria complex proteins from "mitochondrion"; vesicle complex proteins from "intracellular transport vesicle," "peroxisome," and "vacuole or lysosome"; ER complex proteins from "endoplasmic reticulum"; plasma membrane complex proteins from "plasma membrane/membrane attached" and "cell junction"; and cytoplasm complex proteins from "cytoplasm."

HyperLOPIT comparison with Cell Atlas annotations

To compare the subcellular assignments by both methods it was necessary to match the 12 subcellular organelle definitions used by hyperLOPIT to the 30 image categories defined in the Cell Atlas. The comparison was broken down into the following subclasses: all Cell Atlas subnuclear categories ("nucleus," "nucleoplasm," "nuclear speckles," "nuclear bodies," "nucleoli," "nucleoli fibrillar center," and "nuclear membrane") were individually compared with a single hyperLOPIT nuclear class encompassing both hyperLOPIT terms "nucleus" and "nuclear chromatin"; the Cell Atlas term for "vesicles" was compared with the combined hyperLOPIT terms for "lysosome" and "peroxisome" (consistent with the Cell Atlas definition for vesicles); and the Cell Atlas class "cell junctions" was compared with the hyperLOPIT term "plasma membrane." For the Cell Atlas terms called "plasma membrane," "mitochondria," "endoplasmic reticulum," "Golgi apparatus" and "cytosol/cytoplasm," the same terms are also available for hyperLOPIT and thus a direct comparison was performed. Proteins that were assigned by hyperLOPIT to the large protein complexes such as ribosomal subunits and proteasome were excluded from the comparison. A detailed description of the hyperLOPIT approach is provided in the supplementary materials.

Heat maps for protein-protein interaction

Protein-protein interaction pairs were obtained from the independent Reactome database (downloaded 20 September 2016) (*45*). A binomial test was used to compare the observed frequency of a target protein (Protein B) localizing to a given compartment with the expected frequency based on all annotations in the Cell Atlas. Here, the likelihood of localizations of the first protein in the pair (Protein A) can be ignored, as under the null hypothesis it has no impact on the localization of Protein B. The test therefore becomes the probability that we observe at least as many instances of Protein B in a specific compartment given the number of "tries" (instances of Protein A) and the background distribution of proteins over the locations in the Cell Atlas. The background distribution of locations was constructed by taking the frequency of each annotated location for proteins in in the Cell Atlas over the total number of proteins annotated in the Cell Atlas.

The results of the test were visualized using a heat map of *P* values (Fig. 5, A and B) where rows represent the location of Protein A and columns represent the location of Protein B. Values are therefore the probability of seeing Protein B in the given compartment at least as frequently as it was actually observed assuming the background distribution. The Bonferroni multiple-hypothesis correction applied per-row to correct for the number of locations being tested for in each pairing. By then considering the correlation of the protein-protein interaction pair locations, key insights into the nature and quality of the data in the Cell Atlas can be gained.

The Reactome database contains several types of protein-protein interactions that can be used to assess different properties of the Cell Atlas annotations. To assess the quality of annotation, we first analyzed direct interactions reasoning that interacting proteins must occupy the same physical space at some point in the cell cycle and therefore should be localized either to the same compartment or adjacent compartments (Fig. 5A).

The same analysis was further performed for protein pairs listed as belonging to the same reaction pathway as defined by the Reactome database to assess what compartments potentially interact through signal cascades (Fig. 5B). This analysis was created using MATLAB2016a.

Figure generation

Plots were generated using R studio (v. 3.3.1) and the additional ggplot2 package. The cell line hierarchical clustering was based on the Spearman correlation of the RNA sequencing data for each cell line. The average distance was used to determine the hierarchical clusters and visualized then by the R package ggdendro. The circular plots showing distribution of multilocalizing proteins were created using the Circos software (v. 0.69) ([59]). The image montages were created using FIJI ImageJ (v. 2.0.0-rc-49/1.51f).

Gene Ontology terms and functional enrichment

To check the overlap with GO annotations for proteins in the Cell Atlas, the web-based tool QuickGO ([60]) was used to acquire GO annotations for all genes using filters for cellular component and information source (downloaded 15 February 2017). The GO annotations based on data from the Cell Atlas were removed, and the Ensembl IDs for all Cell Atlas genes were then used for checking the overlap of genes with experimental evidence for any GO annotation. The functional annotation clustering for the genes not expressed in the Cell Atlas cell line panel was performed using the web based tool DAVID (Database for Annotation, Visualization, and Integrated Discovery v. 6.8) ([61]). All human genes were used as a background and the GO domain "biological process" terms with Bonferroni value of less than 0.01 were regarded as significantly enriched.

Location-pruned protein-protein interactions

Proteins interactions were obtained from published protein interactome data ([46]); among those protein interactions, only interactions with "signaling," "kinase," "complex," "literature," and "binary" types were taken; this indicates direct protein interactions. Those protein interactions were pruned to proteins localized in the same subcellular locations, in either cytoplasm or plasma membrane, or in either cytoplasm or cytoskeleton. Location-pruned protein interactions were visualized (Fig. 5C) through the edge-weighted spring embedded layout of Cytoscape ([62]) and their nodes were colored by the least frequent one of subcellular locations they have. In each subcellular location, hub proteins from protein interactions of given subcellular locations were examined based on their degree connectivity.

REFERENCES AND NOTES

1. K. Laurila, M. Vihinen, Prediction of disease-related mutations affecting protein localization. *BMC Genomics* **10**, 122 (2009). doi: 10.1186/1471-2164-10-122; pmid: 19309509
2. S. Park *et al*., Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol. Syst. Biol.* **7**, 494 (2011). doi: 10.1038/msb.2011.29; pmid: 21613983
3. T. P. Dunkley, R. Watson, J. L. Griffin, P. Dupree, K. S. Lilley, Localization of organelle proteins by isotope tagging (LOPIT). *Mol. Cell. Proteomics* **3**, 1128–1134 (2004). doi: 10.1074/mcp.T400009-MCP200; pmid: 15295017
4. L. J. Foster *et al*., A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187–199 (2006). doi: 10.1016/j.cell.2006.03.022; pmid: 16615899
5. L. Jakobsen *et al*., Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. *EMBO J.* **30**, 1520–1535 (2011). doi: 10.1038/emboj.2011.63; pmid: 21399614
6. A. Christoforou *et al*., A draft map of the mouse pluripotent stem cell spatial proteome. *Nat. Commun.* **7**, 9992 (2016). doi: 10.1038/ncomms9992; pmid: 26754106
7. D. N. Itzhak, S. Tyanova, J. Cox, G. H. Borner, Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, e16950 (2016). doi: 10.7554/eLife.16950; pmid: 27278775
8. K. J. Roux, D. I. Kim, B. Burke, BioID: A screen for protein-protein interactions. *Curr. Protoc. Protein Sci.* **74**, 19.23.1–19.23.14 (2013). pmid: 24510646
9. H.-W. Rhee *et al*., Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328–1331 (2013). doi: 10.1126/science.1230593; pmid: 23371551
10. V. Hung *et al*., Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging. *Mol. Cell* **55**, 332–341 (2014). doi: 10.1016/j.molcel.2014.06.003; pmid: 25002142
11. S.-Y. Lee *et al*., APEX fingerprinting reveals the subcellular localization of proteins of interest. *Cell Rep.* **15**, 1837–1847 (2016). doi: 10.1016/j.celrep.2016.04.064; pmid: 27184847
12. J. C. Simpson, R. Wellenreuther, A. Poustka, R. Pepperkok, S. Wiemann, Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **1**, 287–292 (2000). doi: 10.1093/embo-reports/kvd058; pmid: 11256614
13. W.-K. Huh *et al*., Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003). doi: 10.1038/nature02026; pmid: 14562095
14. Y. T. Chong *et al*., Yeast proteome dynamics from single cell imaging and automated analysis. *Cell* **161**, 1413–1424 (2015). doi: 10.1016/j.cell.2015.04.051; pmid: 26046442
15. L. Barbe *et al*., Toward a confocal subcellular atlas of the human proteome. *Mol. Cell. Proteomics* **7**, 499–508 (2008). doi: 10.1074/mcp.M700325-MCP200; pmid: 18029348
16. C. Stadler, M. Skogs, H. Brismar, M. Uhlén, E. Lundberg, A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* **73**, 1067–1078 (2010). doi: 10.1016/j.jprot.2009.10.012; pmid: 19896565
17. P. Horton *et al*., WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* **35**, W585–W587 (2007). doi: 10.1093/nar/gkm259; pmid: 17517783
18. K. C. Chou, Z. C. Wu, X. Xiao, iLoc-Hum: Using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. *Mol. Biosyst.* **8**, 629–641 (2012). doi: 10.1039/C1MB05420A; pmid: 22134333
19. UniProt Consortium, UniProt: A hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015). doi: 10.1093/nar/gku989; pmid: 25348405
20. L. Fagerberg *et al*., Mapping the subcellular protein distribution in three human cell lines. *J. Proteome Res.* **10**, 3766–3777 (2011). doi: 10.1021/pr200379a; pmid: 21675716
21. C. Stadler *et al*., Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells. *Nat. Methods* **10**, 315–323 (2013). doi: 10.1038/nmeth.2377; pmid: 23435261
22. M. Jadot *et al*., Accounting for protein subcellular localization: A compartmental map of the rat liver proteome. *Mol. Cell. Proteomics* **16**, 194–212 (2017). pmid: 27923875
23. M. Uhlen *et al*., Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010). doi: 10.1038/nbt1210-1248; pmid: 21139605
24. M. Uhlén *et al*., Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). doi: 10.1126/science.1260419; pmid: 25613900
25. K. D. Pruitt, T. Tatusova, G. R. Brown, D. R. Maglott, NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012). doi: 10.1093/nar/gkr1079; pmid: 22121212
26. A. Yates *et al*., Ensembl 2016. *Nucleic Acids Res.* **44**, D710–D716 (2016). doi: 10.1093/nar/gkv1157; pmid: 26687719
27. H. Kawaji *et al*., Update of the FANTOM web resource: From mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res.* **39**, D856–D860 (2011). pmid: 21075797
28. K. G. Ardlie *et al*., The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015). doi: 10.1126/science.1262110; pmid: 25954001
29. P. Gaudet *et al*., The neXtProt knowledgebase on human proteins: Current status. *Nucleic Acids Res.* **43**, D764–D770 (2015). doi: 10.1093/nar/gku1178; pmid: 25593349
30. M. Beck *et al*., The quantitative proteome of a human cell line. *Mol. Syst. Biol.* **7**, 549–549 (2011). doi: 10.1038/msb.2011.82; pmid: 22068332
31. T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann, Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **11**, 014050 (2012). doi: 10.1074/mcp.M111.014050; pmid: 22278370
32. P. Nilsson *et al*., Towards a human proteome atlas: High-throughput generation of mono-specific antibodies for tissue profiling. *Proteomics* **5**, 4327–4337 (2005). doi: 10.1002/pmic.200500072; pmid: 16237735
33. M. Peplow, Citizen science lures gamers into Sweden's Human Protein Atlas. *Nat. Biotechnol.* **34**, 452–453 (2016). doi: 10.1038/nbt0516-452c
34. J. Bordeaux *et al*., Antibody validation. *Biotechniques* **48**, 197–209 (2010). doi: 10.2144/000113382; pmid: 20359301
35. M. Baker, Reproducibility crisis: Blame it on the antibodies. *Nature* **521**, 274–276 (2015). doi: 10.1038/521274a; pmid: 25993940
36. M. Uhlen *et al*., A proposal for validation of antibodies. *Nat. Methods* **13**, 823–827 (2016). pmid: 27595404
37. C. Stadler *et al*., Systematic validation of antibody binding and protein subcellular localization using siRNA and confocal microscopy. *J. Proteomics* **75**, 2236–2251 (2012). doi: 10.1016/j.jprot.2012.01.030; pmid: 22361696
38. M. Skogs *et al*., Antibody validation in bioimaging applications based on endogenous expression of tagged proteins. *J. Proteome Res.* **16**, 147–155 (2017). pmid: 27723985
39. G. Covini *et al*., Cytoplasmic rods and rings autoantibodies developed during pegylated interferon and ribavirin therapy in patients with chronic hepatitis C. *Antivir. Ther.* **17**, 805–811 (2012). doi: 10.3851/IMP1993; pmid: 22293655
40. B. Vroling *et al*., NucleaRDB: Information system for nuclear receptors. *Nucleic Acids Res.* **40**, D377–D380 (2012). doi: 10.1093/nar/gkq1009; pmid: 22064856
41. S. A. Ochsner, C. M. Watkins, B. S. LaGrone, D. L. Steffen, N. J. McKenna, Research resource: Tissue-specific transcriptomics and cistromics of nuclear receptor signaling: a web research resource. *Mol. Endocrinol.* **24**, 2065–2069 (2010). doi: 10.1210/me.2010-0216; pmid: 20685849
42. A. Ruepp *et al*., CORUM: The comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res.* **38**, D497–D501 (2010). doi: 10.1093/nar/gkp914; pmid: 19884131
43. L. Gatto, L. M. Breckels, S. Wieczorek, T. Burger, K. S. Lilley, Mass-spectrometry-based spatial proteomics data analysis using pRoloc and pRolocdata. *Bioinformatics* **30**, 1322–1324 (2014). doi: 10.1093/bioinformatics/btu013; pmid: 24413670
44. L. M. Breckels *et al*., Learning from heterogeneous data sources: An application in spatial proteomics. *PLOS Comput. Biol.* **12**, e1004920 (2016). doi: 10.1371/journal.pcbi.1004920; pmid: 27175778

45. A. Fabregat *et al*., The Reactome pathway knowledgebase. *Nucleic Acids Res.* **44**, D481–D487 (2016). doi: 10.1093/nar/gkv1351; pmid: 26656494
46. J. Menche *et al*., Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015). doi: 10.1126/science.1257601; pmid: 25700523
47. A. Sakaue-Sawano *et al*., Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008). doi: 10.1016/j.cell.2007.12.033; pmid: 18267078
48. Gene Ontology Consortium, Gene Ontology Consortium: Going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015). doi: 10.1093/nar/gku1179; pmid: 25428369
49. L. C. Crosswell, J. M. Thornton, ELIXIR: A distributed infrastructure for European biological data. *Trends Biotechnol.* **30**, 241–242 (2012). doi: 10.1016/j.tibtech.2012.02.002; pmid: 22417641
50. C. E. Chapple *et al*., Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.* **6**, 7412 (2015). doi: 10.1038/ncomms8412; pmid: 26054620
51. K.-W. Min, S.-H. Lee, S. J. Baek, Moonlighting proteins in cancer. *Cancer Lett.* **370**, 108–116 (2016). doi: 10.1016/j.canlet.2015.09.022; pmid: 26499805
52. M. Lindskog, J. Rockberg, M. Uhlén, F. Sterky, Selection of protein epitopes for antibody production. *Biotechniques* **38**, 723–727 (2005). doi: 10.2144/05385ST02; pmid: 15945371
53. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011). doi: 10.1038/nmeth.1701; pmid: 21959131
54. L. Käll, A. Krogh, E. L. L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004). doi: 10.1016/j.jmb.2004.03.016; pmid: 15111065
55. H. Viklund, A. Bernsel, M. Skwark, A. Elofsson, SPOCTOPUS: A combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**, 2928–2929 (2008). doi: 10.1093/bioinformatics/btn550; pmid: 18945683
56. L. Fagerberg, K. Jonasson, G. von Heijne, M. Uhlén, L. Berglund, Prediction of the human membrane proteome. *Proteomics* **10**, 1141–1149 (2010). doi: 10.1002/pmic.200900258; pmid: 20175080
57. F. Danielsson *et al*., RNA deep sequencing as a tool for selection of cell lines for systematic subcellular localization of all human proteins. *J. Proteome Res.* **12**, 299–307 (2013). doi: 10.1021/pr3009308; pmid: 23227862
58. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016). doi: 10.1038/nbt.3519; pmid: 27043002
59. M. Krzywinski *et al*., Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009). doi: 10.1101/gr.092759.109; pmid: 19541911
60. D. Binns *et al*., QuickGO: A web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009). doi: 10.1093/bioinformatics/btp536; pmid: 19744993
61. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2008). doi: 10.1038/nprot.2008.211; pmid: 19131956
62. P. Shannon *et al*., Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003). doi: 10.1101/gr.1239303; pmid: 14597658

ACKNOWLEDGMENTS

We acknowledge the staff of the Human Protein Atlas program and Science for Life Laboratory for valuable contributions. We also acknowledge support from these facilities of the Science for Life Laboratory: the National Genomics Infrastructure and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure, the National Laboratories for Chemical Biology at Karolinska Institutet for access to imaging infrastructure, and the Advanced Light Microscopy facility for superresolution microscopy. We acknowledge H. Masai (Tokyo Metropolitan Institute of Medical Science) for providing the stable U-2 OS FUCCI cell line; A. Hyman (Max Planck Institute) for HeLa cells expressing GFP-tagged proteins; A. Falk (Karolinska Institute) for the AF22 cells; Evercyte for the ASC TERT1, HUVEC TERT2, HBF TERT88, HTCEpi, LHCN-M2, and RPTEC TERT cells; Horizon Discoveries for the Hap1 CRISPR knockout cells; and R. Pepperkok and B. Neumann for prepared solid-phase siRNA plates. We acknowledge CCP Games and MMOS Srl for the creation of Project Discovery and the players of EVE Online for classification of protein patterns in images through Project Discovery. Funding was provided by the Knut and Alice Wallenberg Foundation and the Erling Persson Foundation to M.U. and the Science for Life Laboratory to E.L. C.M.M. and L.M.B. were funded by Wellcome Trust grant 108441/Z/15/Z. A.G. was funded through the Alexander S. Onassis Public Benefit Foundation, the Foundation for Education and European Culture, and the Embiricos Trust Scholarship of Jesus College Cambridge. M.U., F.P., P.N., and S.H. are cofounders and M.U. and F.P. are on the board of Atlas Antibodies. All antibodies are available from various commercial providers, including Atlas Antibodies. The data are available in the supplementary materials. Images and Cell Atlas transcriptome and proteome data are available in the Human Protein Atlas (www.proteinatlas.org/humancell).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/356/6340/eaal3321/suppl/DC1
Additional Materials and Methods
Figs. S1 to S13
Tables S1 to S19
References (*63*–*66*)

2 November 2016; accepted 31 March 2017
Published online 11 May 2017
10.1126/science.aal3321

RESEARCH ARTICLE SUMMARY

CANCER

A pathology atlas of the human cancer transcriptome

Mathias Uhlen,* Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu,† Fredrik Ponten†

INTRODUCTION: Cancer is a leading cause of death worldwide, and there is great need to define the molecular mechanisms driving the development and progression of individual tumors. The Hallmarks of Cancer has provided a framework for a deeper molecular understanding of cancer, and the focus so far has been on the genetic alterations in individual cancers, including genome rearrangements, gene amplifications, and specific cancer-driving mutations. Using systems-level approaches, it is now also possible to define downstream effects of individual genetic alterations in a genome-wide manner.

RATIONALE: In our study, we used a systems-level approach to analyze the transcriptome of 17 major cancer types with respect to clinical outcome, based on a genome-wide transcriptomics analysis of ~8000 individual patients with clinical metadata. The study was made possible through the availability of large open-access knowledge-based efforts such as the Cancer Genome Atlas and the Human Protein Atlas. Here, we used the data to perform a systems-level analysis of 17 major human cancer types, describing both interindividual and intertumor variation patterns.

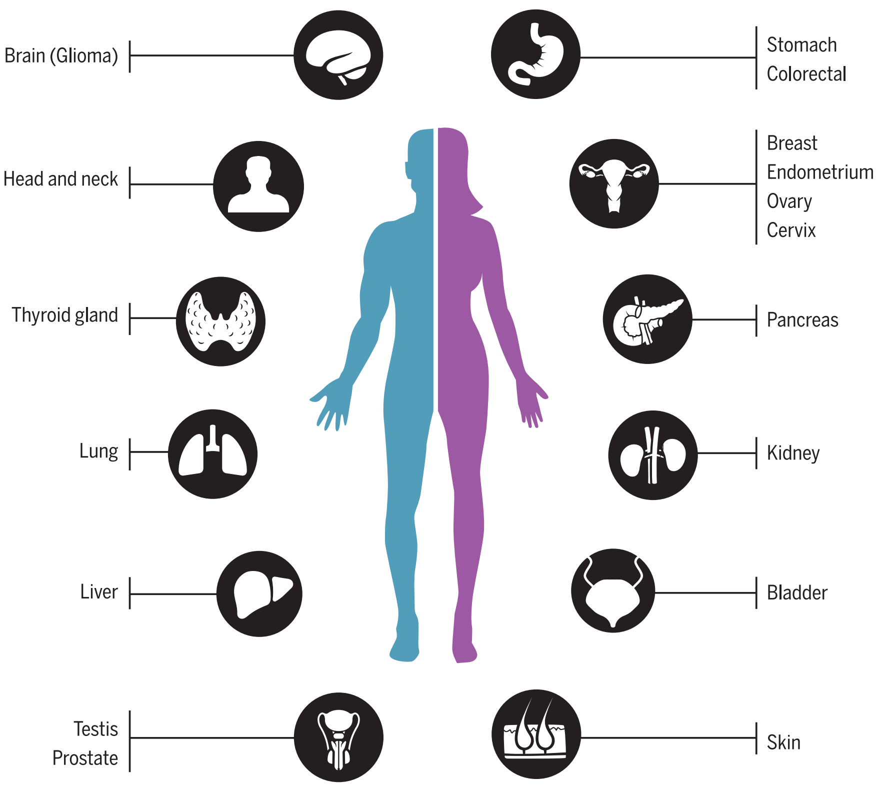
RESULTS: The analysis identified candidate prognostic genes associated with clinical outcome for each tumor type; the results show that a large fraction of cancer protein-coding genes are differentially expressed and, in many cases, have an impact on overall patient survival. Systems biology analyses revealed that gene expression of individual tumors within a particular cancer varied considerably and could exceed the variation observed between distinct cancer types. No general prognostic gene necessary for clinical outcome was applicable to all cancers. Shorter patient survival was generally associated with up-regulation of genes involved in mitosis and cell growth and down-regulation of genes involved in cellular differentiation. The data allowed us to generate personalized genome-scale metabolic models for cancer patients to identify key genes involved in tumor growth. In addition, we explored tissue-specific genes associated with the dedifferentiation of tumor cells and the role of specific cancer testis antigens on a genome-wide scale. For lung and colorectal cancer, a selection of prognostic genes identified by the systems biology effort were analyzed in independent, prospective cancer cohorts using immunohistochemistry to validate the gene expression patterns at the protein level.

CONCLUSION: A Human Pathology Atlas has been created as part of the Human Protein Atlas program to explore the prognostic role of each protein-coding gene in 17 different cancers. Our atlas uses transcriptomics and antibody-based profiling to provide a standalone resource for cancer precision medicine. The results demonstrate the power of large systems biology efforts that make use of publicly available resources. Using genome-scale metabolic models, cancer patients are shown to have widespread metabolic heterogeneity, highlighting the need for precise and personalized medicine for cancer treatment. With more than 900,000 Kaplan-Meier plots, this resource allows exploration of the specific genes influencing clinical outcome for major cancers, paving the way for further in-depth studies incorporating systems-level analyses of cancer. All data presented are available in an interactive open-access database (www.proteinatlas.org/pathology) to allow for genome-wide exploration of the impact of individual proteins on clinical outcome in major human cancers. ■

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aan2507

The Human Pathology Atlas



Schematic overview of the Human Pathology Atlas. A systems-level approach enables analysis of the protein-coding genes of 17 different cancer types from ~8000 patients. Results are available in an interactive open-access database.

The list of author affiliations is available in the full article online.

*Corresponding author. Email: mathias.uhlen@scilifelab.se

†These authors contributed equally to this work.

Cite this article as M. Uhlen *et al.*, *Science* **357**, eaan2507 (2017). DOI: 10.1126/science.aan2507

Originally published 18 August 2017

RESEARCH ARTICLE

CANCER

A pathology atlas of the human cancer transcriptome

Mathias Uhlen,[1,2,3]* Cheng Zhang,[1] Sunjae Lee,[1] Evelina Sjöstedt,[1,4] Linn Fagerberg,[1] Gholamreza Bidkhori,[1] Rui Benfeitas,[1] Muhammad Arif,[1] Zhengtao Liu,[1] Fredrik Edfors,[1] Kemal Sanli,[1] Kalle von Feilitzen,[1] Per Oksvold,[1] Emma Lundberg,[1] Sophia Hober,[3] Peter Nilsson,[1] Johanna Mattsson,[4] Jochen M. Schwenk,[1] Hans Brunnström,[5] Bengt Glimelius,[4] Tobias Sjöblom,[4] Per-Henrik Edqvist,[4] Dijana Djureinovic,[4] Patrick Micke,[4] Cecilia Lindskog,[4] Adil Mardinoglu,[1,3,6]† Fredrik Ponten[4]†

Cancer is one of the leading causes of death, and there is great interest in understanding the underlying molecular mechanisms involved in the pathogenesis and progression of individual tumors. We used systems-level approaches to analyze the genome-wide transcriptome of the protein-coding genes of 17 major cancer types with respect to clinical outcome. A general pattern emerged: Shorter patient survival was associated with up-regulation of genes involved in cell growth and with down-regulation of genes involved in cellular differentiation. Using genome-scale metabolic models, we show that cancer patients have widespread metabolic heterogeneity, highlighting the need for precise and personalized medicine for cancer treatment. All data are presented in an interactive open-access database (www.proteinatlas.org/pathology) to allow genome-wide exploration of the impact of individual proteins on clinical outcomes.

Cancer is one of the leading causes of death worldwide, and both the incidence and prevalence of cancer continue to increase. Most current cancer drugs are effective only in a subgroup of patients owing to interindividual tumor heterogeneity, and large gaps remain in our current understanding of the best treatment approaches and the underlying molecular mechanisms driving cancer pathogenesis (*1*). There is therefore an urgent need for the development of personalized diagnostic and therapeutic strategies using methods such as systems-level analysis (*2*–*4*). Such approaches can be used to study the genome-wide effect of gene rearrangements, amplifications, and specific cancer-driving mutations on protein-coding regions.

Thanks to large open-access knowledge-based efforts, such as The Cancer Genome Atlas (TCGA) (*5*), the Human Protein Atlas (HPA) (*6*), the GTEx consortium (*7*), and recount2 (*8*), it is now possible to explore the genome-wide expression of individual genes in different tissues and cancers (*9*). The database resource from TCGA represents a comprehensive and coordinated effort to accelerate our understanding of cancer (*5*), and the HPA and GTEx represent international efforts to map the expression of protein-coding genes in normal human tissues. Many of the patients included in the TCGA database are also accompanied by clinical survival metadata, allowing clinical outcomes to be associated with genome-wide expression patterns of protein-coding genes and metabolic modeling of individual cancer patients. Such analysis is facilitated by the recent suggestion that there is a gene-specific correlation between RNA and protein levels in human tissues and cells, allowing quantitative analyses of mRNA levels to be used as proxies for the corresponding protein levels (*10*).

Here, we used data from TCGA and the HPA efforts to perform a systems-level analysis of 17 major human cancer types corresponding to 7932 tumor samples, and describe both interindividual and intertumor variation patterns. The analysis identified candidate prognostic genes associated with clinical outcome for each tumor type and generated metabolic models for individual patients. A Human Pathology Atlas has been created as part of the Human Protein Atlas program to explore the prognostic role of each protein-coding gene in each cancer type by means of transcriptomics and antibody-based profiling (Fig. 1A). More than 100 million Kaplan-Meier survival plots were generated as part of the genome-wide analysis of potential prognostic genes in these cancers. More than 900,000 survival plots—each accompanied with statistical significance—can be visualized at the new pathology resource.

To investigate the key prognostic genes affecting patient survival, we generated cancer-specific coexpression networks for each of the studied cancer types and examined the functional relationship between the prognostic genes and the genes associated with Hallmarks of Cancer (*11*). Personalized genome-scale metabolic models (GSMMs) for the tumors in each cancer patient were generated to study the individual metabolic differences among tumors. This analysis also allowed us to study the role of tissue-specific genes in the "dedifferentiation" of cancer and the role of specific cancer testis antigens (CTAs) on a genome-wide scale. For two of the cancer types, lung and colorectal cancer, a selection of prognostic genes identified by the systems biology effort were analyzed in independent prospective cancer cohorts, using immunohistochemistry (IHC) to validate the gene expression patterns at the protein level.

All primary Human Pathology Atlas data are freely available without restrictions in the public open access database (www.proteinatlas.org/pathology) that is part of the Human Protein Atlas program. Significant prognostic genes in each cancer type are highlighted together with Kaplan-Meier plots based on overall survival and accompanied with data for individual gene expression heterogeneity of prognostic genes at the time of diagnosis.

Transcriptome analysis of human cancers

We retrieved RNA sequencing (RNA-seq) data together with clinical metadata corresponding to the 33 different human cancers that are available in TCGA (table S1). As a result, data were collected from 9666 individuals out of the 11,000 cancer patients included in the TCGA project from the Genomic Data Commons (GDC) Data Portal (https://gdc-portal.nci.nih.gov/). First, using hierarchical clustering, we investigated the relationship between the global gene expression patterns of all protein-coding genes in the 33 cancer types (n = 19,571) and the gene expression patterns in 37 normal human tissues obtained from 162 healthy subjects in the HPA project (*6*) (fig. S1). RNA-seq data from all cancer tissues and all normal tissues were processed in the same bioinformatics pipeline and normalized as fragments per kilobase of exon per million fragments mapped (FPKM). We found that a majority of all cancers (26 of 33) clustered in the same group, while the majority of the normal tissues (33 of 37) clustered in a different group, indicating that most cancer types share expression features that render them significantly different from normal tissues. Notably, we found that liver tissue and the primary form of liver cancer, hepatocellular carcinoma, as well as bone marrow and acute myeloid leukemia clustered together, suggesting that these phenotypes are more closely related independent of a benign or malignant status.

We previously classified all protein-coding genes into six different categories according to their expression across normal tissues and organs (*6*). The classification, based on a FPKM cut-off >1, ranged from genes expressed in all tissues to those

[1]Science for Life Laboratory, KTH–Royal Institute of Technology, Stockholm, Sweden. [2]Center for Biosustainability, Danish Technical University, Copenhagen, Denmark. [3]School of Biotechnology, AlbaNova University Center, KTH–Royal Institute of Technology, Stockholm, Sweden. [4]Department of Immunology Genetics and Pathology, Uppsala University, Uppsala, Sweden. [5]Division of Pathology, Lund University, Skåne University Hospital, Lund, Sweden. [6]Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.
***Corresponding author. Email: mathias.uhlen@scilifelab.se**
†These authors contributed equally to this work.

Originally published 18 August 2017

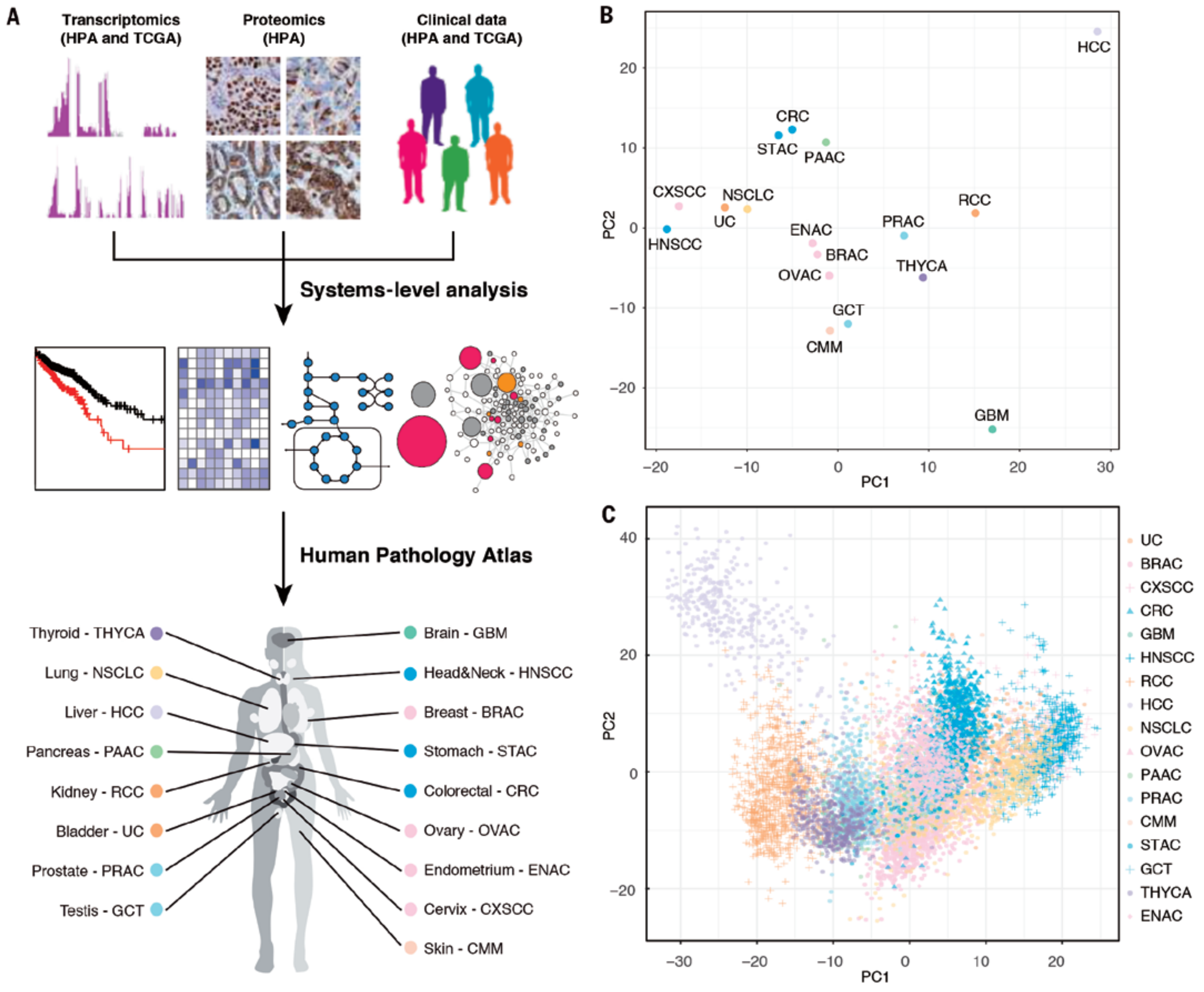


Fig. 1. Analysis of the global expression patterns of protein-coding genes in human cancers. (**A**) Schematic drawing of the Human Pathology Atlas effort described herein. (**B**) Principal components analysis (PCA) showing the similarities in expression of 19,571 protein-coding genes among 17 cancer types. See fig. S4 for additional PCA analysis with more stratified patient cohorts. (**C**) PCA plot showing the individual differences in the genome-wide global expression profiles among the 17 cancer types in 9666 individual patients.

with tissue-restricted expression and those not detected in any of the analyzed tissues. The transcriptomics data for the 33 different cancers allowed us to classify the protein-coding genes into six different categories based on the expression level. Our analysis revealed that a large fraction (41%) of the protein-coding genes were expressed in all analyzed cancers, while approximately 46% (*n* = 9057) displayed more tumor type-restricted expression. Among the protein-coding genes, 13% were not detected in any tumor types investigated (fig. S2 and table S2). The majority of the genes (*n* = 5772) detected in all samples were shared between cancers and normal tissues, whereas 2401 additional genes were expressed in all cancers analyzed, but with more restricted expression in the normal tissues. These "housekeeping" genes in tumors are enriched in biological functions related to DNA replication and the regulation of apoptosis and mitosis (table S3 and fig. S3).

Subsequently, we focused our analysis on 17 tumor types with large numbers of patients available in the TCGA data set accompanied by clinical metadata (Fig. 1A and table S4). The connectivity among these 17 cancers was determined using principal components analysis (PCA) based on the expression pattern of all protein-coding genes (Fig. 1B and fig. S4). We observed a relationship among cancer types that shared a similar tissue type of origin or similar morphological features and phenotypic expression patterns. For example, cancers with a dominating squamous cell carcinoma phenotype, such as cervical or head and neck cancer, clustered together close to the related urothelial cell carcinoma and non–small cell lung cancer (NSCLC), which also contains a large fraction of squamous cell carcinoma. Adenocarcinomas that originate from the gastrointestinal tract, including pancreatic cancer, also clustered separately from the cluster containing the three adenocarcinomas representing female cancer (i.e., breast, endometrial, and ovarian cancer). Interestingly, testicular germ cell tumors were located close to melanoma and were well separated from the more classical epithelial tumor types, whereas glioma (brain) and hepatocellular (liver) carcinoma clearly represented the most divergent tumor types in this global expression analysis.

Individual variation among cancers

To determine the individual gene expression patterns within and among certain cancer types, we used PCA to visualize the global expression patterns for all 9666 individual tumors that were included in the patient cohorts, representing the 17 major cancer types (Fig. 1C). The results showed that the interindividual variation within each type of cancer was considerable, and that there was a large overlap in expression among individuals with different cancer types. One exception was liver cancer (Fig. 1C, upper left), in which the individual tumors showed relatively unique global expression patterns with little overlap with the other cancer types. Thus, gene expression varies considerably in individual tumors within a particular cancer subtype. For some patient tumors, the global expression pattern resembles other cancer types more than it does the given type of diagnosed cancer, which reinforces previous discoveries (12).

Clinical outcome based on gene expression analysis

First, we analyzed the survival data from the TCGA metadata (fig. S5 and table S4). Prostate cancer and testis cancer (germ cell tumors) have the most favorable 3-year survival rates (98% and 97%, respectively), while high-grade glioma and pancreatic cancer have the lowest 3-year survival rates (8% and 35%, respectively). The patient survival data and matched transcriptomic data enabled us to perform gene-centric and genome-wide survival analyses to identify prognostic genes across the 17 cancer types. For each cancer, all patients with survival data were included in the Kaplan-Meier survival analysis spanning 10 years as extracted from the metadata. The RNA levels at the time of diagnosis were plotted against the survival data as extracted from the follow-up clinical data (see examples in Fig. 2A). For each gene and cancer type, the patient cohort was stratified into two groups with the highest and lowest expression (FPKM) based on individual expression levels. To choose the best FPKM cutoffs for grouping the patients most significantly, we used all FPKM values from the 20th to 80th percentiles to group the patients, examined significant differences in the survival outcomes of the groups, and selected the value yielding the lowest log-rank P value. In total, more than 100 million Kaplan-Meier plots were generated that corresponded to all 19,571 protein-coding genes across the 17 cancer types. As a comparison, we also tested the method described by Hothorn and Lausen (13) and the results were highly similar (fig. S6). Two examples of genes in the liver cancer cohort are shown in Fig. 2B, including the survival data for the individual patients in the liver cancer cohort.

We identified two types of prognostic marker genes in terms of clinical outcome: (i) unfavorable prognostic genes, for which higher expression of a given gene was correlated with a poor patient survival outcome, and (ii) favorable prognostic genes, for which higher expression of a given gene was correlated with a longer patient survival outcome. A prognostic gene for a given cancer was defined as a gene for which the expression level above or below the experimentally determined cutoff in an individual patient yields a significant ($P < 0.001$) difference in overall survival. The ratios of favorable and unfavorable prognostic genes varied among the different types of cancer. In Fig. 2C, the numbers of prognostic genes for each of the 17 cancer types are shown, with more detailed information provided in table S5. It is noteworthy that 2375 genes showed opposite effects on prognosis depending upon cancer type and location, highlighting the need to perform functional studies of prognostic genes. See table S6 for a complete list of the prognostic association of all genes in all cancers.

In Fig. 2A, examples of favorable and unfavorable prognostic genes are shown for five of the cancer types, based on the optimal stratification P value calculated for each gene and cancer. In each case, a significant separation ($P < 0.001$) of the survival rate could be observed on the basis of differences in the expression levels of the respective gene. For some genes, the prognostic value has previously been reported in the literature; one example is RBM3 (RNA binding motif protein 3) (Fig. 2A), which has been implicated in survival of colorectal cancer (14). However, most of the identified prognostic genes lacked prior reports of a survival link to a given cancer, making them potential candidates for follow-up studies.

We extended the survival analysis by constructing panels of the five most significant favorable and unfavorable prognostic genes (table S7) for each tumor type and used them to predict the clinical outcome (Fig. 2A). Each of the five panels generated a prognostic panel of high significance ($P < 10^{-5}$). Similarly, all of the other 12 cancer types yielded prognostic panels in the same manner with very high significance (table S7). It is noteworthy that for cancers with more favorable survival rates (e.g., testicular or prostate cancer), a limited number of prognostic genes have been identified, perhaps because the 3-year survival probability for these cancers exceeds 95% and thus larger patient cohorts are needed to obtain prognostic genes with high significance. For two of the tumors (i.e., renal and liver cancer), the numbers of prognostic genes were much larger than for the other cancers (6070 and 2892, respectively) (Fig. 2C). This observation is interesting because both are cancers with distinct features and morphology, and liver cancer especially appears to be distantly related to other cancer types (Fig. 1B). For renal cancer, the number of favorable (n = 2782) and unfavorable genes (n = 3288) was balanced, whereas there were a large number of unfavorable prognostic genes (n = 2629) for liver cancer. An earlier study of renal cancer based on TCGA data showed distinctly different groups of patients that are not reflected by morphological subtypes (e.g., clear cell, papillary, and chromophobe phenotypes) (15). Thus, the large number of prognostic genes may simply reflect large global expression differences between these two subtypes, resulting in a large number of "passenger" genes and a much smaller set of driver genes affecting the clinical course of the patient.

Overlap of prognostic genes across cancer types

We examined the extent of overlap of prognostic genes among different cancer types. The correlation among the 17 cancer types for favorable and unfavorable prognostic genes was investigated in a pairwise manner (Fig. 3A). For most cancers, little correlation was observed, suggesting a relatively limited number of common prognostic genes. In contrast, a significant overlap of favorable prognostic genes was observed for other cancers (e.g., renal, liver, lung, and pancreatic cancers). Similarly, unfavorable prognostic genes for some cancers, including renal, breast, lung, and pancreatic cancer, clustered together. However, a detailed analysis revealed that no prognostic genes were shared among more than 7 of the cancer types (table S8).

Functional analysis of prognostic genes

A functional gene ontology (GO) analysis was performed for the most significant prognostic genes shared among the 17 major cancers, including both favorable and unfavorable genes (table S9). The results (Fig. 3B) suggest that many of the common unfavorable genes are related to cell proliferation, including mitosis, cell cycle regulation, and nucleic acid metabolism. In contrast, few GO functions were significantly overrepresented by the common favorable genes; the most enriched GO functions were positive regulation of cell activation, regulation of immune cell activation, and cell-cell adhesion.

Because genes associated with proliferation were identified by the functional analysis, we investigated the prognostic effect of all 314 cell cycle genes defined by the Molecular Signature database (16) in various cancer types. Interestingly, more than 60% (n = 194) of these genes were associated with an unfavorable clinical outcome, with increased expression in at least one of the analyzed cancer types (table S10). However, these prognostic cell cycle genes were generally only shared among a few cancers (Fig. 3C), which suggests that although cell cycle genes are commonly unfavorable genes, the use of a particular set of cell cycle genes and their effect on clinical outcome may differ among individual cancer types.

Tissue-enriched genes and dedifferentiation in cancer

We further analyzed genes with high relative expression that correlated with prolonged overall survival, for which a high expression level of a particular gene was associated with a good clinical outcome. Many of these favorable genes have previously (6) been classified as elevated in certain normal tissues (table S11), as exemplified in liver cancer (Fig. 3D), for which more than half (n = 150) of the 263 favorable prognostic genes were defined as tissue-elevated. To further investigate the molecular signatures related to differentiation, we analyzed alterations in liver-enriched genes (n = 154) defined by tissue-wide expression studies of normal hepatocytes. Samples from normal liver tissue were analyzed and compared with the transcriptomics patterns of the primary liver cancer biopsies and the liver cancer–derived HepG2 cell line. To further compare the expression levels of the tissue-enriched proteins, we plotted the genome-wide transcriptomics data using the relative changes between cancer/normal tissue and cell line/normal tissue, respectively, for all genes expressed in the normal liver. The liver-enriched genes (red), liver group–enriched genes (orange), and all other expressed genes (black) are summarized in Fig. 4A. The global analysis demonstrates a down-regulation in both the liver cancer and the cancer cell line as compared with the expression levels in normal liver tissue (lower left quadrant). This quadrant contains 102 of the 154 liver-enriched genes (66%), which suggests that liver-enriched genes are down-regulated as a sign of dedifferentiation in both liver cancer and liver cancer cell lines.

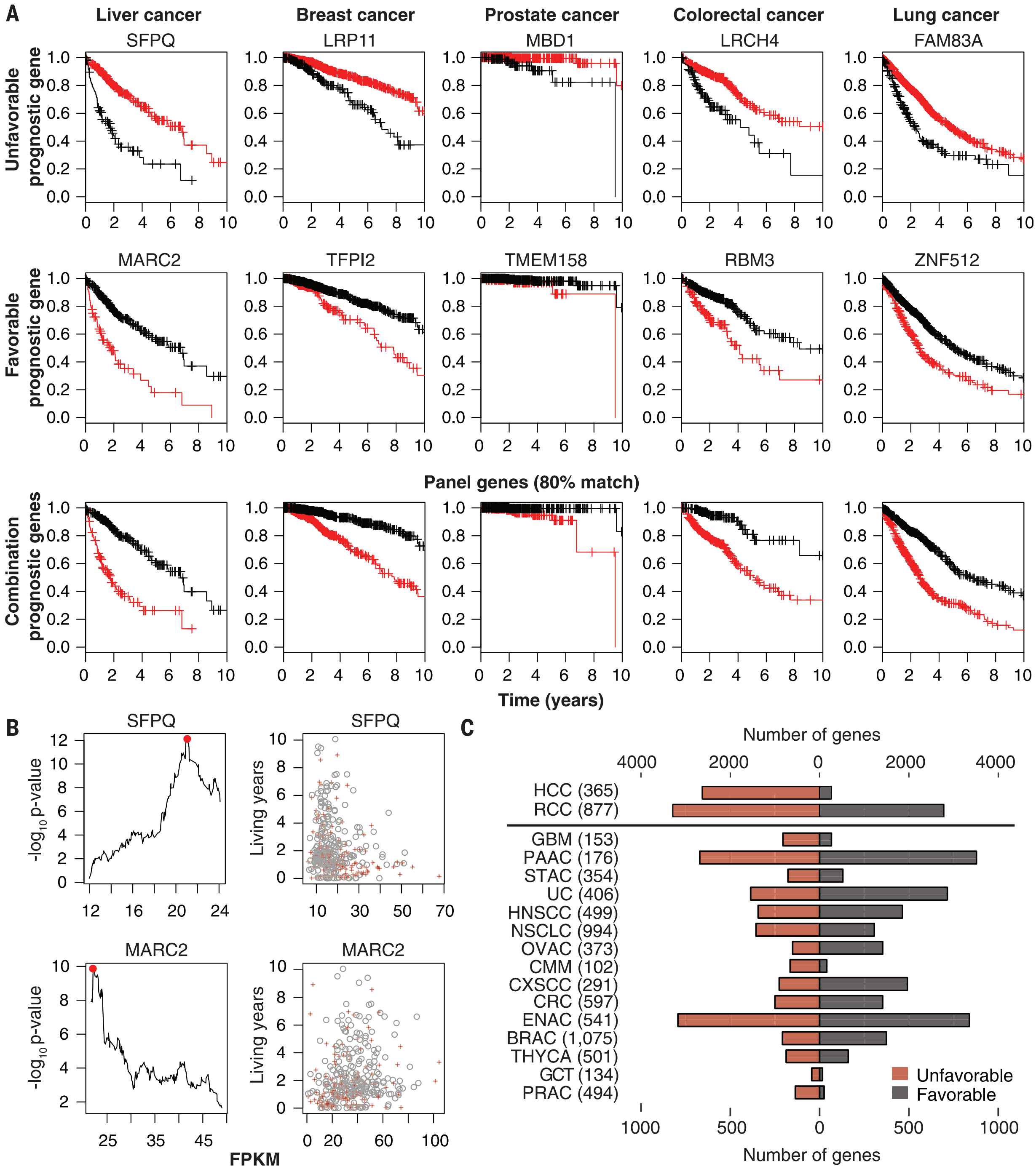


Fig. 2. Identification of prognostic genes based on expression coupled with clinical survival for 17 different cancer types. (**A**) Examples of Kaplan-Meier plots for five major cancer patients stratified by the expression of an unfavorable prognostic gene (first row), a favorable prognostic gene (second row), and a combination of 10 prognostic genes (third row). The selected unfavorable and favorable genes had the best log-rank *P* value based on the Kaplan-Meier analysis, with average RNA expression levels more than the median average expression of all protein-coding genes; the 10 marker genes were a combination of the top five favorable and unfavorable genes with expression higher than the median average expression. Black and red lines show high and low (or, in the third row, favorable and unfavorable) expression, respectively. (**B**) Examples of two prognostic genes in liver cancer. Left: Distribution of log-rank *P* values against the RNA expression with different RNA-level (FPKM) cutoffs. Right: Patient-centric scatterplot showing the relationships between living years and RNA expression of the prognostic genes. (**C**) Numbers of genes showing favorable and unfavorable prognostic effects in the 17 Human Pathology Atlas cancer types. Patient numbers for each cancer are shown in parentheses.

Metadata for the grade of malignancy (i.e., the degree of differentiation) are available in the TCGA database, and this allowed us to analyze the relative expression level of liver-enriched genes in liver cancer and to compare different grades of malignancy. The tumor grade was scored using the modified nuclear grading scheme outlined by Edmondson and Steiner (17), with the tumor grade categorized as low, intermediate, or high. The malignancy grade (G1 to G3) (18) was available for 341 cases. The analysis revealed a significant correlation between the malignancy grade and the expression pattern of liver-enriched genes that were significantly down-regulated in liver cancer. In Fig. 4B, examples of IHC-based protein expression levels of a liver-enriched gene (CYP2C9) are displayed for normal liver versus liver cancer with differing tumor grade. The gene expression levels of CYP2C9 across all patients are also shown as box plots for different tumor grades (Fig. 4C). In addition, we analyzed the distribution of correlation coefficients for all analyzed liver-enriched genes compared with that of a randomly selected set of genes (Fig. 4D). Randomly selected genes showed no correlation (median rho = 0.07), whereas the tissue-enriched genes showed a negative correlation, with reduced

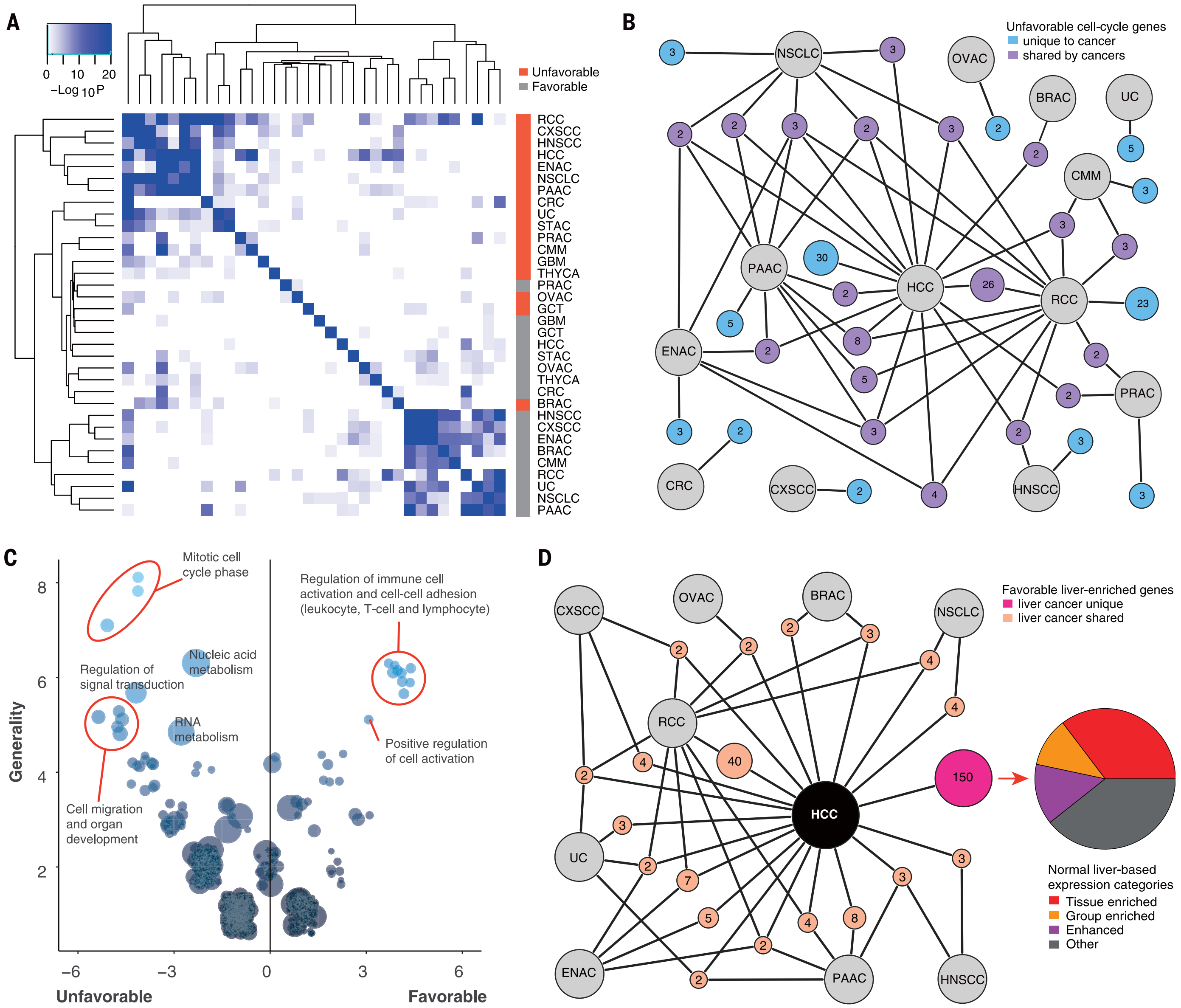


Fig. 3. Network analysis of prognostic genes. (**A**) Heat map showing the hypergeometric *P* value for the pairwise overlap of prognostic genes between the cancer types. (**B**) Bubble plot showing the common enriched Gene Ontology (GO) functions among the 17 Human Pathology Atlas cancer types. Bubble sizes represent numbers of genes in GO function; the *x* and *y* axes indicate the directionalities and generalities of the GO terms. Generality is defined by the number of cancers with their prognostic genes overrepresenting the GO function; directionality is defined by the number of cancers with their favorable genes overrepresenting the GO function minus the number of cancers with unfavorable genes overrepresenting the GO function. Note that only functions with more than five generalities are labeled. All GO terms for each cancer are provided in table S9. Results based on optional *P* value or hazard ratio cutoff–defined prognostic genes are provided in fig. S7 and table S9. (**C**) Network plot showing the number of cancer-specific and shared unfavorable cell cycle genes in all cancer types. Note that all groups with only one gene were removed from the plot. (**D**) Network plot showing the number of liver cancer–specific favorable genes and the favorable genes shared among liver and other cancers in the Human Pathology Atlas. Inset: Pie chart showing the fraction of elevated normal liver genes among the liver cancer–specific favorable genes.

expression of tissue-enriched genes in high-grade tumors (grade G3). The results demonstrated a molecular correlation between the expression levels of tissue-enriched genes and tumor grade, supporting the concept that dedifferentiated cancers are associated with decreased patient survival.

Cancer testis antigens in liver cancer

Cancer testis antigens are expressed in a wide range of cancer types, whereas their expression in normal tissues is restricted to immune-privileged sites such as the testis and placenta. To explore this observation further, we investigated the differential expression patterns of testis-enriched genes in normal liver, primary liver biopsies, and a liver cancer–derived cell line (HepG2). A global analysis, shown in Fig. 4E (upper right quadrant), showed that many of the testis-enriched genes had higher expression in the patient biopsy and cell line than in normal liver tissue. The results support many previous studies (19) that testis-enriched genes have higher expression in cancer than in the corresponding normal tissues.

Coexpression networks of human cancers

The Hallmarks of Cancer (11) has laid an important foundation for understanding cancer pathogenesis, and from the corresponding cellular processes, 2172 genes have recently been defined as hallmark-related genes (16, 20). We thus decided to investigate their relationship with the prognostic genes reported here. Approximately two-thirds (65%) of the "hallmark genes" were predictive for clinical outcome in at least one of the cancers analyzed, but a network analysis revealed that none of the genes were shared among the majority of cancers, with most genes consequently affecting only a few of the cancer types (Fig. 5A and figs. S8 and S9). Subsequently, a cancer-specific coexpression network analysis for all 17 major cancers (table S12; available at http://inetmodels.com) was performed to identify genes that are expressed concurrently during tumorigenesis. Figure 5B shows a coexpression cluster in the lung cancer cohort, with enrichment for both prognostic and hallmark genes. Within this cluster, the hub genes (located in the center) are generally more prognostic than those with less coexpression. It is tempting to speculate

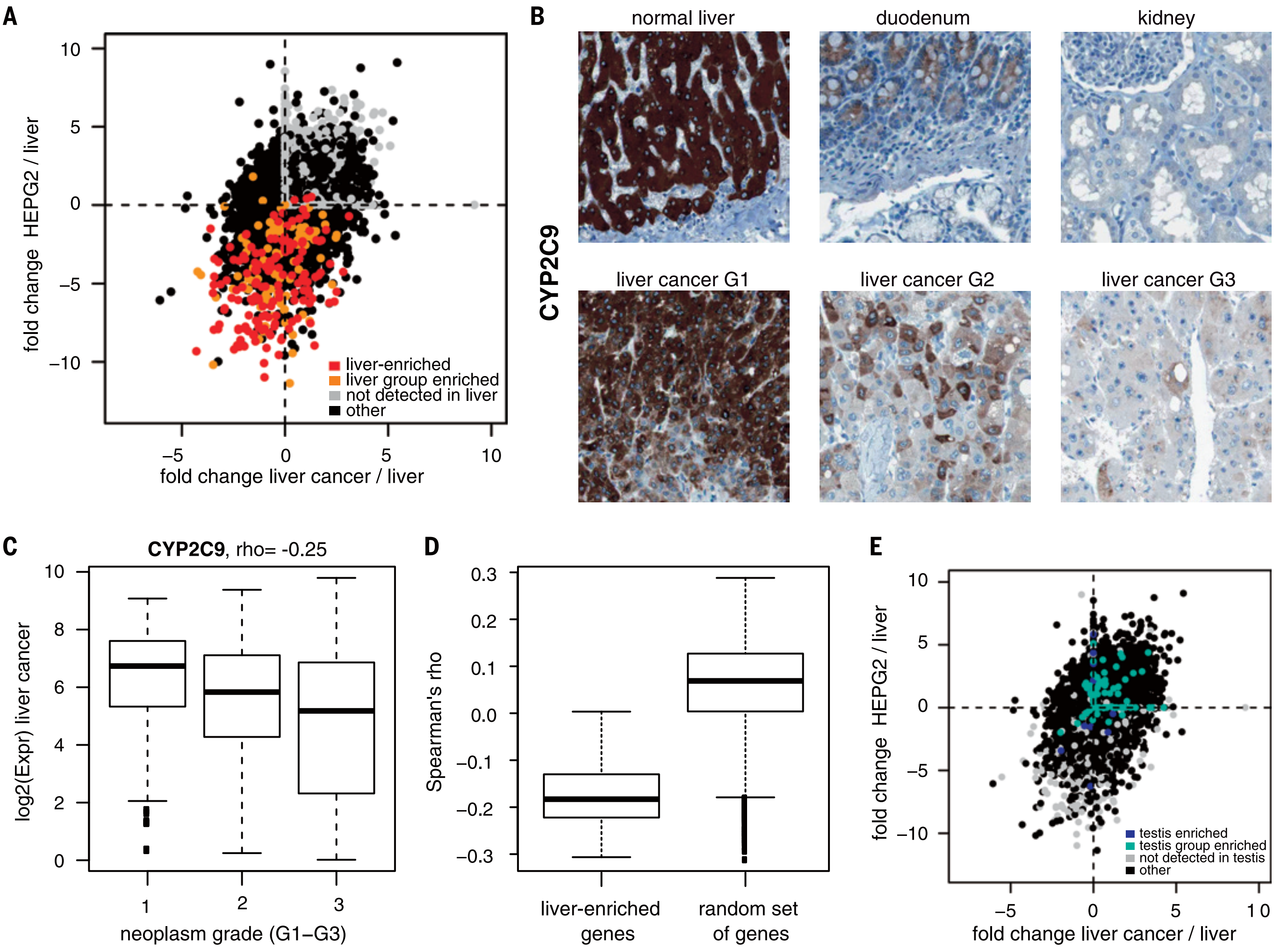


Fig. 4. Correlation between tumor differentiation and expression of liver-enriched genes. (**A**) Scatterplots showing the relative (fold) change between the transcript expression level in liver cancer and normal liver tissue (*x* axis) and the HepG2 cell line and normal tissue (*y* axis) for all protein-coding genes. Individual genes are colored according to their expression-based category in liver. All FPKM values less than 1 were set to 1 for the fold change calculation. (**B**) IHC staining of CYP2C9 proteins in four normal tissues and different hepatocellular carcinoma samples. For full IHC protein profiles, view the gene at www.proteinatlas.org/pathology. (**C**) Box plots showing the expression levels of liver tumor samples of different neoplasm grades for three representative liver-enriched genes for CYP2C9. (**D**) Box plot showing the distribution of correlation coefficients (Spearman's rho) between the neoplasm grade and expression for a random set of genes and all liver-enriched genes in liver tumors. (**E**) Scatterplots for all protein-coding genes showing the fold change in testis-specific antigen in liver cancer and normal liver tissue (*x* axis) and in the HepG2 cell line and normal liver tissue (*y* axis). Individual genes are colored according to their expression-based category in the testis.

that the hub genes in this cluster are lung cancer "drivers" and that the genes located around the outer boundary are lung cancer "passengers." Using somatic copy number alteration data in a TCGA pan-cancer analysis, we found that 36.4% of the genes in this cluster (table S13) were amplified or deleted in their chromosomal regions (21).

Among cancer-specific coexpression clusters, those that were significantly enriched with prognostic genes (hypergeometric test, $P \leq 0.05$) were named prognostic clusters, and an average of 13.9 clusters per cancer were enriched with prognostic genes (fig. S10 and table S14). A functional analysis, as exemplified by lung cancer (Fig. 5C and fig. S9), showed that many prognostic clusters were enriched with genes associated with the hallmark genes, such as those involved in DNA repair, cell proliferation, angiogenesis, and cell-cell signaling, implying that those processes or pathways may be associated with lung cancer progression. Across the 17 cancer types, the fractions of prognostic genes associated with the hallmark genes were determined (Fig. 5D and fig. S9); more than half (57% on average) of the prognostic genes were not identified as hallmark genes but were coexpressed with hallmark genes. It remains to be determined whether many of the prognostic genes identified herein have a passive or dominant role in the development of cancer.

Personalized metabolic networks for cancer patients

Tumors increase the nutrient import from the environment to fulfill biosynthetic demands

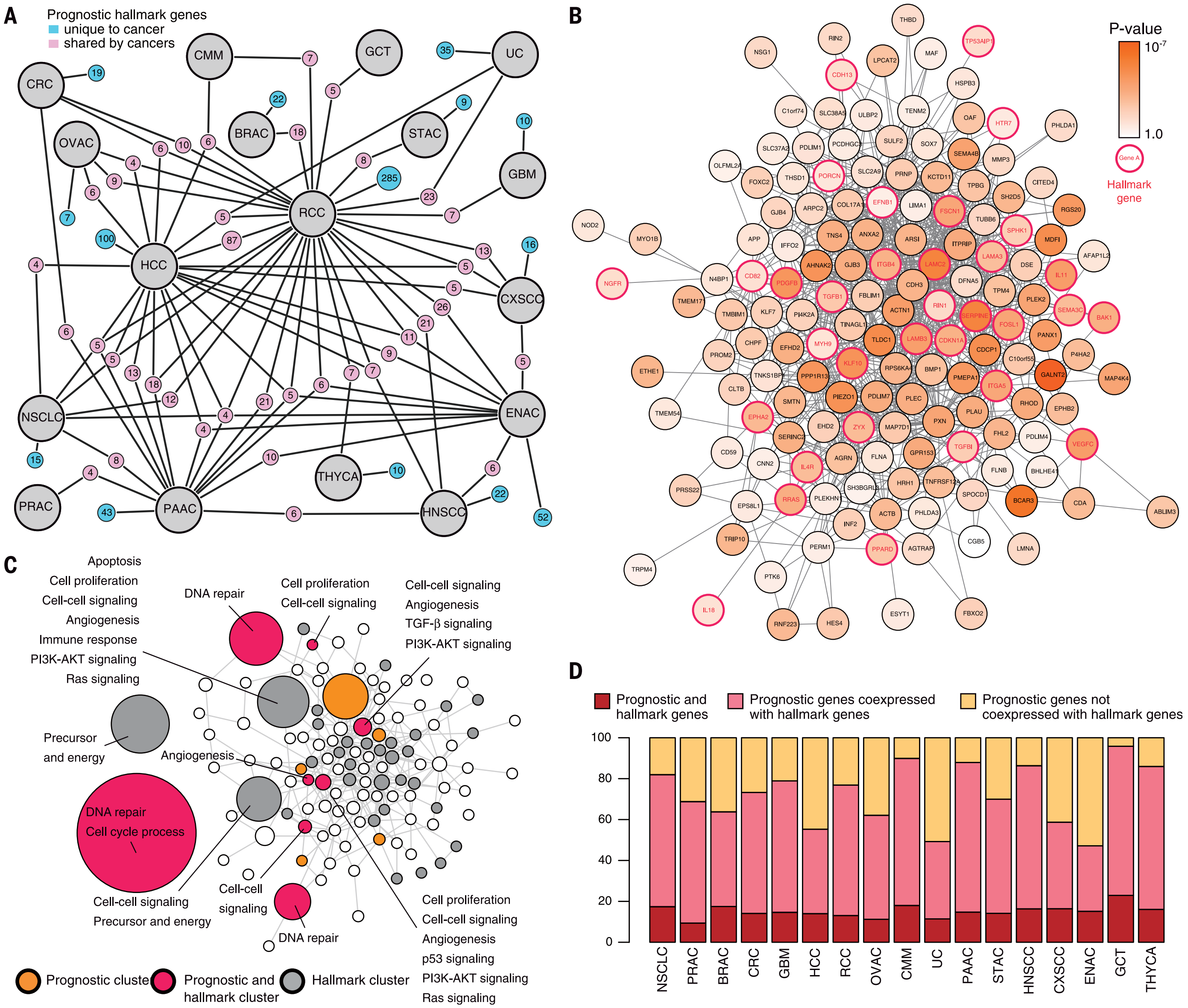


Fig. 5. Coexpression analysis reveals the relationship with the Hallmarks of Cancer and clues for drivers among prognostic genes. Gene coexpression of 17 cancers was investigated on the basis of established cancer coexpression networks. (**A**) Network plot showing the number of cancer-specific and shared prognostic cancer hallmark genes in all cancer types. Note that all groups with fewer than four genes were removed from the plot. (**B**) A gene coexpression cluster from the coexpression network of lung cancer enriched with both hallmark and prognostic genes. (**C**) Network plot showing coexpression clusters of lung cancer. All nodes indicate gene coexpression clusters; edges indicate significant coexpression links between clusters. The gray, yellow, and red color of the nodes indicates that the cluster was significantly enriched with hallmark genes, prognostic genes, and both cases, respectively. (**D**) Bar plot showing the fraction of prognostic genes that are mere hallmark genes (red), coexpressed in hallmark gene clusters (pink), or not coexpressed with hallmark genes (gold).

associated with proliferation, making use of these nutrients to both maintain viability and build new biomass (*22*–*24*). To investigate the metabolic reprogramming of each tumor, we generated personalized GSMMs for tumors from more than 7000 of the 17 major cancer patients based on transcriptomics data and generic human GSMM HMR2 (*25*) as previously described (*26*) (Fig. 6A). The resulting personalized GSMMs ranged in size from 2070 to 4058 metabolites, 2093 to 5261 reactions, and 978 to 2102 associated genes (fig. S11 and table S15). A total of 4889 metabolites, 6977 reactions, and 2760 genes were shared across the models; 1419 metabolites, 1020 of the reactions, and 334 of the genes were present in all personalized GSMMs. The personalized GSMMs are available in the BioModels Database (www.ebi.ac.uk/biomodels) with accession numbers MODEL1707110000 to MODEL1707116752.

Personalized GSMMs may allow for the investigation of common and unique biological functions for each patient (*27*). Using personalized GSMM and constraint-based modeling, we investigated heterogeneities of individual cancers by identifying genes within a tumor that are important for its growth (*3*). This method is suitable for studying cancer metabolism because it assumes an increase in tumor growth rate under optimal conditions and hence searches for metabolic flux distributions to produce essential biomass precursors at high rates (*2*, *28*, *29*). We found significant differences in the essential genes catalyzing tricarboxylic acid (TCA) cycle metabolism in liver cancer (Fig. 6B). As shown, the enzyme FH (fumarate hydratase) is identified as a conserved gene for tumor growth in all liver cancer patients analyzed, whereas SDHA (succinate dehydrogenase complex, subunit A) is important for tumor growth in ~60% of liver cancer patients, and ACLY (ATP citrate lyase) is key for tumor growth in fewer than 5% of liver cancer patients. In total, we identified 2553 essential genes that can inhibit or kill tumor growth in any of the analyzed samples and found that 55 (2%) of the key genes are common in all cancer patients analyzed, regardless of the cancer type (table S14). Notably, we found that only 10% to 25% of the essential genes were conserved in more than 80% of patients of each cancer type (Fig. 6C).

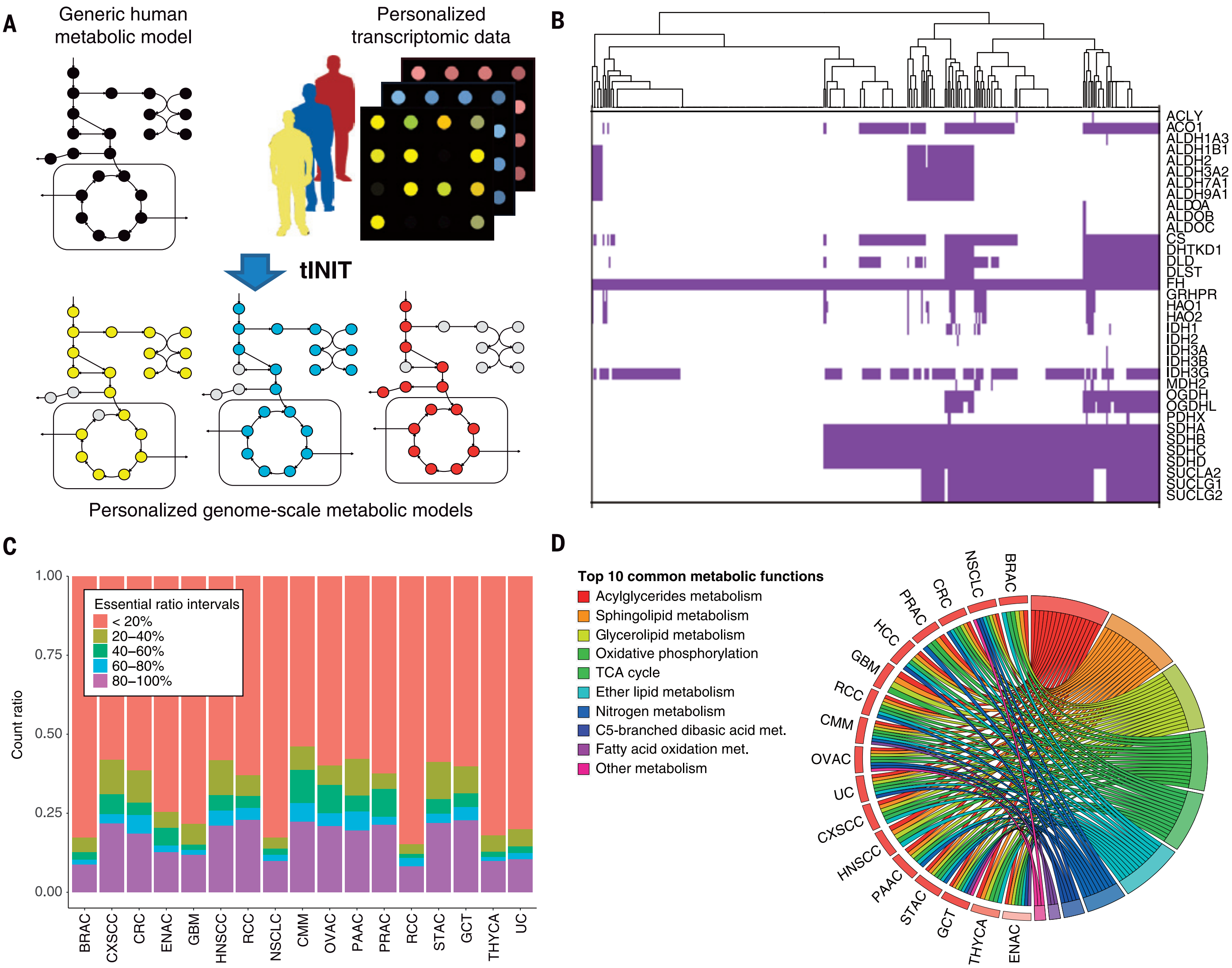


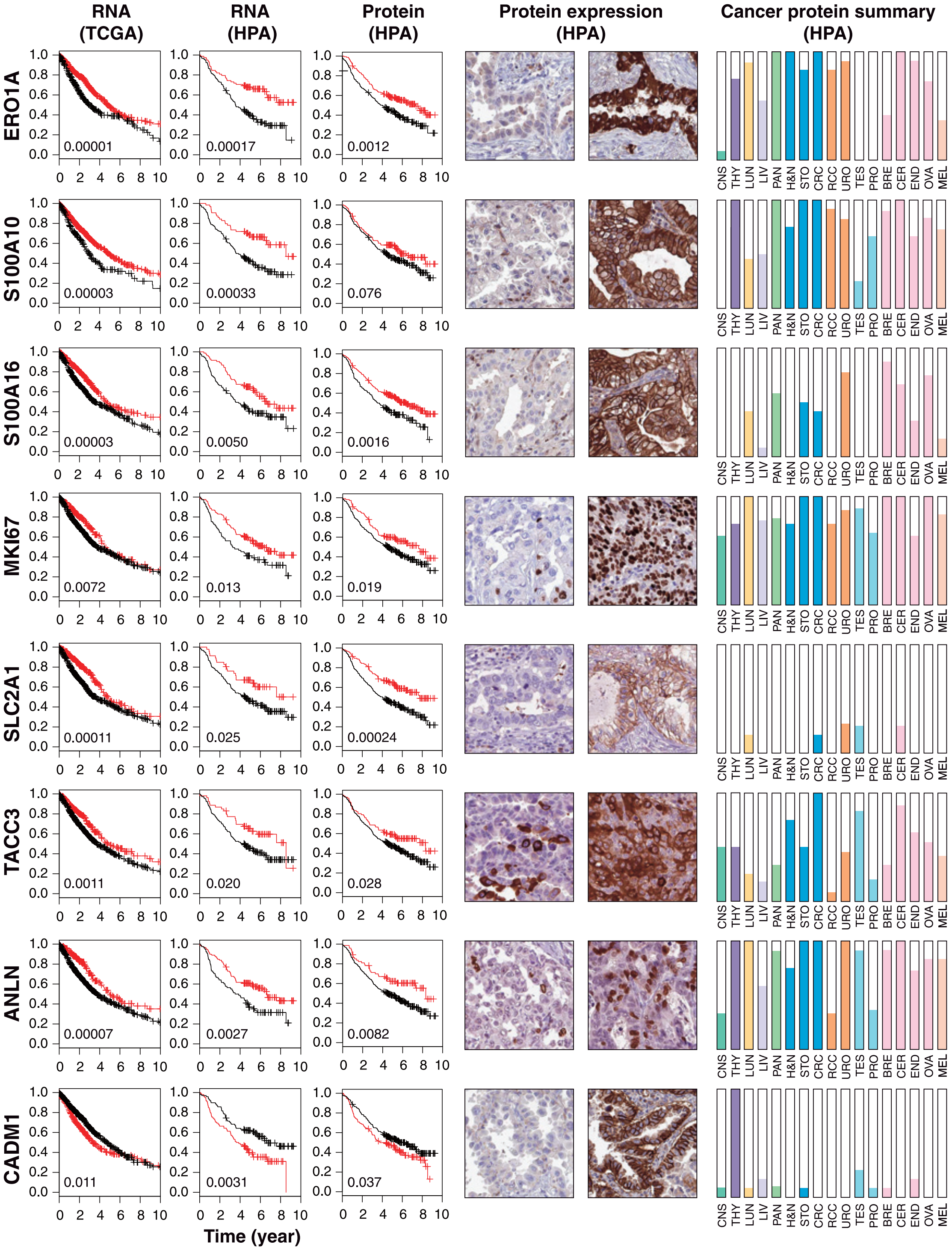
Fig. 6. Genome-scale metabolic models (GSMMs) of cancers. (**A**) Concept of personalized GSMMs, which are comprehensive compilations of all the metabolic reactions within a particular cell, tissue, organ, or organism. By mapping the transcriptomic data from cancer patients, personalized GSMMs could be reconstructed for investigation of the specific metabolic viabilities for each individual. (**B**) Heat map showing the essential enzymes in the TCA cycle for all glioma patients to exemplify the heterogeneity within the same cancer patient group. Only enzymes that were key in at least one patient are shown. (**C**) Bar plot showing the fraction of genes that were common in key genes in different proportions of patients for 17 Human Pathology Atlas cancers. (**D**) Circos plot showing the top 10 common metabolic pathways that were overrepresented by key genes in 17 Human Pathology Atlas cancers. Abbreviated names are provided in Fig. 1A and table S17.

When we investigated the associated biological functions, a vast majority of these genes were associated with central metabolic functions that are essential for normal tissues (Fig. 6D and table S16), and the corresponding proteins are thus not suitable as targets for drug development. Therefore, we performed toxicity tests using the models generated for healthy tissues and observed that, in many cases, the potential inhibition of 76 to 81% of these targets could be predicted to have severe side effects, because the target is essential in at least some normal tissues. Moreover, we also predicted that 32 gene targets that are mainly involved in nucleotide metabolism were predicted to be nontoxic in healthy tissues (fig. S12) but key in more than 80% of the tumor of the patient, regardless of the cancer type. These genes may therefore hold promise as potential targets for cancer treatment. In general, gene targets with less toxicity in normal tissue were key for tumor growth in fewer than 20% of cancer patients. Our analysis thus demonstrates the large heterogeneities in different cancer patients from a metabolic perspective and shows a path to individualized treatment of patients based on metabolic modeling, thereby highlighting the importance of systems-level analysis for precision cancer treatment.

Examination of genes in lung cancer

Further validation of prognostic genes identified through analyses of TCGA data was performed using an independent cohort of lung cancer

Fig. 7. Validation of selected genes with a prognostic effect in lung cancer. Kaplan-Meier plots for RNA level separation from the TCGA cohort, RNA level separation from the HPA cohort, and protein-level separation are shown in the first, second, and third columns, respectively. The log-rank *P* values are shown in the lower left corner of each Kaplan-Meier plot. IHC stained tissues representing high and low protein expression are shown in the fourth and fifth columns, respectively. The protein expression levels across 17 cancer types analyzed by IHC in the Human Pathology Atlas are shown at the right.



(NSCLC) patients (n = 357). We used available RNA-seq data from 199 individual tumors (30) and paraffin-embedded tumor tissue material in a tissue microarray (TMA) format from 357 patients (31). On the basis of transcriptomic data, the 100 most significant lung cancer prognostic genes identified in the TCGA analysis showed a high degree of overlap with prognostic genes in the independent NSCLC cohort (74% with P < 0.1, 45% with P < 0.01). In addition, the panel for lung cancer shown in Fig. 2A was also validated in this independent cohort (fig. S13).

To further investigate whether prognostic genes identified through genome-wide transcriptomics analyses could be verified at the protein level, we performed antibody-based IHC analyses of TMAs with tumor tissue (n = 357) for eight selected targets (Fig. 7). The IHC-based analysis confirmed that the corresponding protein expression pattern was also significantly associated with prognosis, and this was also supported by the RNA-seq data in the independent NSCLC cohort. Examples (Fig. 7) include the endoplasmic reticulum oxidoreductase α protein ERO1A (32) and two members of the S100 family (S100A10 and S100A16). The latter two proteins have been suggested as prognostic markers at the protein level in NSCLC adenocarcinoma (33, 34). We could confirm the prognostic association of both S100A10 and S100A16 in the NSCLC cohort containing both adenocarcinomas and squamous cell carcinomas. The proliferation marker MKI67 has been studied in a number of cancer types; however, its clinical application has been debated (35), and MKI67 has not been included in routine NSCLC diagnostics (36). In the present investigation, MKI67 was associated with an unfavorable prognosis in the TCGA data set, which was also confirmed at both the RNA and protein level in the independent NSCLC cohort. SLC2A1 (solute carrier family 2 member 1), also known as GLUT1, is a downstream gene of the hypoxic marker HIF1A and plays a role in glucose transport. TACC3 (transforming acidic coiled coil–containing protein 3) is involved in controlling normal cell growth and differentiation. Overexpression of SLC2A1 and TACC3 was previously associated with a poor prognosis in lung cancer (37, 38), and here we found that expression level associates with clinical outcome in lung cancer. Anillin (ANLN), an actin-binding protein required for cytokinesis, plays an important role in cell division and has been suggested as a prognostic marker in breast cancer (39) and lung cancer (40). Here, our TCGA analysis show prognostic value in lung, renal, pancreatic, and liver cancers, and the analysis of the independent lung cohort implies that this may be a favorable prognostic gene for clinical outcome.

Examination of genes in colon cancer

We investigated a large, independent, prospectively collected population-based cohort of colorectal cancer patients available in TMA format to assess possible prognostic protein signatures. In this cohort, mRNA expression data (RNA-seq) were also available for a smaller subset of the patients (n = 60). Six targets with prognostic significance in colorectal cancer based on TCGA data were selected for IHC staining on the TMAs. All six genes were verified as related to prognosis at both the RNA level (n = 60) and protein level (n = 745) (fig. S14).

The Human Pathology Atlas

As part of this publication, we launch a new open-access resource named the Human Pathology Atlas as part of the Human Protein Atlas (www.proteinatlas.org/pathology), presenting the Kaplan-Meier survival plots for all protein-coding genes in 17 different tumor types. A survival plot of the patient cohort, with the respective cancer and gene divided into two equal groups (median), is presented on the basis of RNA levels. More than 900,000 survival plots (as exemplified by Fig. 2C) are presented in the new pathology resource to allow investigators to explore the clinical significance of patient survival related to specific genes in specific cancers, together with the associated transcriptomic, proteomic, and clinical information. A total of 13,088 Kaplan-Meier plots with high significance (P < 0.001) are highlighted, and the data are presented in a gene-centric manner for all human protein-coding genes across the analyzed cancer types. Each prognostic gene for a given cancer type is shown, including the Kaplan-Meier plots (Fig. 2A), together with the underlying data for the selection of suitable FPKM cutoffs for patient stratification (Fig. 2B) and the individual survival data for all patients (Fig. 2B). In addition, IHC analysis using a TMA-based analysis of the corresponding proteins in patients with the respective cancer types is presented for a majority of the protein-coding genes. More than 5 million IHC-based cancer tissue images are included in the atlas, showing protein expression patterns for individual tumors of each cancer type. All IHC images have been manually annotated by certified pathologists. Thus, the resource allows researchers to explore the possible prognostic value of all human protein-coding genes related to expression levels in different forms of human cancer.

Discussion

Our results demonstrate the power of large systematic "big data" efforts that make use of publicly available resources, such as the TCGA database used herein. The compiled data show that a large fraction of human protein-coding genes are differentially expressed in cancer and that this differential expression in many cases has an impact on patient survival. Prognostic genes appear to be restricted to only a few cancer types, and no genes were informative across a large set of cancer patients. A general pattern emerged, with unfavorable genes showing an up-regulation associated with mitosis and cell growth, whereas the down-regulation of genes associated with cellular differentiation was associated with shorter patient survival. However, it is important to point out that for a given prognostic gene, we observe a huge variation in terms of clinical outcome for the corresponding patient, implying the need for further sophisticated studies to better comprehend the concept of prognostic genes.

The prognostic genes we identified should be validated in independent patient cohorts, as exemplified by the validation using antibody-based TMAs of a selection of the genes identified in lung cancer. The clinical metadata in the TCGA resource did not include therapeutic regimens for the patients, nor whether death was related to the diagnosed cancer. In addition, the different sample and effect sizes for different cancers would affect the number of prognostic genes obtained by survival analysis and log-rank test. Moreover, the purity of the tumor samples should also affect the survival analysis, as previously reported (41). Hence, there is a need for follow-up validation studies in more controlled independent cancer cohorts before a potential prognostic gene can be confirmed. An important quest for the near future is to identify which prognostic genes are functionally important ("drivers") with functional consequences that are required for carcinogenesis and tumor progression, and which of the apparent prognostic genes are merely coexpressed with these "driver" genes.

We generated cancer-specific coexpression networks to study the functional relationship between the prognostic genes and genes associated with Hallmarks of Cancer. This network-dependent analysis enabled the identification of genes with a key role in the survival of patients. The personalized genome-scale GSMMs allowed us to identify genes that were critical for tumor growth by demonstrating a huge heterogeneity among patients from a metabolic perspective, highlighting the need for precise and personalized medicine for cancer treatment. In this context, the new Human Pathology Atlas is a useful standalone resource for cancer precision medicine. With its more than 900,000 Kaplan-Meier plots, this resource enables insights concerning the specific involvement of genes in clinical outcome for all the major cancers, paving the way for further in-depth studies incorporating systems-level analyses of cancer. All data presented herein are available in an interactive open-access database (www.proteinatlas.org/pathology) to allow for genome-wide exploration of the impact of individual proteins on clinical outcome in major human cancer types.

REFERENCES AND NOTES

1. D. J. Brennan, D. P. O'Connor, E. Rexhepaj, F. Ponten, W. M. Gallagher, Antibody-based proteomics: Fast-tracking molecular diagnostics in oncology. *Nat. Rev. Cancer* **10**, 605–617 (2010). doi: 10.1038/nrc2902; pmid: 20720569
2. E. Björnson *et al.*, Stratification of hepatocellular carcinoma patients based on acetate utilization. *Cell Rep.* **13**, 2014–2026 (2015). doi: 10.1016/j.celrep.2015.10.045; pmid: 26655911
3. A. Mardinoglu, J. Nielsen, New paradigms for metabolic modeling of human cells. *Curr. Opin. Biotechnol.* **34**, 91–97 (2015). doi: 10.1016/j.copbio.2014.12.013; pmid: 25559199
4. S. Lee, A. Mardinoglu, C. Zhang, D. Lee, J. Nielsen, Dysregulated signaling hubs of liver lipid metabolism reveal hepatocellular carcinoma pathogenesis. *Nucleic Acids Res.* **44**, 5529–5539 (2016). doi: 10.1093/nar/gkw462; pmid: 27216817
5. J. N. Weinstein *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013). doi: 10.1038/ng.2764; pmid: 24071849

6. M. Uhlén *et al*., Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). doi: 10.1126/science.1260419; pmid: 25613900
7. J. Lonsdale *et al*., The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013). doi: 10.1038/ng.2653; pmid: 23715323
8. L. Collado-Torres *et al*., Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017). doi: 10.1038/nbt.3838; pmid: 28398307
9. L. Peng *et al*., Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Sci. Rep.* **5**, 13413 (2015). doi: 10.1038/srep13413; pmid: 26292924
10. F. Edfors *et al*., Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol. Syst. Biol.* **12**, 883 (2016). doi: 10.15252/msb.20167144; pmid: 27951527
11. D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011). doi: 10.1016/j.cell.2011.02.013; pmid: 21376230
12. C. Kandoth *et al*., Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013). doi: 10.1038/nature12634; pmid: 24132290
13. T. Hothorn, B. Lausen, On the exact distribution of maximally selected rank statistics. *Comput. Stat. Data Anal.* **43**, 121–137 (2003). doi: 10.1016/S0167-9473(02)00225-6
14. B. Hjelm *et al*., High nuclear RBM3 expression is associated with an improved prognosis in colorectal cancer. *Proteomics Clin. Appl.* **5**, 624–635 (2011). doi: 10.1002/prca.201100020; pmid: 21956899
15. C. J. Creighton *et al*., Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013). doi: 10.1038/nature12222; pmid: 23792563
16. A. Subramanian *et al*., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005). doi: 10.1073/pnas.0506580102; pmid: 16199517
17. H. A. Edmondson, P. E. Steiner, Primary carcinoma of the liver: A study of 100 cases among 48,900 necropsies. *Cancer* **7**, 462–503 (1954). doi: 10.1002/1097-0142(195405)7:3<462::AID-CNCR2820070308>3.0.CO;2-E; pmid: 13160935
18. T. M. Pawlik *et al*., Preoperative assessment of hepatocellular carcinoma tumor grade using needle biopsy: Implications for transplant eligibility. *Ann. Surg.* **245**, 435–442 (2007). doi: 10.1097/01.sla.0000250420.73854.ad; pmid: 17435551
19. A. J. Simpson, O. L. Caballero, A. Jungbluth, Y. T. Chen, L. J. Old, Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer* **5**, 615–625 (2005). doi: 10.1038/nrc1669; pmid: 16034368
20. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017). doi: 10.1093/nar/gkw1092; pmid: 27899662
21. T. I. Zack *et al*., Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013). doi: 10.1038/ng.2760; pmid: 24071852
22. N. N. Pavlova, C. B. Thompson, The emerging hallmarks of cancer metabolism. *Cell Metab.* **23**, 27–47 (2016). doi: 10.1016/j.cmet.2015.12.006; pmid: 26771115
23. M. G. Vander Heiden, R. J. DeBerardinis, Understanding the intersections between metabolism and cancer biology. *Cell* **168**, 657–669 (2017). doi: 10.1016/j.cell.2016.12.039; pmid: 28187287
24. P. Ghaffari, A. Mardinoglu, J. Nielsen, Cancer metabolism: A modeling perspective. *Front. Physiol.* **6**, 382 (2015). doi: 10.3389/fphys.2015.00382; pmid: 26733270
25. A. Mardinoglu *et al*., Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 3083 (2014). doi: 10.1038/ncomms4083; pmid: 24419221
26. R. Agren *et al*., Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.* **10**, 721 (2014). doi: 10.1002/msb.145122; pmid: 24646661
27. A. Mardinoglu *et al*., Personal model-assisted identification of NAD(+) and glutathione metabolism as intervention target in NAFLD. *Mol. Syst. Biol.* **13**, 916 (2017). doi: 10.15252/msb.20167422; pmid: 28254760
28. L. Jerby-Arnon *et al*., Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell* **158**, 1199–1209 (2014). doi: 10.1016/j.cell.2014.07.027; pmid: 25171417
29. C. Zhang, Q. Hua, Applications of genome-scale metabolic models in biotechnology and systems medicine. *Front. Physiol.* **6**, 413 (2016). doi: 10.3389/fphys.2015.00413; pmid: 26779040
30. D. Djureinovic *et al*., Profiling cancer testis antigens in non-small-cell lung cancer. *Jci Insight* **1**, e86837 (2016). doi: 10.1172/jci.insight.86837; pmid: 27699219
31. P. Micke *et al*., The impact of the Fourth Edition of the WHO Classification of Lung Tumours on histological classification of resected pulmonary NSCCs. *J. Thorac. Oncol.* **11**, 862–872 (2016). doi: 10.1016/j.jtho.2016.01.020; pmid: 26872818
32. T. Tanaka *et al*., Cancer-associated oxidoreductase ERO1-α drives the production of VEGF via oxidative protein folding and regulating the mRNA level. *Br. J. Cancer* **114**, 1227–1234 (2016). doi: 10.1038/bjc.2016.105; pmid: 27100727
33. K. Katono *et al*., Clinicopathological significance of S100A10 expression in lung adenocarcinomas. *Asian Pac. J. Cancer Prev.* **17**, 289–294 (2016). doi: 10.7314/APJCP.2016.17.1.289; pmid: 26838226
34. K. Saito *et al*., S100A16 is a prognostic marker for lung adenocarcinomas. *Asian Pac. J. Cancer Prev.* **16**, 7039–7044 (2015). doi: 10.7314/APJCP.2015.16.16.7039; pmid: 26514487
35. F. Penault-Llorca, N. Radosevic-Robin, Ki67 assessment in breast cancer: An update. *Pathology* **49**, 166–171 (2017). doi: 10.1016/j.pathol.2016.11.006; pmid: 28065411
36. J. N. Jakobsen, J. B. Sørensen, Clinical impact of Ki-67 labeling index in non-small cell lung cancer. *Lung Cancer* **79**, 1–7 (2013). doi: 10.1016/j.lungcan.2012.10.008; pmid: 23137549
37. M. Younes, R. W. Brown, M. Stephenson, M. Gondo, P. T. Cagle, Overexpression of Glut1 and Glut3 in stage I nonsmall cell lung carcinoma is associated with poor survival. *Cancer* **80**, 1046–1051 (1997). doi: 10.1002/(SICI)1097-0142(19970915)80:6<1046::AID-CNCR6>3.0.CO;2-7; pmid: 9305704
38. C. K. Jung *et al*., Expression of transforming acidic coiled-coil containing protein 3 is a novel independent prognostic marker in non-small cell lung cancer. *Pathol. Int.* **56**, 503–509 (2006). doi: 10.1111/j.1440-1827.2006.01998.x; pmid: 16930330
39. K. Magnusson *et al*., ANLN is a prognostic biomarker independent of Ki-67 and essential for cell cycle progression in primary breast cancer. *BMC Cancer* **16**, 904 (2016). doi: 10.1186/s12885-016-2923-8; pmid: 27863473
40. C. Suzuki *et al*., ANLN plays a critical role in human lung carcinogenesis through the activation of RHOA and by involvement in the phosphoinositide 3-kinase/AKT pathway. *Cancer Res.* **65**, 11314–11325 (2005). doi: 10.1158/0008-5472.CAN-05-1507; pmid: 16357138
41. D. Aran, M. Sirota, A. J. Butte, Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015). doi: 10.1038/ncomms9971; pmid: 26634437

ACKNOWLEDGMENTS

The data reported in this manuscript are tabulated in the main paper and the supplementary materials. We thank the entire staff of the Human Protein Atlas program and the Science for Life Laboratory, the National Genomics Infrastructure, and Swedish National Infrastructure for Computing at UPPMAX and C3SE for providing assistance in massive parallel sequencing and computational infrastructure. Supported by the Swedish Cancer Foundation (F.P.), Science for Life Laboratory infrastructure funding (C.L., E.L. and P.N.), the Erling Persson Foundation (M.U.), Elixir (EU infrastructure funding) (K.F. and M.U.), the Swedish Cancer Foundation and the Uppsala Lions Cancer Foundation (P.M.), the Knut and Alice Wallenberg Foundation, and the strategic research area “U-CAN” through Swedish Research Council grant CancerUU. We thank the Uppsala Biobank and Clinical Department of Pathology at the Uppsala University Hospital for providing specimens. We also thank The Cancer Genome Atlas for providing access to data. M.U. and F.P. are inventors on patents held by Atlas Antibodies that cover the analysis of RBM3 (U.S. patent 8728739) and PODXL (U.S. patent 8999656) as a prognostic marker in several cancers. The authors declare that they have no other conflict of interest. An interactive pathology atlas with gene-specific data for all human genes (including more than 900,000 Kaplan-Meier plots) is available at www.proteinatlas.org/pathology.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/357/6352/eaan2507/suppl/DC1
Materials and Methods
Figs. S1 to S14
Tables S1 to S21
References (*42*–*51*)

22 March 2017; resubmitted 2 June 2017
Accepted 14 July 2017
10.1126/science.aan2507

BIOCHEMISTRY

The human secretome

Mathias Uhlén[1,2,3]*, Max J. Karlsson[1], Andreas Hober[1], Anne-Sophie Svensson[4], Julia Scheffel[4], David Kotol[1], Wen Zhong[1], Abdellah Tebani[1], Linnéa Strandberg[1], Fredrik Edfors[1,5], Evelina Sjöstedt[3], Jan Mulder[3], Adil Mardinoglu[1], Anna Berling[4], Siri Ekblad[4], Melanie Dannemeyer[4], Sara Kanje[4], Johan Rockberg[4], Magnus Lundqvist[4], Magdalena Malm[4], Anna-Luisa Volk[4], Peter Nilsson[1], Anna Månberg[1], Tea Dodig-Crnkovic[1], Elisa Pin[1], Martin Zwahlen[1], Per Oksvold[1], Kalle von Feilitzen[1], Ragna S. Häussler[1], Mun-Gwan Hong[1], Cecilia Lindskog[6], Fredrik Ponten[6], Borbala Katona[6], Jimmy Vuu[6], Emil Lindström[6], Jens Nielsen[7], Jonathan Robinson[7], Burcu Ayoglu[1], Diana Mahdessian[1], Devin Sullivan[1], Peter Thul[1], Frida Danielsson[1], Charlotte Stadler[1], Emma Lundberg[1], Göran Bergström[8,9], Anders Gummesson[8], Bjørn G. Voldborg[2], Hanna Tegel[4], Sophia Hober[4], Björn Forsström[1], Jochen M. Schwenk[1], Linn Fagerberg[1], Åsa Sivertsson[1]

The proteins secreted by human cells (collectively referred to as the secretome) are important not only for the basic understanding of human biology but also for the identification of potential targets for future diagnostics and therapies. Here, we present a comprehensive analysis of proteins predicted to be secreted in human cells, which provides information about their final localization in the human body, including the proteins actively secreted to peripheral blood. The analysis suggests that a large number of the proteins of the secretome are not secreted out of the cell, but instead are retained intracellularly, whereas another large group of proteins were identified that are predicted to be retained locally at the tissue of expression and not secreted into the blood. Proteins detected in the human blood by mass spectrometry–based proteomics and antibody-based immunoassays are also presented with estimates of their concentrations in the blood. The results are presented in an updated version 19 of the Human Protein Atlas in which each gene encoding a secretome protein is annotated to provide an open-access knowledge resource of the human secretome, including body-wide expression data, spatial localization data down to the single-cell and subcellular levels, and data about the presence of proteins that are detectable in the blood.

INTRODUCTION

An important class of human proteins are those that are actively transported within the secretory pathway for destinations outside the cytoplasm and nucleus of the cell. The collection of actively secreted proteins, herein referred to as the "human secretome," constitutes a large fraction of the targets for pharmaceutical drugs, but these are also important as diagnostic targets both for classical clinical chemistry and as potential targets for future precision medicine efforts (*1*, *2*). Many of these proteins are also involved in signaling functions both locally and systemically, including proteins such as cytokines, growth factors, and hormones. Despite the huge interest in this class of proteins, there have been few attempts to define the size and constituents of the entire human secretome. A prediction of the number of putatively secreted proteins, defined as having a signal sequence and no transmembrane regions, was previously estimated to correspond to 2918 protein-coding genes, thus involving approximately 15% of all human genes (*3*). Attempts to characterize the proteins present in blood (the "plasma proteome") have led to the development of multiplex assays involving thousands of protein targets using nucleic acid–based technologies (*4*, *5*), immune-based assays (*6*), or mass spectrometry (MS) (*7*, *8*). However, these plasma proteome efforts do not normally distinguish between actively secreted proteins (here defined as being part of the secretome) and proteins that are leaked by the millions of cells undergoing cell death at any given moment. In addition, many of the proteins secreted from human cells are not destined for the peripheral blood.

We therefore decided to stratify the actively secreted proteins in humans to define the spatial distribution of each protein with regard to its origin of expression and to provide a comprehensive list of annotated proteins based on their final localization in the body. Starting with a bioinformatics definition of the secretome, the proteins were classified into three major categories: (i) the blood proteins, (ii) the locally secreted proteins, and (iii) the intracellular proteins. The latter might sound counterintuitive, but this category reflects the fact that many proteins secreted into the endoplasmic reticulum (ER) are sorted to various intracellular compartments, such as mitochondria and lysosomes, or they are even retained in the ER or Golgi. These proteins are thus not secreted out of the cell. The locally secreted proteins can further be classified into the various sublocalizations, such as the brain, as well as male and female tissues, and they also include proteins secreted to the digestive tract or those that end up in the extracellular matrix. Last, we annotated the proteins predicted to end up in peripheral blood, thus having important effects in the

[1]Department of Protein Science, Science for Life Laboratory, KTH–Royal Institute of Technology, Stockholm, Sweden. [2]Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark. [3]Department of Neuroscience, Karolinska Institute, Stockholm, Sweden. [4]Department of Protein Science, AlbaNova University Center, KTH–Royal Institute of Technology, Stockholm, Sweden. [5]Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA. [6]Department of Pathology, Uppsala University, Uppsala, Sweden. [7]Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden. [8]Department of Molecular and Clinical Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [9]Region Västra Götaland, Sahlgrenska University Hospital, Department of Clinical Physiology, Gothenburg, Sweden.
*Corresponding author. Email: mathias.uhlen@scilifelab.se

Originally published 26 November 2019

human body, contributing to the systems-level control of homeostasis, transport of nutrients, inflammatory response, defense mechanisms, hormone regulation, and many other functions. To complement this annotation of the human secretome, we also investigated the proteins found in blood using various technology platforms. All results are presented in a new version of the Human Protein Atlas (HPA) (www.proteinatlas.org/blood) with a "secretome" part including data on each gene encoding these proteins.

RESULTS

Revised list of the human secretome

We first performed a protein-centric transcriptomics scan to define a revised set of human secreted proteins (secretome) based on 19,670 protein-coding genes predicted by Ensembl (*9*). For each protein-coding gene, all protein isoforms (splice variants) were annotated (Fig. 1A) on the basis of the presence of a signal peptide, transmembrane regions, or both, and each protein isoform was classified as being secreted, membrane bound, or intracellular. The secreted proteins are here defined according to the HPA classification (*3*) as those proteins that have a signal peptide but lack a transmembrane region. The 338 Ensembl immunoglobulin genes (file S1) were excluded because of their complex genetic structure, which consists of many partial genes with variable (V), diversity (D), and joining (J) regions. In addition, proteins annotated as being secreted by UniProt (*10*) were added, which resulted in a list of 3513 genes with at least one predicted secreted isoform (transcript). From this list, proteins with no corresponding entry in the current Ensembl version (v92) and genes for which the predicted secreted isoform has low evidence for existence were excluded. The revised number of genes encoding potentially secreted proteins (the human secretome) was 2641 (file S2). This list comprises approximately 13% of all human protein-coding genes, and it serves as a resource for all researchers interested in secreted proteins as targets for diagnostic and therapeutic drugs.

Annotation of the human secretome into localization classes

On the basis of the published literature, bioinformatics analysis, and experimental evidence, all genes coding for the putative secretome were subsequently manually annotated, taking into account their spatial distribution in the human body. Starting with a bioinformatics definition of the secretome, the proteins were classified into three major categories: (i) the blood proteins, (ii) the locally secreted proteins, and (iii) the intracellular proteins. For the latter category, bioinformatics evidence for intracellular locations was also used, such as ER retention signals (*11*), a peroxisome-targeting signal (*12*), and a mitochondrial target signal (*13*). The locally secreted proteins were further subdivided into seven classes to yield altogether nine protein categories (Fig. 1A). Note that the annotation considers the possible functional role of each protein and, as an example, the well-known plasma protein PSA (*KLK3*) was here annotated as being localized to male reproductive tissue and not the blood, because it was here assumed that the primary functional role of this protein is in the prostate. The annotation of all 2641 protein-coding genes identified 730 proteins as blood secretome proteins, and about 500 were annotated as being secreted to local compartments, that is, male or female reproductive tissues, the brain, or other tissues, such as the eye or the skin. Eighty-eight proteins were identified as being secreted to the gastrointestinal tract, including 30 proteins produced in the pancreas and 25 in the salivary glands (fig. S1). More than 200 proteins were further annotated to belong to the group of proteins involved in the forming and function of the extracellular matrix, including both structural proteins, such as laminins, collagens, elastin, and fibronectin, and the matricellular proteins (*14*). More than 900 proteins were annotated to be intracellular or membrane related, with 254 of these proteins predicted to be localized to the Golgi, ER, or both and 270 proteins being membrane associated. Last, about 170 genes encoded proteins that lacked supporting data for their location. These are again interesting proteins for further studies to explore their function and location. A list of the genes in each category was compiled (file S2) and is also available at the open-access Blood Atlas resource (www.proteinatlas.org/blood). Given that 932 genes were annotated to encode intracellularly localized proteins, despite having a predicted secreted isoform, the genome-wide annotation of the subcellular location of the protein-coding genes was revised (Fig. 1B). The resulting list of the human secretome consists of 1709 genes with at least one isoform (transcript) coding for a predicted secreted protein. Note that many genes encode multiple transcripts predicted to be in more than one location.

Functional analysis of those proteins predicted to be actively secreted to the blood

A functional analysis (Fig. 1C) of the 730 proteins predicted to be secreted to human blood revealed many well-characterized proteins, such as cytokines, interleukins, interferons, and chemokines (altogether 154 proteins), complement and coagulation factors (n = 68 proteins), hormones (n = 75), growth factors (n = 33), and enzymes (n = 83). Many of these proteins are interesting pharmaceutical targets, and 72 are already the products or targets of U.S. Food and Drug Administration (FDA)–approved drugs (file S3) (*15*). Furthermore, almost 100 of the predicted secreted blood proteins have currently no functional annotation in UniProt; thus, these are interesting targets for further exploration of their functional role in the blood.

The tissue distribution of the human secretome proteins

The tissue expression pattern of the human secretome genes was subsequently analyzed on the basis of transcriptomics data from the HPA. A classification according to tissue specificity as previously described (*3*) was performed for all of the secretome genes in the different annotation categories. The results showed that genes predicted to be locally expressed mainly consisted of tissue-enriched genes. Furthermore, genes encoding intracellular proteins generally showed low tissue specificity (Fig. 2A), which suggests that they have a more "house-keeping" role in the cell. Similarly, the matrix proteins are widely expressed across the analyzed tissues. A large portion of the blood proteins are produced in the liver, whereas others are generated either by blood cells or more generally across all tissues.

A cluster analysis (*16*) suggests that the expression profiles across the 730 blood proteins can be stratified into nine expression clusters (Fig. 2, B and C), and these can further be grouped according to tissue origin, including 139 proteins mainly originating from the liver (clusters A and D) with classical plasma proteins, such as albumin, transferrin, the apolipoproteins, and complement factors. Another group consists of the 174 proteins mainly originating from human blood cells, lymphoid tissues, or both (clusters C and F), which include many chemokines and granzymes. The 60 proteins that are mainly enriched in the brain (cluster E) constitute a third group to which several hormones and neuropeptides belong, including oxytocin, gonadotropin-releasing hormone, and pro-melanin

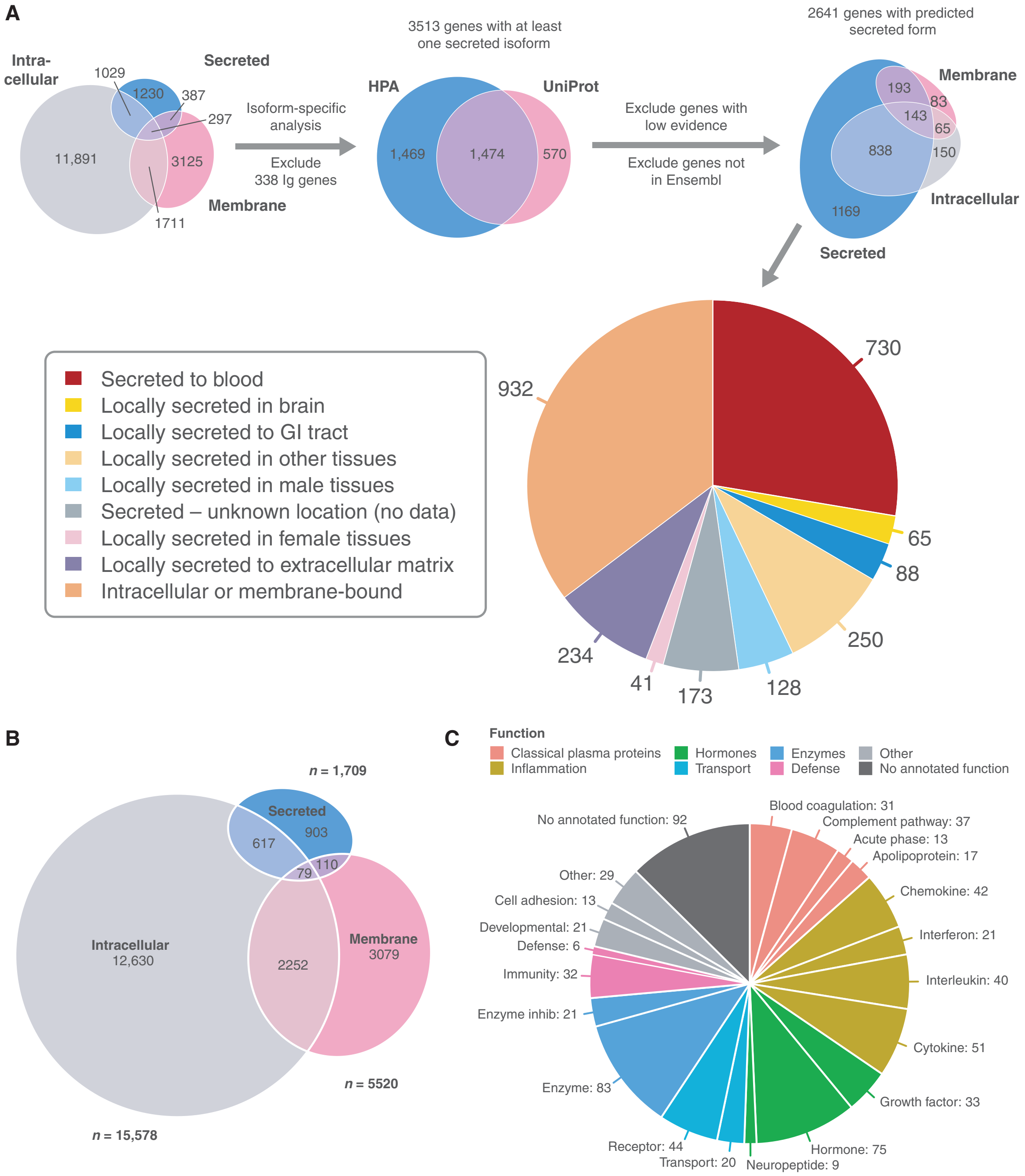


Fig. 1. Annotation of the human secretome. (**A**) Workflow used to predict the final location of components of the human secretome ($n = 2641$ proteins). The proteins are classified according to their predicted final location as seen in the pie chart (see file S2 for a complete list). (**B**) Predicted number of genes in the three indicated categories after considering the intracellular and membrane-bound annotations. (**C**) Functional analysis of the 730 proteins annotated as being blood proteins. The color code represents eight functional classes, including the subclasses denoted in the pie chart.

concentrating hormone. A fourth group includes 104 proteins with selective expression (cluster B), including interferons and hormones produced in specialized tissues, such as the placenta or endocrine glands. Last, there is a large group of 253 proteins that are expressed ubiquitously across many tissues (clusters G, H, and I). Many of these proteins are highly expressed in endothelial cells present throughout the body, which could explain their diverse expression, but it is also possible that some of the expression is related to the nonsecreted isoforms of the respective gene products and more in-depth studies are needed to resolve the isoform-specific transcript abundances for these genes. The relationship between the clusters and protein functions are shown as a chord diagram (Fig. 2D), demonstrating that the genes in cluster B mainly encode cytokines and growth factors, whereas genes in cluster A encode not only complement, coagulation factors, and acute phase proteins but also transport proteins and enzymes.

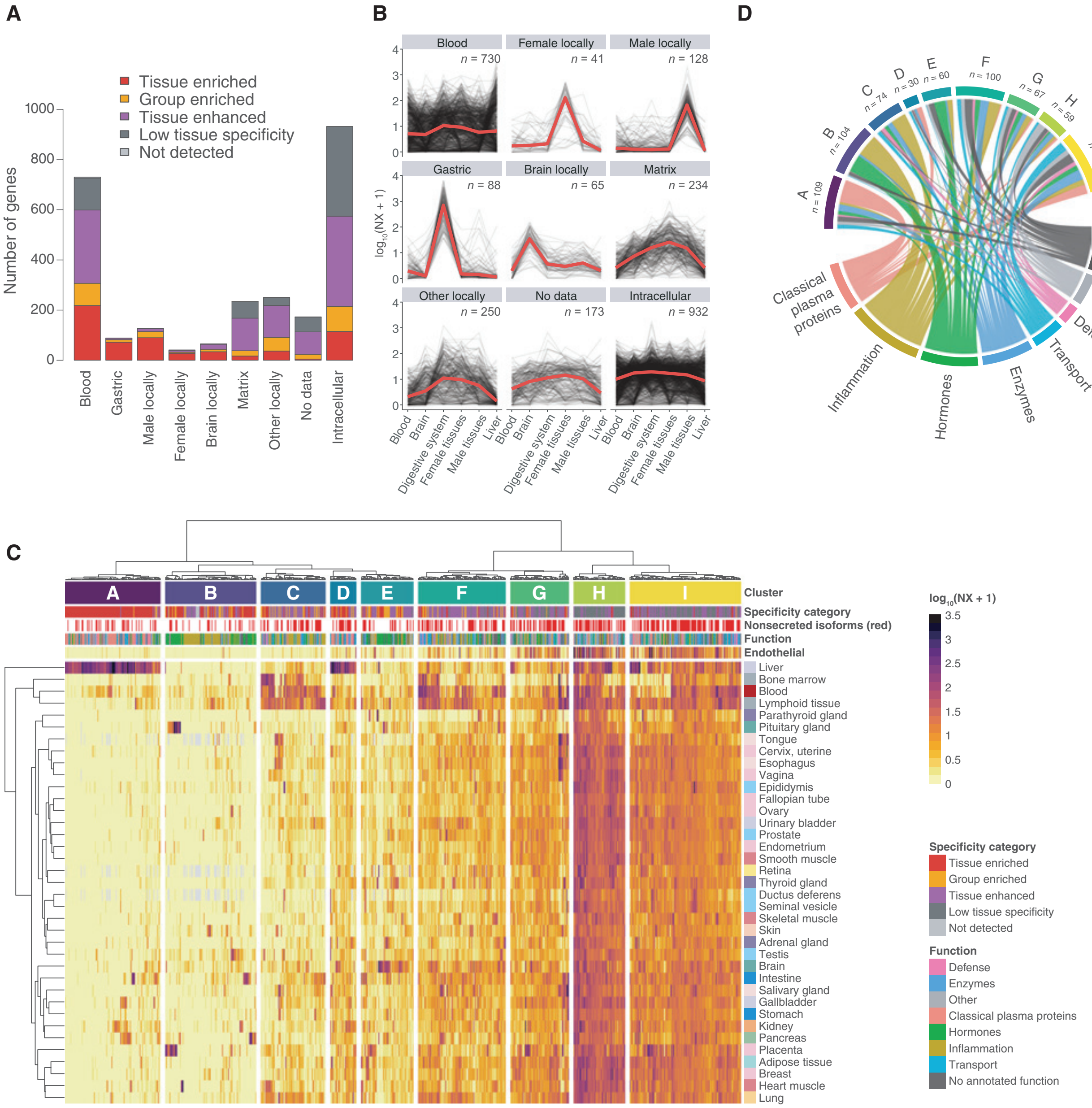
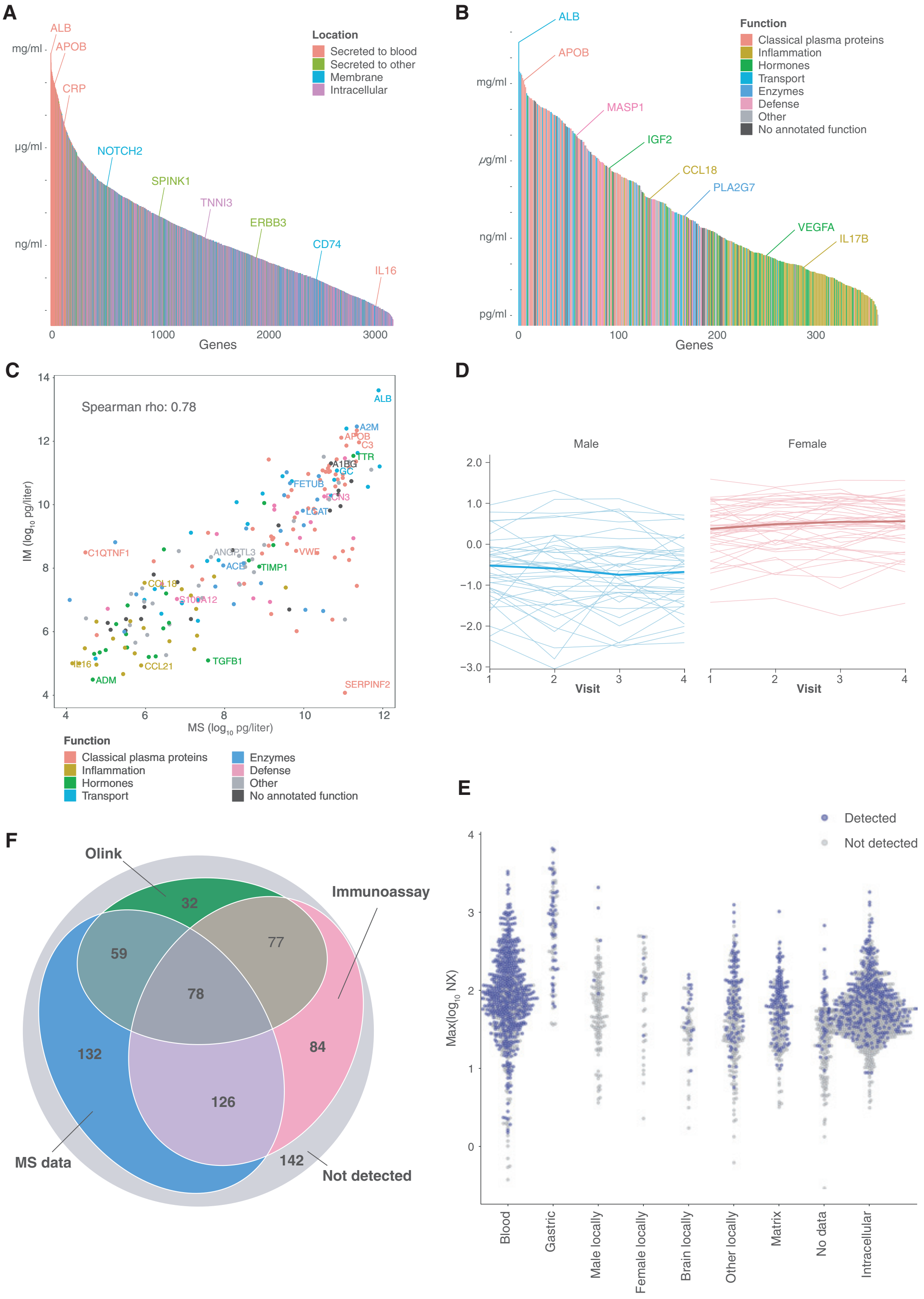


Fig. 2. The tissue distribution of the human secretome proteins. (**A**) Tissue specificity classification based on transcriptomics data for the genes in the indicated annotation categories. (**B**) Expression values (NX) across six selected tissues (see fig. S2 for the corresponding heatmaps) for the nine annotation categories (A to H). The number of protein-coding genes in each category is shown (*n*). (**C**) Heatmap based on the relative expression of all blood secretome protein-coding genes ($n = 730$) with nine expression clusters shown on top. (**D**) Chord diagram showing the relationship between the functional annotation of the 730 blood secretome proteins (bottom) with the nine expression clusters (top).

Proteins detected in blood by MS

We subsequently decided to investigate the presence of the predicted secretome proteins in human blood based on MS and antibody-based immunoassays and to complement this with "in-house" analysis using a sensitive proximity extension assay (PEA) (*17*). For MS-based proteomics, protein concentrations in blood (Fig. 3A) were inferred by spectral counting (*18*) from a combined dataset of more than 170 publicly available experiments hosted in the Human Plasma PeptideAtlas (*19*). More than 3000 proteins were detected in this diverse set of experiments (file S4). The detected proteins were colored according to their predicted localization, with proteins annotated to be intracellular, membrane bound, or secreted and with the latter group stratified into those secreted to blood and those secreted to other compartments (according to Fig. 1). In this regard, note that there are a few abundant proteins that make up most of the protein mass in blood, with 99% of the protein mass corresponding to only a few proteins (*20*). However, many proteins are present at lower concentrations either resulting from active secretion or proteins leaked from cells due to normal cell turnover or pathology involving various diseases. The analysis showed that most of the proteins detected by MS were here defined as leakage proteins.

Fig. 3. Proteins detected in human blood. The predicted amounts of proteins detected in human blood by three different technology platforms. (**A**) Plasma concentrations based on MS assays for proteins detected in blood. The plot shows plasma concentrations estimated from spectral counting of 3223 proteins belonging to MS experiments in the Human Plasma PeptideAtlas. The bars are colored on the basis of the classification of the corresponding proteins into one of four localization categories: secreted to blood, secreted to other, membrane bound, or intracellular. Some examples of proteins are shown as reference. (**B**) Plasma concentrations based on immunoassays from reference articles for 365 proteins actively secreted to blood. The bar plot shows ranked median plasma concentrations with bars colored on the basis of the classification of the corresponding proteins into one of the eight indicated functional categories. (**C**) Plot showing the correlation of the estimated concentrations of 205 blood secretome proteins with data from both proteomics and immunoassays. (**D**) The protein profiles of 86 individuals during a 1-year period were analyzed using a set of multiplex PEAs targeting 748 proteins. The image shows an example of the average amounts of a protein (leptin) for four visits spanning 1 year. The individuals were stratified according to females (red) and males (blue). (**E**) Maximal transcript abundances (NX) of the tissue of origin for all secretome proteins stratified according to annotated localization. Each protein is visualized as to whether it was detected (blue) or not detected (gray) in at least one of the assay platforms. (**F**) Number of blood secretome proteins detected by the three different methods.



Proteins detected in the blood by antibody-based immunoassays

We also compiled the plasma concentrations of proteins as detected by antibody-based immunoassays with a focus on the 730 proteins here defined as actively secreted to blood. A literature search was performed for each protein, and references were collected for those studies that reported the absolute concentrations in human plasma. In total, reference plasma or serum concentrations were found for 365 of the 730 proteins secreted to the blood, and these values are presented on the Blood Atlas page of each gene and are combined here for all of these proteins (Fig. 3B). The proteins are colored according to the functional annotation, and the results showed the expected high abundances for coagulation and complement factors, whereas cytokines, hormones, and growth factors often are present at low concentrations. Note that some of the actively secreted proteins with no annotated function were detected at relatively high concentrations in the blood. A comparison of the concentrations for 205 proteins determined by both proteomics and immunoassays showed a good correlation with a Spearman's ρ value of 0.79 (Fig. 3C).

Proteins detected in the blood by PEAs

Last, we analyzed blood samples from a healthy cohort with a multiplex PEA (*17*), covering 748 proteins involving both secreted and leakage products to investigate the longitudinal variability of these proteins during 1 year. As an example, the abundances of the protein leptin were measured for 86 individuals across 1 year and four consecutive visits with 3-month intervals (Fig. 3D). Although the average protein abundances are greater in females, the leptin concentrations of individuals were highly variable, with several males having consistently higher amounts during the assay period. These data support an individual-based definition of reference values for diagnostic applications and suggest that caution should be taken when using population-based data as a reference for health. The protein abundances across the individuals during 1 year for all 748 proteins analyzed are presented as part of the new Blood Atlas.

We subsequently assessed the detectability of each of the 2641 secretome proteins based on the maximal transcript abundance in their respective tissue of origin (Fig. 3E). As expected, most of the proteins classified as blood proteins have been detected in blood, but it is also evident that many of the proteins classified in the other categories have been detected in blood, in particular many intracellular proteins and matrix proteins. The question arises as to whether this is due to these proteins having functional roles as circulating proteins or due to their leakage from the extracellular matrix. Furthermore, many of the proteins not detected by any assay were found to have low abundances in their predicted tissue of origin. On the basis of these three assay platforms, the overall detection of proteins secreted into the blood (*n* = 730) was summarized (Fig. 3F). Note that 142 of the blood secretome proteins were not detected (or lacked specific assays) across all three platforms, and none of the assay platforms was able to detect more than half of the blood secretome proteins. This finding demonstrates the importance of developing assays for this important group of blood proteins and highlights the need for systematic efforts to develop assays for the "missing proteins" by respective assay platform.

DISCUSSION

Here, we present an analysis of all of the proteins predicted to be secreted in humans based on sequence analysis of the corresponding transcripts, including all protein isoforms having a signal peptide and no transmembrane-spanning regions. We identified 2641 genes through this bioinformatics approach, and these genes were subsequently annotated individually to reflect the destinations of their products in the human body. The annotation provides a view of the actively secreted proteins in humans; however, the quality of the annotation varies across the different proteins, mainly due to lack of experimental evidence, in particular limited information regarding genes with multiple transcript isoforms. The annotations of the individual proteins presented here will thus be revised when more information becomes available in the future. The results are presented in an updated version 19 of the HPA (www.proteinatlas.org/blood) to provide an open-access knowledge resource of the human secretome.

Some conclusions can be made from the results presented here. First, note that the number of predicted actively secreted blood proteins is surprisingly low (*n* = 730), corresponding to less than 4% of all human protein-coding genes. Among the proteins identified in this category of the secretome are not only the classical plasma proteins, the inflammation proteins (cytokines and interleukins), well-known hormones, and receptors but also close to 100 proteins with no annotated function yet. Similarly, 88 proteins were annotated to be secreted to the gastric tract, which again is a low number, but the list includes not only well-known digestive enzymes and defense proteins but also proteins that are much less studied. The analysis also revealed that the largest number of proteins was annotated to remain locally after secretion, including many matrix proteins (*n* = 234) and proteins that end up in specific tissues, including the brain and male and female tissues. However, the most unexpected finding from our analysis is that more than one-third (*n* = 932) of the proteins are predicted not to be secreted out of the cell. This intracellular category reflects the fact that many proteins are retained in the ER, Golgi, or both or are sorted to various intracellular compartments. More in-depth studies are needed to annotate the functions of many of these proteins.

The proteins detected in human blood were also analyzed, and the study resulted in an annotated list of proteins that have thus far been detected by MS-based proteomics and antibody-based immunoassays. The analysis revealed that for a large fraction of the secreted blood proteins, assays are lacking and a quest for the future is therefore to extend the respective assays to cover all the actively secreted proteins defined here, including relevant leakage proteins, to provide a secretome-wide toolbox of assays for the proteins in the blood to expand our capabilities for functional analysis and precision medicine efforts. In conclusion, the open-access resource presented here includes a genome-wide classification of all protein-coding genes with regard to the predicted spatial location in the human body of the corresponding proteins, as well as showing the cellular and/or tissue origin of each of the secretome proteins to facilitate basic and applied research involving this important class of proteins.

MATERIALS AND METHODS

Definition of the human secretome

The human secretome gene set includes Ensembl genes with either at least one splice variant encoding a protein with a signal peptide and no transmembrane regions according to HPA predictions (*3*) or for which UniProt has at least one isoform that is annotated as being secreted. From this gene set, the immunoglobulin-encoding genes were removed as were genes that encoded proteins for which there was little evidence of a secreted isoform.

Annotation of the human secretome

The genes were annotated into one of nine different categories based on the published literature, subcellular localization data from the HPA and UniProt, functional data from UniProt and HPA, RNA expression data from HPA, GTEx, and FANTOM, and protein expression data from HPA and UniProt. The annotation was performed using an in-house annotation system displaying functional, expression, and subcellular location data from the different sources and links to relevant publications for each gene and the corresponding ensemble gene model with predicted signal peptides and transmembrane regions based on several algorithms (fig. S3). The annotation was made on a gene level, and for genes that were found to have both secreted and intracellular or membrane-bound protein isoforms, only the secreted variant was annotated. For each gene, a single category was selected together with an explanatory comment and references to relevant articles. A short description and the hierarchy of the categories are shown in table S1.

Collection of protein plasma concentration references based on immunoassays

A literature search was performed for the 730 proteins annotated as being actively secreted to the blood, and the plasma concentrations of these proteins from up to three independent references were compiled. The search was limited to studies in which the protein concentrations had been determined using an antibody-based assay such as enzyme-linked immunosorbent assay, radioimmunoassay, or immunoturbidimetry. If reference concentrations from a mixed-sex healthy control group could be found, then these values were used, but if no such values were available, then concentrations from condition-specific groups were accepted, for example, from pregnant women or different diseases. In some studies, the plasma concentrations were not explicitly stated and median concentrations had to be estimated from plots in the articles. The complete list of collected plasma protein concentrations and their corresponding reference articles can be found in file S5.

Retrieving MS-based plasma concentration data from the PeptideAtlas

For MS-based plasma proteomics, the data from >170 studies present in the PeptideAtlas (www.peptideatlas.org) were used. We queried for the "Human Plasma Non-Glyco 2017-04" build and set presence levels "canonical," "no redundant relationships" for redundancy, and "show estimated abundances" as display options. This search revealed 3694 entries, from which the 22 labeled as contaminants ("CONTAM") were removed. Of the remaining 3672 entries, 3484 (93%) were listed with a protein concentration estimated by spectral counts column ("estimated_ng_per_ml") (*18*). These data were subsequently mapped to the ensemble gene set used in the HPA, and the 3223 remaining protein-coding genes were used in the analysis.

Plasma profiling using PEA

Plasma from 86 healthy individuals being part of the Swedish SCAPIS SciLifeLab Wellness Profiling (S3WP) (*21*) project was used for this analysis. The S3WP study is an observational study with the aim to collect longitudinal data in a community-based cohort, and the subjects were recruited from the ongoing Swedish Cardio Pulmonary bioImage Study (SCAPIS) (*22*), which includes randomly selected subjects aged 50 to 65 years from the general Swedish population. Before sampling, all subjects fasted overnight, for at least 8 hours. Blood samples were drawn in EDTA tubes using standard operating procedures. All samples were stored at −80°C until used. Plasma proteins were analyzed using a multiplex PEA (Olink Bioscience, Uppsala, Sweden). Each kit provides a microtiter plate for measuring 92 protein biomarkers in 90 samples (one panel), and in this study, 11 panels were used, including Cardiometabolic, Cell Regulation, Cardiovascular II (CVD II), Cardiovascular III (CVD III), Development, Immune Response, Immuno-Oncology, Oncology II, Inflammation, Metabolism, Neurology, and Organ Damage. Each well in a kit contains 96 pairs of DNA-labeled antibody probes. Samples were incubated in the presence of proximity antibody pairs tagged with DNA reporter molecules. When the antibody pairs bind to their corresponding antigens, the corresponding DNA tails form an amplicon by proximity extension, which can be quantified by high-throughput, real-time polymerase chain reaction (PCR). Briefly, 1 μl of each sample was mixed with 3 μl of probe solution containing a set of 92 protein target–specific antibodies conjugated with distinctive DNA oligonucleotides. The mixture was incubated overnight at 4°C, and then 96 μl of extension solution containing extension enzyme and PCR reagents was added. The generated fluorescent signal enabled the quantification of the protein using the BioMark HD System (Fluidigm Corporation). To minimize inter- and intra-run variation, the data were normalized using both an internal control (extension control) and an interplate control and were then transformed using a predetermined correction factor. The preprocessed data were provided in the arbitrary unit Normalized Protein eXpression (NPX) on a $\log_2$ scale, which were then linearized by using the formula 2NPX. Thus, a high NPX value corresponds to a high protein concentration. The limit of detection for each protein was defined as three SDs above the background. A list of the 748 analyzed proteins can be found in file S6.

SUPPLEMENTARY MATERIALS

stke.sciencemag.org/cgi/content/full/12/609/eaaz0274/DC1

Fig. S1. Tissue of origin for proteins in different annotation categories.

Fig. S2. Expression profiles for the annotation categories.

Fig. S3. Some of the data displayed for the gene *LALBA* in the in-house annotation tool.

Table S1. Description of the annotation categories.

Data file S1. The immunoglobulin genes not annotated in this study.

Data file S2. The human secretome genes and their annotated categories.

Data file S3. Human blood secretome proteins that are targets for or act as FDA-approved pharmaceutical drugs according to DrugBank.

Data file S4. Proteins detected in blood by MS and their estimated concentrations.

Data file S5. Proteins detected in blood by immunoassays and the corresponding reference articles and concentrations.

Data file S6. Proteins detected by PEAs in a longitudinal study of 86 healthy individuals.

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. M. Stastna, J. E. Van Eyk, Secreted proteins as a fundamental source for biomarker discovery. *Proteomics* **12**, 722–735 (2012).
2. H. F. Clark, A. L. Gurney, E. Abaya, K. Baker, D. Baldwin, J. Brush, J. Chen, B. Chow, C. Chui, C. Crowley, B. Currell, B. Deuel, P. Dowd, D. Eaton, J. Foster, C. Grimaldi, Q. Gu, P. E. Hass, S. Heldens, A. Huang, H. S. Kim, L. Klimowski, Y. Jin, S. Johnson, J. Lee, L. Lewis, D. Liao, M. Mark, E. Robbie, C. Sanchez, J. Schoenfeld, S. Seshagiri, L. Simmons, J. Singh, V. Smith, J. Stinson, A. Vagts, R. Vandlen, C. Watanabe, D. Wieand, K. Woods, M.-H. Xie, D. Yansura, S. Yi, G. Yu, J. Yuan, M. Zhang, Z. Zhang, A. Goddard, W. I. Wood, P. Godowski, A. Gray, The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: A bioinformatics assessment. *Genome Res.* **13**, 2265–2270 (2003).
3. M. Uhlen, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Ponten, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
4. B. B. Sun, J. C. Maranville, J. E. Peters, D. Stacey, J. R. Staley, J. Blackshaw, S. Burgess, T. Jiang, E. Paige, P. Surendran, C. Oliver-Williams, M. A. Kamat, B. P. Prins, S. K. Wilcox, E. S. Zimmerman, A. Chi, N. Bansal, S. L. Spain, A. M. Wood, N. W. Morrell, J. R. Bradley, N. Janjic, D. J. Roberts, W. H. Ouwehand, J. A. Todd, N. Soranzo, K. Suhre, D. S. Paul, C. S. Fox, R. M. Plenge, J. Danesh, H. Runz, A. S. Butterworth, Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

5. V. Emilsson, M. Ilkov, J. R. Lamb, N. Finkel, E. F. Gudmundsson, R. Pitts, H. Hoover, V. Gudmundsdottir, S. R. Horman, T. Aspelund, L. Shu, V. Trifonov, S. Sigurdsson, A. Manolescu, J. Zhu, O. Olafsson, J. Jakobsdottir, S. A. Lesley, J. To, J. Zhang, T. B. Harris, L. J. Launer, B. Zhang, G. Eiriksdottir, X. Yang, A. P. Orth, L. L. Jennings, V. Gudnason, Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).
6. A. A. Ellington, I. J. Kullo, K. R. Bailey, G. G. Klee, Antibody-based protein multiplex platforms: Technical and operational challenges. *Clin. Chem.* **56**, 186–193 (2010).
7. P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, M. Mann, Plasma proteome profiling to assess human health and disease. *Cell Syst.* **2**, 185–195 (2016).
8. N. J. Wewer Albrechtsen, P. E. Geyer, S. Doll, P. V. Treit, K. N. Bojsen-Moller, C. Martinussen, N. B. Jorgensen, S. S. Torekov, F. Meier, L. Niu, A. Santos, E. C. Keilhauer, J. J. Holst, S. Madsbad, M. Mann, Plasma proteome profiling reveals dynamics of inflammatory and lipid homeostasis markers after roux-En-Y gastric bypass surgery. *Cell Syst.* **7**, 601–612.e3 (2018).
9. D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, P. Flicek, Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
10. UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
11. H. R. Pelham, The retention signal for soluble proteins of the endoplasmic reticulum. *Trends Biochem. Sci.* **15**, 483–486 (1990).
12. J. E. Legakis, S. R. Terlecky, PTS2 protein import into mammalian peroxisomes. *Traffic* **2**, 252–260 (2001).
13. G. Schatz, The protein import system of mitochondria. *J. Biol. Chem.* **271**, 31763–31766 (1996).
14. P. Bornstein, Matricellular proteins: An overview. *J. Cell Commun. Signal.* **3**, 163–165 (2009).
15. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
16. F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014).
17. E. Assarsson, M. Lundberg, G. Holmquist, J. Björkesten, S. B. Thorsen, D. Ekman, A. Eriksson, E. Rennel Dickens, S. Ohlsson, G. Edfeldt, A.-C. Andersson, P. Lindstedt, J. Stenvang, M. Gullberg, S. Fredriksson, Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLOS ONE* **9**, e95192 (2014).
18. T. Farrah, E. W. Deutsch, G. S. Omenn, D. S. Campbell, Z. Sun, J. A. Bletz, P. Mallick, J. E. Katz, J. Malmstrom, R. Ossola, J. D. Watts, B. Lin, H. Zhang, R. L. Moritz, R. Aebersold, A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell. Proteomics* **10**, M110.006353 (2011).
19. J. M. Schwenk, G. S. Omenn, Z. Sun, D. S. Campbell, M. S. Baker, C. M. Overall, R. Aebersold, R. L. Moritz, E. W. Deutsch, The human plasma proteome draft of 2017: Building on the human plasma PeptideAtlas from mass spectrometry and complementary assays. *J. Proteome Res.* **16**, 4299–4310 (2017).
20. R. S. Tirumalai, K. C. Chan, D. A. Prieto, H. J. Issaq, T. P. Conrads, T. D. Veenstra, Characterization of the low molecular weight human serum proteome. *Mol. Cell. Proteomics* **2**, 1096–1103 (2003).
21. M. Neiman, C. Hellström, D. Just, C. Mattsson, L. Fagerberg, I. Schuppe-Koistinen, A. Gummesson, G. Bergstrom, O. Kallioniemi, A. Achour, R. Sallinen, M. Uhlén, P. Nilsson, Individual and stable autoantibody repertoires in healthy individuals. *Autoimmunity* **52**, 1–11 (2019).
22. G. Bergström, G. Berglund, A. Blomberg, J. Brandberg, G. Engström, J. Engvall, M. Eriksson, U. de Faire, A. Flinck, M. G. Hansson, B. Hedblad, O. Hjelmgren, C. Janson, T. Jernberg, A. Johnsson, L. Johansson, L. Lind, C. G. Löfdahl, O. Melander, C. J. Östgren, A. Persson, M. Persson, A. Sandstrom, C. Schmidt, S. Söderberg, J. Sundström, K. Toren, A. Waldenström, H. Wedel, J. Vikgren, B. Fagerberg, A. Rosengren, The Swedish CArdioPulmonary BioImage study: Objectives and design. *J. Intern. Med.* **278**, 645–659 (2015).

Acknowledgments: We acknowledge the entire staff of the HPA program and the Science for Life Laboratory for their valuable contributions. **Funding:** Main funding was provided from the Knut and Alice Wallenberg Foundation (WCPR), the Erling Persson Foundation (KCAP), and the Novo Nordisk Foundation (CFB). **Author contributions:** M.U. conceived and designed the analysis. M.U., Å.S., M.J.K., A.H., B.F., F.E., F.P., J.N., L.S., J. Rockberg, J. Robinson, B.G.V., and D.K. collected and contributed data to the study. M.U., Å.S., W.Z., A.T., M.J.K., B.F., and L.F. performed the data analysis. M.U., Å.S., H.T., L.F., J. Rockberg, M.D., M.M., E.S., J.M., A.M., J.M.S., P.N., C.S., C.L., A.-L.V., M.L., A.B., A.-S.S., S.K., S.E., J.S., B.K., J.V., A.M., T.D.-C., E.P., R.S.H., M.-G.H., E. Lindström, B.A., D.M., D.S., P.T., F.D., and E. Lundberg performed the secretome annotation. K.v.F., P.O., and M.Z. provided the infrastructure for the data. G.B. and A.G. supplied clinical material. M.U. and Å.S. drafted the manuscript. M.U. and Å.S. revised the manuscript. All authors discussed the results and contributed to the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or Supplementary Materials and are available without restriction at the HPA portal (www.proteinatlas.org).

Submitted 6 August 2019
Accepted 12 November 2019
Published 26 November 2019
10.1126/scisignal.aaz0274

Citation: M. Uhlén, M. J. Karlsson, A. Hober, A.-S. Svensson, J. Scheffel, D. Kotol, W. Zhong, A. Tebani, L. Strandberg, F. Edfors, E. Sjöstedt, J. Mulder, A. Mardinoglu, A. Berling, S. Ekblad, M. Dannemeyer, S. Kanje, J. Rockberg, M. Lundqvist, M. Malm, A.-L. Volk, P. Nilsson, A. Månberg, T. Dodig-Crnkovic, E. Pin, M. Zwahlen, P. Oksvold, K. von Feilitzen, R. S. Häussler, M.-G. Hong, C. Lindskog, F. Ponten, B. Katona, J. Vuu, E. Lindström, J. Nielsen, J. Robinson, B. Ayoglu, D. Mahdessian, D. Sullivan, P. Thul, F. Danielsson, C. Stadler, E. Lundberg, G. Bergström, A. Gummesson, B. G. Voldborg, H. Tegel, S. Hober, B. Forsström, J. M. Schwenk, L. Fagerberg, Å. Sivertsson, The human secretome. *Sci. Signal.* **12**, eaaz0274 (2019).

RESEARCH ARTICLE SUMMARY

TRANSCRIPTOMICS

A genome-wide transcriptomic analysis of protein-coding genes in human blood cells

Mathias Uhlen*, Max J. Karlsson, Wen Zhong, Abdellah Tebani, Christian Pou, Jaromir Mikes, Tadepally Lakshmikanth, Björn Forsström, Fredrik Edfors, Jacob Odeberg, Adil Mardinoglu, Cheng Zhang, Kalle von Feilitzen, Jan Mulder, Evelina Sjöstedt, Andreas Hober, Per Oksvold, Martin Zwahlen, Fredrik Ponten, Cecilia Lindskog, Åsa Sivertsson, Linn Fagerberg†, Petter Brodin†

INTRODUCTION: Blood is the predominant source for molecular analyses in humans, both in clinical and research settings, and is the target for many therapeutic strategies, emphasizing the need for comprehensive molecular maps of the cells constituting human blood. The Human Protein Atlas program (www.proteinatlas.org) is an open-access database that aims to map all human proteins by integrating various omics technologies, including antibody-based imaging. Previously, the Human Protein Atlas included gene expression information from peripheral blood mononuclear cells but not the many subpopulations of blood cells within this cell type. To increase the resolution, we performed an in-depth characterization of the constituent cells in blood to provide a detailed view of the gene expression in individual human blood cells and relate these to the other tissues in the body.

RATIONALE: A quantitative transcriptomics-based expression analysis was performed in 18 canonical immune cell populations (Fig. 1) isolated by flow cytometric sorting. The blood cell expression profiles are presented in combination with expression profiles of tissues, including transcriptomics data from external sources to expand the number of tissue types as well as brain regions included in the database. A genome-wide classification of the protein-coding genes has been performed in terms of expression specificity and distribution, both in blood cells and tissues.

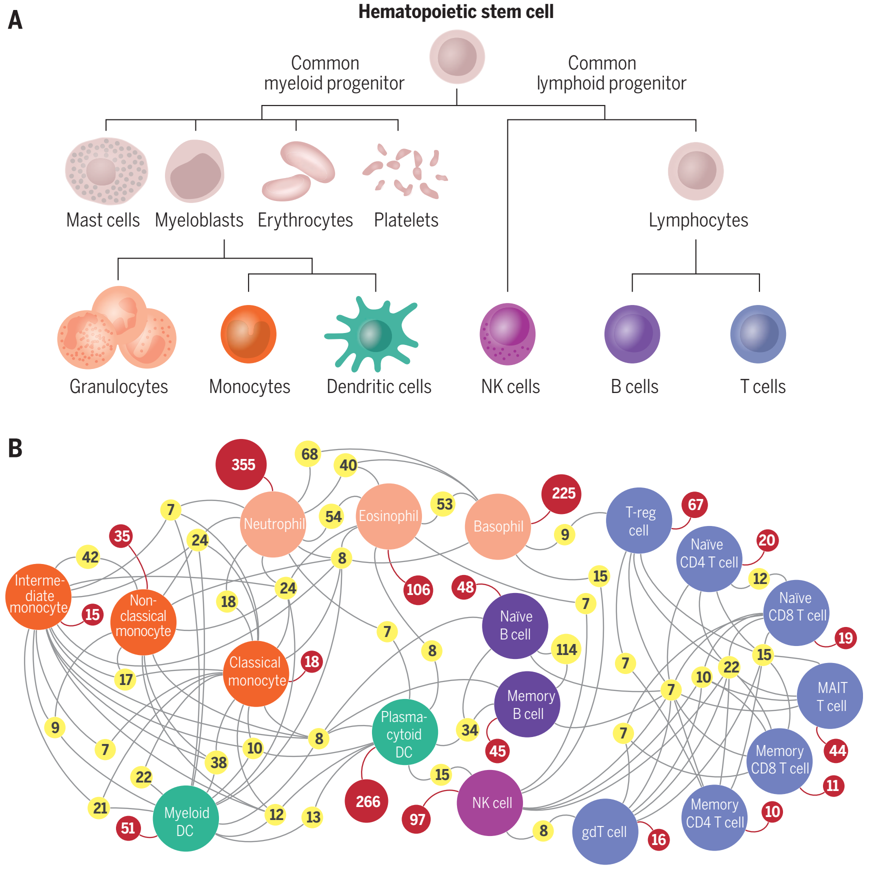


Fig. 1. Outline of the analysis of human single blood cell types. (**A**) A schematic view of the hematopoietic differentiation. This study analyzes the cell types shown in the bottom row. NK, natural killer. (**B**) Network plot showing the number of cell type– (red) and group-enriched (yellow) genes in the 18 cell types. The network is limited to nodes with a minimum of seven genes. DC, dendritic cell; T-reg, regulatory T cell; gdT cell, gamma delta T cell; MAIT, mucosal associated invariant.

RESULTS: We present an atlas of the expression of all protein-coding genes in human blood cells, integrated with a classification of the specificity and distribution of all protein-coding genes in all major tissues and organs in the human body. A genome-wide analysis of blood cell RNA expression profiles allowed the identification of genes with elevated expression in various immune cells, confirming well-known protein markers, but also identified novel targets for in-depth analysis. There are 1448 protein-coding genes that have enriched expression in a single immune cell type. It will be interesting to study the corresponding proteins further to explore the biological functions linked to the respective cell phenotypes. A network plot of all cell type–enriched and group-enriched genes (Fig. 1B) reveals that many of the cell type–enriched genes are in neutrophils, eosinophils, and plasmacytoid dendritic cells, while many of the elevated genes in T and B cells are group-enriched across subpopulations of these lymphocytes. To illustrate the usefulness of this resource, we show the cellular distribution of genes known to cause primary immunodeficiencies in humans and find that many of these genes are expressed in cells not currently implicated in these diseases, illustrating how this global atlas can help us better understand the function of specific genes across cells and tissues in humans.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aax9198

CONCLUSION: In this study, we have performed a genome-wide transcriptomic analysis of protein-coding genes in sorted blood immune cell populations to characterize the expression levels of each individual gene across all cell types. All data are presented in an interactive, open-access Blood Atlas as part of the Human Protein Atlas and are integrated with expression profiles across all major tissues to provide spatial classification of all protein-coding genes. This allows for a genome-wide exploration of the expression profiles across human immune cell populations and all major human tissues and organs. ■

The list of author affiliations is available in the full article online.
***Corresponding author. Email: mathias.uhlen@scilifelab.se**
†These authors contributed equally to this work.

Cite this article as M. Uhlen *et al*., *Science* 366, eaax9198 (2019). DOI: 10.1126/science.aax9198

Originally published 20 December 2019

RESEARCH ARTICLE

TRANSCRIPTOMICS

A genome-wide transcriptomic analysis of protein-coding genes in human blood cells

Mathias Uhlen[1,2,3]*, Max J. Karlsson[1], Wen Zhong[1], Abdellah Tebani[1], Christian Pou[4], Jaromir Mikes[4], Tadepally Lakshmikanth[4], Björn Forsström[1], Fredrik Edfors[1], Jacob Odeberg[1,5], Adil Mardinoglu[1,6], Cheng Zhang[1], Kalle von Feilitzen[1], Jan Mulder[2], Evelina Sjöstedt[2], Andreas Hober[1], Per Oksvold[1], Martin Zwahlen[1], Fredrik Ponten[7], Cecilia Lindskog[7], Åsa Sivertsson[1], Linn Fagerberg[1]†, Petter Brodin[4,8]†

Blood is the predominant source for molecular analyses in humans, both in clinical and research settings. It is the target for many therapeutic strategies, emphasizing the need for comprehensive molecular maps of the cells constituting human blood. In this study, we performed a genome-wide transcriptomic analysis of protein-coding genes in sorted blood immune cell populations to characterize the expression levels of each individual gene across the blood cell types. All data are presented in an interactive, open-access Blood Atlas as part of the Human Protein Atlas and are integrated with expression profiles across all major tissues to provide spatial classification of all protein-coding genes. This allows for a genome-wide exploration of the expression profiles across human immune cell populations and all major human tissues and organs.

Resolving the molecular details of proteome variation in the different cells, tissues, and organs of the human body may considerably increase our knowledge of human biology and disease. Several efforts to map the molecular components of the human body in a comprehensive manner have been initiated, including efforts to generate experimental data such as the Human Cell Atlas (*1*), the Human Biomolecular Atlas Program (HuBMAP) (*2*), the Biohub (*3*), the Genotype-Tissue Expression (GTEx) project (*4*), the Functional Annotation of the Mammalian Genome (FANTOM) project (*5*), and the Allen Brain Atlas (*6*), involving many alternative technologies, including single-cell genomics (*7*), in situ analysis (*8*), transcriptomics (*9*), proteomics (*10*), and antibody-based profiling (*11*). In addition, several knowledge resources have been created to annotate, assemble, and integrate data from such sources, such as UniProt (*12*), ELIXIR (*13*), ArrayExpress (*14*), Peptide Atlas (*15*), and ImmPort (*16*). The combined efforts of these resources have the potential to allow a systematic knowledge base of the molecular components of human life that will aid a systems biology understanding of human biology and diseases.

A complement to these efforts is the Human Protein Atlas program (*17*), which is exploring the human proteome using gene-centric and genome-wide antibody-based profiling on tissue microarrays. This allows for spatial pathology-based annotation of protein expression that is performed in combination with deep sequencing transcriptomics profiling of the same tissue types. The aim is to map all human proteins in cells, tissues, and organs using integration of various omics technologies, including antibody-based imaging, mass spectrometry–based proteomics, and transcriptomics. The earlier version of the Human Protein Atlas consists of three separate parts, each focusing on a particular aspect of the genome-wide analysis of human proteins: the Tissue Atlas (*17*), showing the distribution of proteins across all major tissues and organs in the human body; the Cell Atlas (*18*), showing the subcellular localization of proteins in single cells; and the Pathology Atlas (*19*), showing the impact of different protein levels in tumor tissue on the survival of cancer patients. However,

[1] Science for Life Laboratory, KTH–Royal Institute of Technology, Stockholm, Sweden. [2] Department of Neuroscience, Karolinska Institute, Stockholm, Sweden. [3] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark. [4] Science for Life Laboratory, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden. [5] Coagulation Unit, Department of Hematology, Karolinska University Hospital, Stockholm, Sweden. [6] Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral and Craniofacial Sciences, King's College London, London, UK. [7] Department of Immunology, Genetics and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden. [8] Unit of Pediatric Rheumatology, Karolinska University Hospital, Stockholm, Sweden.
*Corresponding author. Email: mathias.uhlen@scilifelab.se

†These authors contributed equally to this work.

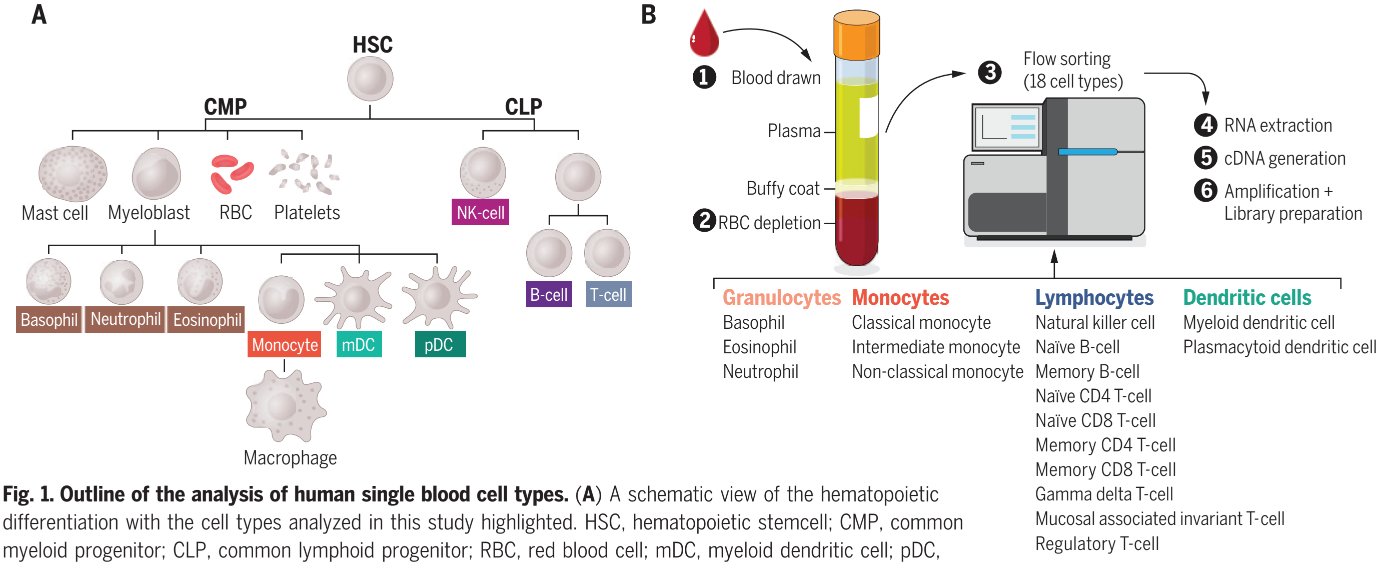


Fig. 1. Outline of the analysis of human single blood cell types. (**A**) A schematic view of the hematopoietic differentiation with the cell types analyzed in this study highlighted. HSC, hematopoietic stemcell; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; RBC, red blood cell; mDC, myeloid dendritic cell; pDC, plasmacytoid dendritic cell. (**B**) A schematic view of the experimental procedure to analyze the transcript expression levels in human single cell types. The 18 cell types listed include seven subsets of T cells, two variants of B cells, three different monocytic cell types, and the three known forms of granulocytes.

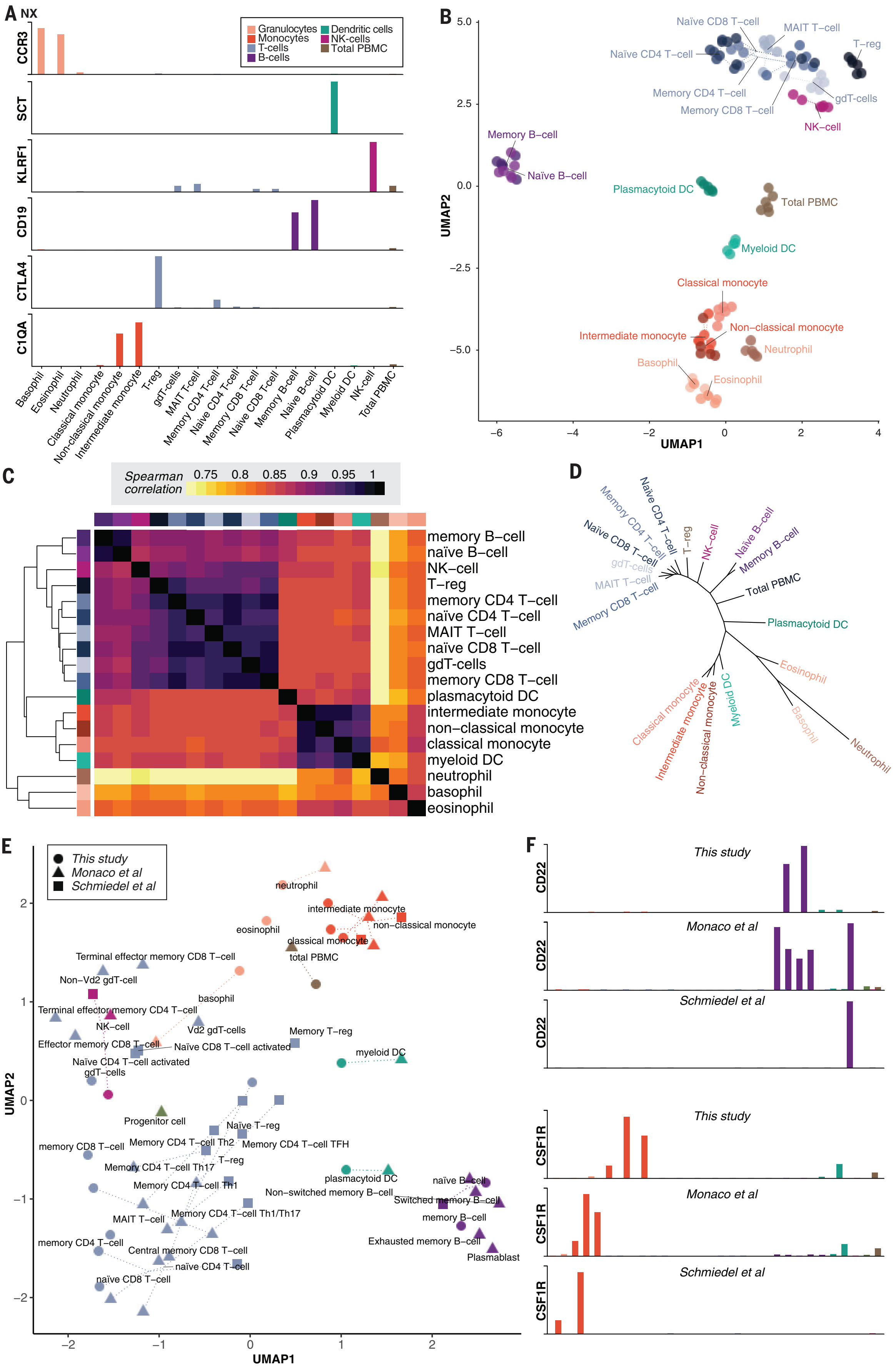


Fig. 2. The expression profiles of the protein-coding genes in human single blood cell types. (**A**) Examples of expression profiles for six genes enriched in one of the cell lineages (see www.proteinatlas.org for details). (**B**) A UMAP analysis of the relationship between the global expression patterns in all the 109 blood cell samples analyzed here. (**C**) A heatmap showing the pairwise Spearman correlation between the global expression profiles for the 18 analyzed cell types. (**D**) Transcriptomics-derived hematopoietic tree showing the similarities in global expression patterns between different human blood cell types. (**E**) UMAP analysis showing the relationship between all the blood cell samples from three different sources. Cell types overlapping between two or all three datasets are connected by dotted lines. (**F**) Comparison of expression profiles for the three datasets, as exemplified for the genes *CD22* and *CSF1R* (see www.proteinatlas.org for details).

there is a lack of data regarding protein expression levels in human blood cells. Given that blood is the most commonly used material for molecular analyses in clinical labs and in research, characterizing the constituents of blood and updating the Human Protein Atlas with a more fine-grained view of the immune cells in blood will be of importance.

In this study, we performed a quantitative expression analysis of 18 canonical immune cell populations, as well as total peripheral blood mononuclear cells (PBMCs) from human blood separated by flow cytometric sorting. The data are integrated with recent transcriptomics efforts involving flow sorting of blood cells, including the analysis in 15 blood cell types by Schmiedel *et al.* (*20*) and 29 blood cell types as well as total PBMCs by Monaco *et al.* (*21*). We presented the expression profiles in specific cell populations and combined the new single-cell blood data with the data from the Tissue Atlas (*17*) by incorporating transcriptomics data from the GTEx (*4*) and the FANTOM5 (*5*) projects. Moreover, we expanded the set of normal tissue samples by adding tissues such as retina and tongue, as well as extensive data covering the different regions of the brain. A genome-wide classification of the protein-coding genes with regard to tissue and cell distribution as well as specificity has been performed using between-sample normalized data (*22*, *23*). The results are presented in an interactive database (www.proteinatlas.org) that can serve as a reference for researchers interested in spatial expression profiles of human blood cells in relation to the body-wide profiles in all major tissues and organs.

Transcriptome analysis of isolated human immune cell populations

We used flow cytometric sorting to allow whole-genome transcriptome analysis of the major blood cell types from human blood (Fig. 1A). Whole blood was collected from six healthy individuals, and 18 immune cell types were separated by flow cytometric sorting, as outlined in Fig. 1B. The cell types recovered included naïve and memory B cells, CD4 and CD8 T cell populations, natural killer (NK) cells, three monocyte subsets, neutrophils, eosinophils, and basophils, as well as plasmacytoid and myeloid dendritic cells. These can be classified into six different blood cell lineages consisting of granulocytes, monocytes, T cells, B cells, dendritic cells, and NK cells. The sorted cells were immediately processed using RNA extraction and cDNA generation followed by deep mRNA sequencing. The RNA expression levels were determined for all protein-coding genes (n = 19,670) across the 18 immune cell populations and visualized in a newly created Blood Atlas, launched here as an extended edition of the open-access Human Protein Atlas (www.proteinatlas.org/blood).

In the Blood Atlas, the expression levels for each of the 19,670 genes are displayed for the 18 cell types and PBMC as exemplified in Fig. 2A. The first example is the G-coupled C-C motif chemokine receptor 3 (CCR3), involved in allergic reactions, showing distinct expression in basophil and eosinophils, with much lower levels in neutrophils. Next, the secretin propeptide (SCT), previously described (*24*) as being produced in the gastrointestinal tract (duodenum and colon), is here found to also be expressed in the human plasmacytoid dendritic cells. The killer cell lectin like receptor F1 (KLRF1), known to stimulate cytotoxicity and cytokine release in NK cells (*25*), is an example of an NK cell enriched gene, but the data also show expression in gamma delta T (gdT) cells and mucosal-associated T invariant (MAIT) cells. The purity of our sorting is verified by known marker expression patterns, such as the canonical cell surface receptor CD19, exclusively expressed in B cells, and the cytotoxic T lymphocyte–associated protein 4 (CTLA4), expressed on regulatory T cells ($\mathrm{T_{regs}}$). The complement C1q A chain (C1QA) of the complement system is instead enriched in monocytes, and the profiling shows high expression in intermediate and nonclassical monocytes but no expression in classical monocytes. In addition to the sorted single cell type populations, the mixed PBMCs were collected from the individuals, as described before (*26*), and the transcriptome determined.

Global expression profiles for the blood cell types

The relationships between all blood cell samples on the basis of their global expression profiles were analyzed using different algorithms, including principal components analysis (PCA) (*27*) and uniform manifold approximation and projection (UMAP) (*28*), and the UMAP results for all samples for all cell types are shown in Fig. 2B. The samples from the different cell types showed similar global expression profiles with the multitude of different B cell and T cell types clustering together. A heatmap based on pairwise Spearman correlation of the expression profiles of the 18 cell types (Fig. 2C) showed that cells of similar origin have similar overall expression profiles, with the three granulocyte cell types having the most distinct expression profiles. All lymphocytes form a separate cluster, including all seven T cells clustering together with the NK cells, and naïve and mature B cells clustering together. The monocytes are most closely related to the myeloid dendritic cells and the plasmacytoid dendritic cells. To analyze the similarities between the cell types of different origins in more detail, we constructed a transcriptomics-derived hematopoietic tree (Fig. 2D) to further illustrate the relation in global expression profiles between the different single blood cell types.

The transcript expression profiles from the recent studies by Schmiedel *et al.* (*20*) and Monaco *et al.* (*21*), having partially overlapping data for 13 and 27 blood cell types, respectively, are also included in the Blood Atlas. UMAP results for all cell types from the three different data sources are shown in Fig. 2E, confirming the distinct expression profiles between various types of blood cells. A summary of the genome-wide expression levels from all three datasets is visualized for all protein-coding genes in the Blood Atlas resource online (Fig. 2F). More in-depth analyses are needed to establish whether the differences seen are due to differential activation states based on sample handling, differences in sample handling and cell sorting, or whether they reflect biological differences among cohorts, representing individuals from Europe (this study), the United States (*20*), and Asia (*21*).

Genome-wide transcriptomics profiles across all major organs and tissues

With the new data covering the blood cell expression profiles as well as an expanded set of normal tissue types, the body-wide tissue profiling performed earlier (*29*) was revised. Because the brain regions were only superficially covered in the earlier analysis, we also decided to include more brain regions using publicly available data from the GTEx (*4*) and FANTOM (*5*) consortia to allow for more in-depth coverage of the different regions of the human brain. Altogether, 1710 samples from selected human brain regions were added to the classification covering 23 human subregions and summarized into 12 main structures of the brain (Fig. 3A). The detailed analysis of the protein expression in these brain structures will be described elsewhere, but here the expression profiles were used in the body-wide tissue classification of all genes. In addition, the five tissues dominated by immune cells (thymus, appendix, spleen, lymph node, and tonsil) were summarized into "lymphoid tissues," and the four highly related tissues from the gut (duodenum, small intestine, colon, and rectum) were summarized into "intestine," as outlined in Fig. 3A. Some additional tissues, including lactating breast, vagina, retina, ductus deferens, and tongue, were also added to the comparative analysis. The expression data for the 18 blood cell types as well as PBMC described above were summarized into "blood." A body-wide classification based on the genome-wide expression profiles of the protein-coding genes was performed with 171 different cells, tissues, and organs, which are summarized into 37 tissue types.

The transcriptomics data was normalized by applying two different strategies with the main objective to allow (i) within-sample comparisons and (ii) between-sample comparisons, respectively, as outlined in fig. S1. For the within-sample comparisons, the fraction of transcripts corresponding to a particular gene

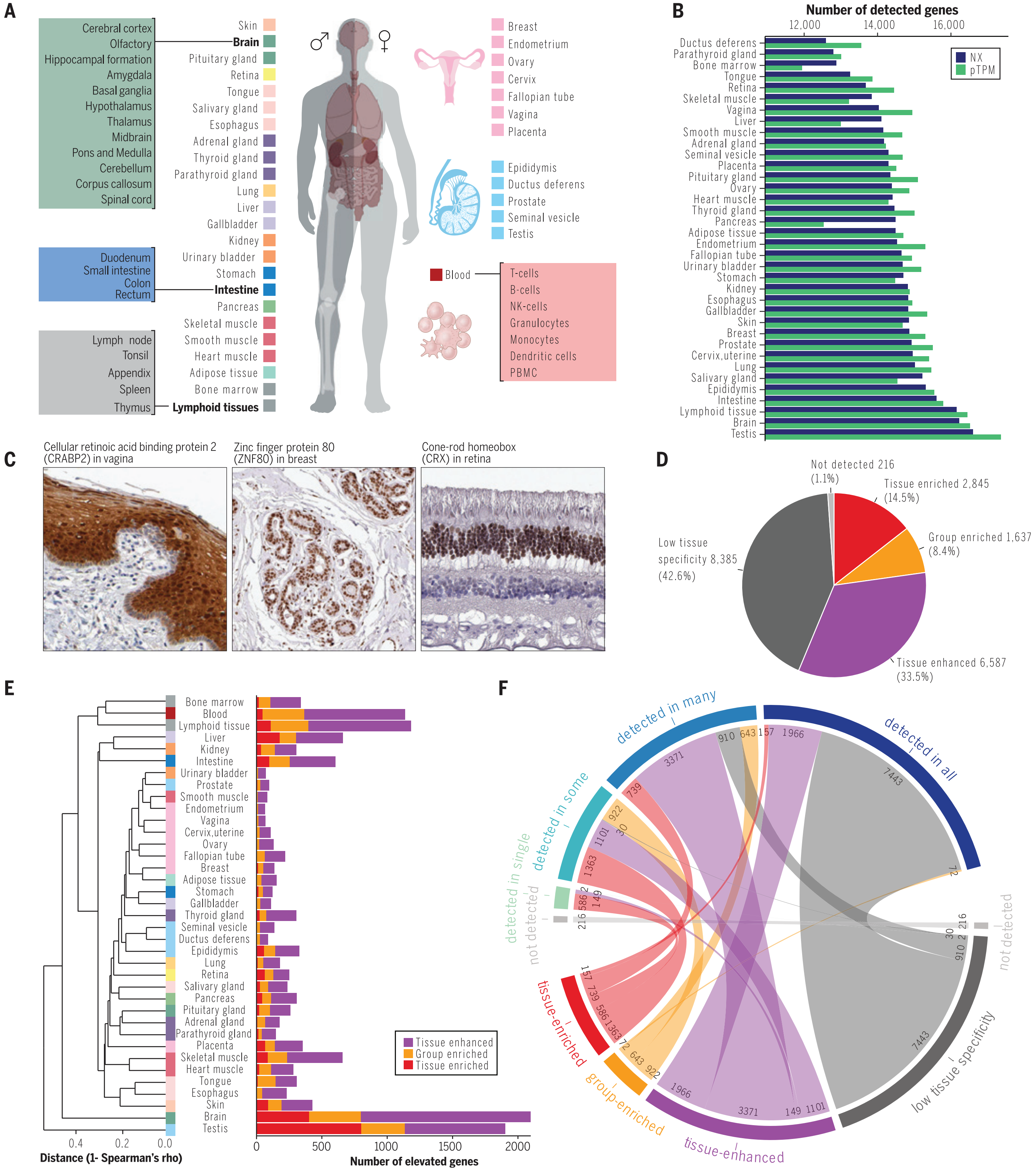


Fig. 3. Classification of the human global gene expression profiles across all major tissues and organs and the immune cell types. (**A**) Schematic view of all human tissues and organs analyzed. (**B**) The number of detected genes in selected tissues based on pTPM and NX values, respectively. (**C**) Three examples of tissues introduced in this study. (**D**) Pie chart showing the number of genes classified according to the specificity categories. (**E**) (Left) A dendrogram based on the correlation of global expression profiles across all tissues and organs, including blood. (Right) Barplot displaying the number of elevated genes for each tissue type. (**F**) Chord diagram showing the relationship between the distribution classification and the specificity classification. Each link represents the number of genes with the linked distribution category and specificity category.

is used. We focus on the protein-coding transcripts and the fraction of transcripts per million of total transcripts from protein-coding genes (pTPM) calculated for each individual gene in every sample. The pTPM value is visualized on the Blood Atlas page of the Human Protein Atlas across the samples for each of the genes. The pTPM values can be considered as the within-sample normalized data from the deep sequencing, in which noncoding RNA has been excluded from the analysis. The pTPM values can be used to investigate the abundance of a particular gene, gene family, or gene class relative to all other transcripts in a particular cell, tissue, or organ.

The second normalization strategy is carried out to allow for comparisons across samples and to avoid batch effects caused by sampling, technology platforms, or the difference in transcriptome size between different types of tissues, as exemplified by pancreas and salivary gland, where a small number of genes are very highly expressed (22, 23). This is particularly important when tissue samples based on different transcriptomic technology platforms have been used, as described for the tissue analysis where RNA sequencing data from multiple sources as well as cap analysis of gene expression data from the FANTOM5 program have been combined. Here, we used a normalization based on trimmed mean of M values (TMM) (30), Pareto scaling (31), and the Limma R package (32) to calculate a normalized expression value (NX) for each gene in every sample. In the Human Protein Atlas, the NX value for each gene is visualized in parallel with the pTPM value for all tissues and cell types. The objective of using the NX value is to facilitate the analysis of differences in expression of genes between cells, tissues, and organs and to allow for a specificity classification based on the genome-wide expression of all genes across the human blood cells, tissues, and organs.

The number of detected genes in the different tissues and organs was investigated using both the within-sample normalization (pTPM) and the between-sample normalization (NX), in both cases using a cutoff value of 1, as described previously (17). In Fig. 3B and fig. S2, the results for selected tissues are shown, and the analysis demonstrated a similar number of detected genes for most samples, with some notable exceptions, including tissues with a small fraction of highly abundant transcripts, such as bone marrow (hemoglobin), pancreas (digestive enzymes), liver (albumin), and salivary gland (digestive proteins).

The revised tissue classification of all human genes

The extended data allowed us to refine the classification for the putative protein-coding genes on the basis of their expression across all 37 cells, tissues, and organs. Some examples of genes detected in the recently added tissues are shown in Fig. 3C. The first example, CRABP2 in vagina, plays a role in the vitamin A signaling pathway, with tissue-enhanced expression in squamous mucosa and with nuclear and cytoplasmic positivity in suprabasal squamous epithelia. Another example is breast with ZNF80, a protein with unknown function that here shows nuclear positivity with tissue enhanced expression in blood and breast tissue. Also shown is retinal epithelium with cone-rod homeobox protein (CRX), showing nuclear positivity in the cone-and-rod photoreceptor layer.

All 19,670 genes were classified according to a strategy based on scoring both tissue specificity and tissue distribution (tables S1 and S2; full list of results in data S1). Of all protein-coding genes, 56% (n = 11,069) showed elevated expression in at least one of the analyzed tissues, and these were further subdivided into (i) tissue-enriched genes with at least fourfold higher expression levels (based on NX values) in one tissue type as compared with any other analyzed tissue; (ii) group-enriched genes with enriched expression in a small number of tissues (2 to 5); and (iii) tissue-enhanced genes with only moderately elevated expression (table S1). 2845 genes (14%) of the protein-coding genes were found to be enriched in one of the analyzed tissues (Fig. 3D), and only 216 genes were not detected in any of the analyzed tissues. Our classification shows the number of tissue-enriched genes for each tissue type, as well as the number of genes enriched in different groups of tissues (Fig. 3E). The largest number of tissue-enriched genes are found in the testes, as shown in our previous results (17); however, the largest number of elevated genes is now found in the brain, most likely owing to the inclusion of many more brain regions as compared with earlier versions of the atlas. Whereas the specificity classification showed us the enrichment of genes, the distribution classification showed us the fraction of tissues where the gene is expressed. Only 737 genes (4%) are restricted to a single tissue, while almost half of the protein-coding genes are expressed in all tissues (n = 9638) (fig. S3).

The global expression profiles were investigated using the between-sample normalized values (NX) using PCA (fig. S4), UMAP (fig. S5), and hierarchical clustering based on genome-wide correlation between the cells, organs, and tissue types (fig. S6). The resulting dendrogram (Fig. 3E) shows that testis and brain have the most distinct expression profiles compared with all other tissues, and that blood is most highly correlated with lymphoid tissues and bone marrow. The overall results corresponded well with the origin and function of each tissue, as exemplified by many of the female tissues clustering together and the close connectivity of the two tissues composed of striated muscle (cardiac and skeletal muscle).

In Fig. 3F and table S3, a summary of all 19,670 genes with regard to both tissue specificity and distribution classification is shown with the genome-wide relationship of the two classification schemes introduced, showing that only 586 genes are "tissue specific," meaning they are tissue-enriched and, at the same time, only detected in a single tissue (this list is available at www.proteinatlas.org). Relatively few genes (n = 1637) were found to be group-enriched, and this lower number compared with earlier results (17) is most likely explained by the fact that some tissues have now been grouped together, such as lymphoid tissues, intestine, and brain. 43% (n = 8385) of the genes were classified as "low tissue specificity," and most of these are found in the "detected in all" category. All 19,670 protein-coding genes in humans have now been analyzed with respect to their tissue specificity and distribution across all major organs, tissues, and blood cells in the human body, and the results are available in the Human Protein Atlas.

Transcriptome usage in different cells and tissues

An analysis of the transcriptome allowed us to determine the fraction of transcripts corresponding to different genes in each analyzed cell type and tissue. Here, we report the transcriptome usage for some representative blood cell types and tissues on the basis of within-sample normalized pTPM values (Fig. 4A and fig. S7A) and between-sample NX normalized values (Fig. 4B and fig. S7B). These are further stratified according to genes coding for secreted, membrane-bound, and intracellular proteins. It is notable that, for pancreas and salivary gland, as much as 80 and 50%, respectively, of the transcripts (based on pTPM) encode for secreted proteins. This demonstrates the extreme specialization of these "secretory cell factories" for production of extracellular proteins, with a few genes dominating the transcriptome load. The most abundant proteins in pancreas code for digestive enzymes, such as lipases (PNLIP, CLPS), proteases (PRSS1, CELA3A), and peptidases (CPA1, CPB1). The most abundant proteins in salivary gland are a protein with essentially unknown function (submaxillary gland androgen regulatory protein 3B, SMR3B) and statherin (STATH), which prevents the precipitation of calcium phosphate in saliva, maintaining a high calcium level in saliva that is necessary for remineralization of tooth enamel. The second- and fourth-most abundant proteins in salivary gland are antimicrobial peptides (HTN3 and HTN1). Similarly, the liver has a large fraction of secreted proteins with the most abundant being albumin (ALB), haptoglobin (HP), and apolipoprotein A2 (APOA2).

In contrast, >60% of all pTPM values for cardiac muscle code for membrane proteins, mainly consisting of mitochondrial proteins,

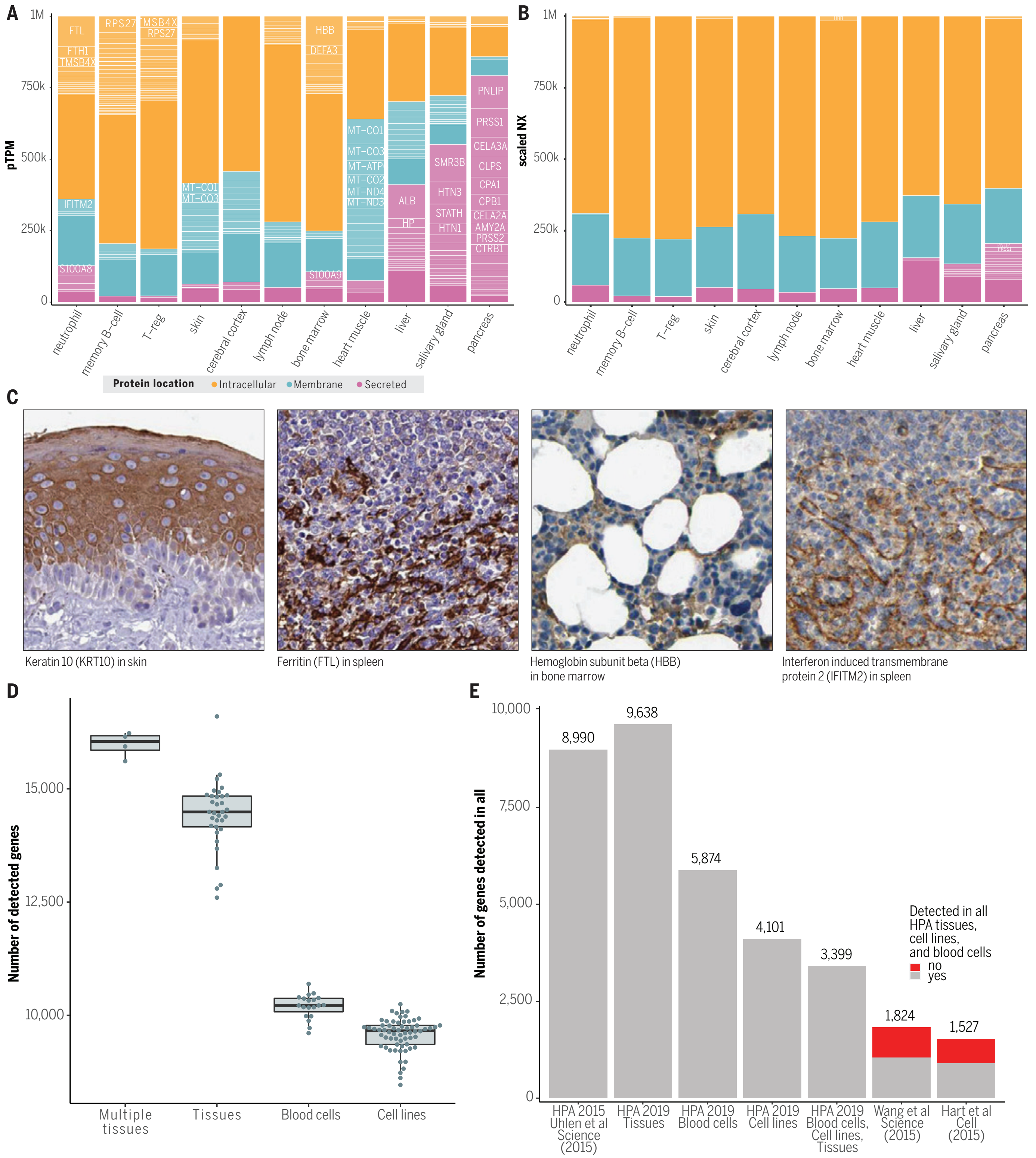


Fig. 4. Analysis of the global expression profiles in the various tissues. (**A**) The transcriptional load based on pTPM in some selected cells and tissues stratified according to protein location: secreted, membrane-bound, or intracellular. The genes with most abundant transcripts are labeled. (**B**) Same as (A), but based on the between-sample normalized NX values scaled to a sum of one million. (**C**) Immunohistochemistry (IHC) images from the Human Protein Atlas for four examples of the most abundant genes in some selected tissues. (**D**) Boxplot showing the distribution of the number of detected genes for the combined groups of tissue types (brain, blood, intestine, and lymphoid tissues), all single tissue types, the 18 blood cell types, and cell lines (*18*). (**E**) The number of genes expressed in all samples is shown based on the earlier analysis (*17*), and in all tissues, the immune cell types reported here as well as for 60 cell lines. Also shown is the number of genes when including all these three sample types. We also compare the number of genes identified as "essential" using CRISPR knock-out strategies (*33*, *34*) and highlight the number of genes not "detected in all" for all samples covering the cell lines, tissues, and blood cells.

which is not unexpected given the extreme requirement of energy in the cardiac muscle. For most tissues and for all the single blood cells, the intracellular proteins instead constitute most of the transcriptome load, as exemplified by bone marrow with hemoglobin (HBB) and the skin with keratin (KRT10) (Fig. 4C) as the most abundant transcript, respectively. In the blood cells, there are fewer genes with a dominant abundance, although the most abundant transcript in neutrophils is the gene encoding the intracellular protein ferritin light chain (FTL), a subunit of ferritin, the major protein responsible for intracellular iron storage. A notable example of a gene with abundant transcripts, but with almost no known functional information, is the interferon-induced transmembrane protein 2 (IFITM2), which is highly expressed in neutrophils and here is shown in spleen. The transcriptome maps demonstrate the high specialization of each tissue with a large portion of the transcript burden devoted to functions of relevance for the corresponding cells in respective tissue type.

Number of detected genes and the "housekeeping" genes

An analysis of the number of detected genes in the various samples (Fig. 4D) shows that ~16,000 genes are detected in the four combined groups of multiple tissue types (blood, brain, intestine, and lymphoid tissues), while the analysis of single tissues shows a slightly smaller number of genes (~14,000 on average) —with the exception of testis, in which 16,598 genes are detected. This is in contrast to the much smaller number of detected genes when analyzing cell lines (~9500 genes per cell line) and single blood cell types (~10,000 genes). The fact that more genes are detected in tissues as compared with the single cell type analysis is not unexpected, as it reflects the presence of a multitude of different cell types present in composite tissues. The observation that a slightly smaller number of genes are detected in the cell lines as compared with the single blood cells is interesting, and it is tempting to speculate that this is due to the in vitro specialization of the cell lines.

Almost half (49%) of the protein-coding genes (n = 9638) were detected in all analyzed tissues (Fig. 4E), and these genes include known "housekeeping" genes encoding mitochondrial proteins, and proteins involved in overall cell structure, translation, transcription, and replication. An analysis of the human cell lines shows that 4101 genes are detected in all samples. Similarly, the analysis of the 18 single blood cell types shows that 5874 genes are ubiquitously detected across all immune cells. If the tissues, cell lines, and single blood cell types are combined, the number of protein-coding genes detected in all samples is decreased to 3399 (Fig. 4E). This is still a much larger number when compared with the determination of essential genes using genome-wide CRISPR-Cas9 knock outs ([33], [34]), which identified 1824 and 1527 genes with unconditional importance for cell survival, respectively. This suggests that many genes are present in all cells but that they perform redundant functions in cell lines. Altogether, we identified genes that are both essential in genome-wide knock-out screens and here detected in all blood cells, cell lines, tissues, and organs. This list of genes (available at www.proteinatlas.org) contains many well-known housekeeping genes involved in replication, translation, and cellular processes, and more in-depth studies are needed to explore the function of the genes detected in all tissues and yet not identified as essential by the knock-out screen.

It is reassuring that the number of "missing genes," i.e., those not detected in any tissue or cell type, is now reduced to 216, which is only ~1% of the total number of predicted protein-coding genes. We therefore revised ([35]) the number of genes for which evidence at protein level is present by combining our antibody-based data with the manual annotation of literature by the UniProt consortium ([36]) and the results from mass spectrometry–based proteogenomics analyses ([37]). The analysis showed that there are 17,660 protein-coding genes with proteins identified from at least one of the three efforts and 15,155 genes with experimental evidence from at least two of the efforts (fig. S8; see www.proteinatlas.org/humanproteome/proteinevidence for details). Furthermore, there are 1794 additional genes with evidence only at the RNA level, and these genes are obvious targets for more comprehensive functional protein studies. It is notable that chromosome 11 has many more missing genes than the other chromosomes, likely owing to its high number of olfactory genes. A summary of the supporting data in a chromosome-centric manner is shown in the new version of the Human Protein Atlas launched as part of this publication.

Classification of cell type–specific expression profiles in human blood immune cells

We next performed a genome-wide analysis with regard to expression profiles in the blood cells for the identification of proteins with an elevated expression in immune cells. This was performed both on the cell type level (n = 18 cell types) and on cell lineage level in which the various cell types were combined into six groups, including T cells, B cells, and granulocytes (see full list of results in data S1). The number of genes in each of the five specificity categories is shown in Fig. 5A, with 1448 genes classified as cell type–enriched in one of the cell types and 5934 (30%) of all protein-coding genes elevated in at least one of the human blood cell types. Many genes (n = 3797) were not detected in any of the blood cells, while 9939 showed low specificity for expression in blood cells. The cell type distribution (fig. S9) showed that only 1713 genes were detected in a single cell type, while 5934 were detected in all 18 cell types. The relationship of the two classification schemes is compared in fig. S10 and table S4, showing that 889 genes are cell type–enriched and detected in a single cell type. These genes are of interest for further study to explore the biological functions linked to the respective different cell phenotypes. A heatmap showing the transcript expression profiles for all 1448 immune cell type–enriched genes shows that most are found in neutrophils, basophils, and plasmacytoid dendritic cells (Fig. 5B), while the group-enriched genes are more evenly distributed across the 18 cell types.

A network plot of all cell type–enriched and group-enriched genes (Fig. 5C) reveals a cluster of genes enriched in T cells and another cluster enriched in myeloid cells. Many genes (n = 114) are also shared between the two types of B cell populations (mature and naïve). In Fig. 5D, the number of elevated genes in the different blood cell types, clustered on the basis of the expression profiles, is shown, again highlighting the many cell type–enriched genes in neutrophils, eosinophils, and plasmacytoid dendritic cells, while many of the elevated genes in T and B cells are group-enriched across subpopulations of these lymphocytes. In fig. S11, all group-enriched and tissue-enriched genes are visualized and the relationship of sharing enriched expression between the cell types can be observed.

The extensive data generated here also allowed us to investigate the relationship between (body-wide) tissue expression and the expression in the single blood cell types. In Fig. 5E, a summary of all individual genes is shown with classification based on distribution in all tissues and blood cell types, respectively, and a summary of the genes that are enriched both on tissue level and blood cell level can be found in fig. S15. Some, but not a majority, of the genes expressed in a single or several blood cell types are shown to be predominately expressed in blood cells even when all major tissues and organs are considered. It is notable that many of the genes detected in all tissues are only detected in some of the blood cell types, suggesting that they are not necessary for cell survival.

Enriched genes among the blood immune cell types

Using our definition of cell population enrichment and cell group enrichment of genes, we analyzed the enriched genes among the 18 immune cell populations. Figure 6A shows the top five genes most enriched for each cell population, colored by their predicted protein location either in the membrane, secreted, or intracellular. Notable examples (Fig. 6B) include

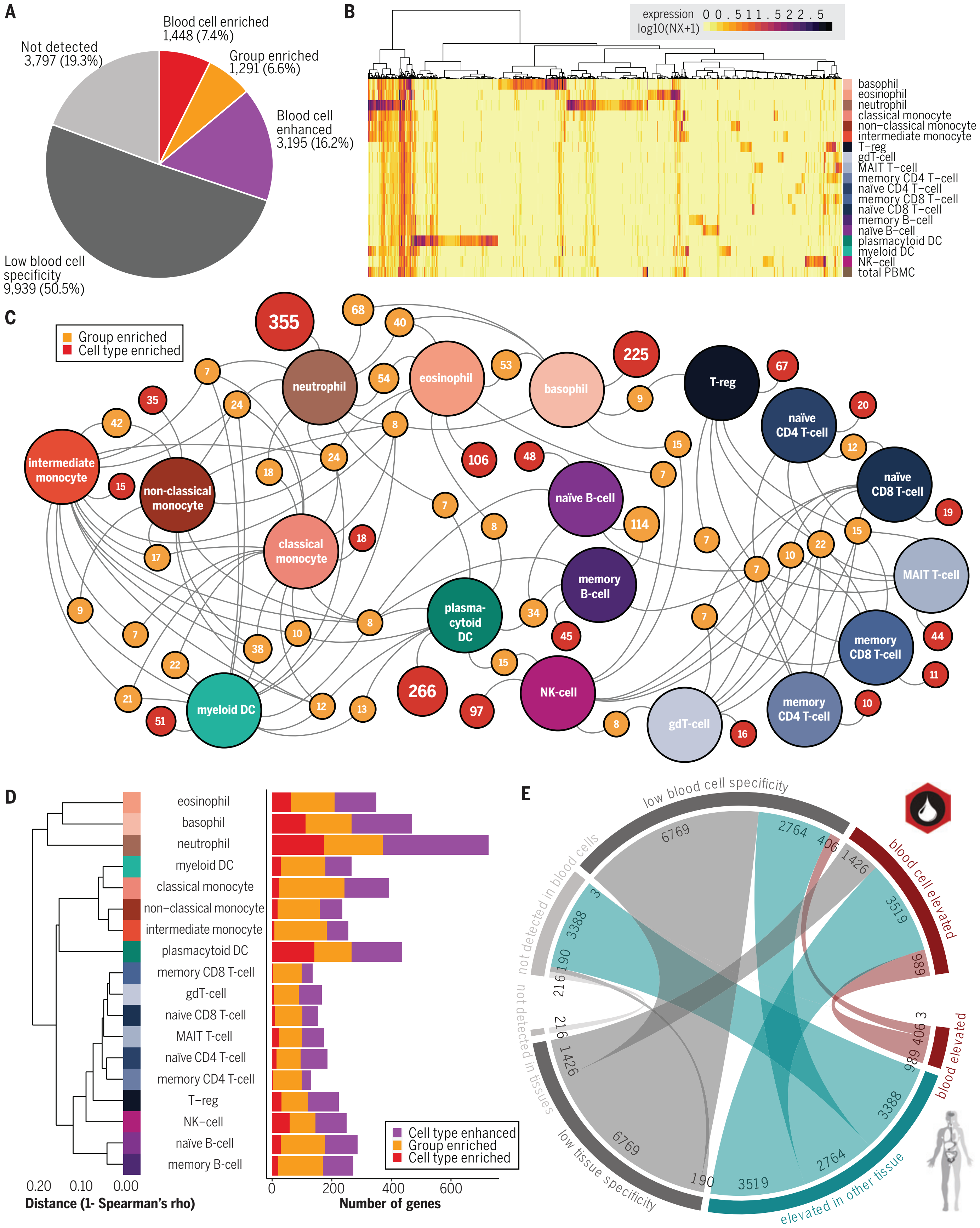

A
Blood cell enriched
1,448 (7.4%)
Group enriched
1,291 (6.6%)
Blood cell
enhanced
3,195 (16.2%)
Not detected
3,797 (19.3%)
Low blood cell
specificity
9,939 (50.5%)
B
expression
log10(NX+1)
0 0.5 1 1.5 2 2.5
basophil
eosinophil
neutrophil
classical monocyte
non−classical monocyte
intermediate monocyte
T−reg
gdT-cell
MAIT T−cell
memory CD4 T−cell
naïve CD4 T−cell
memory CD8 T−cell
naïve CD8 T−cell
memory B−cell
naïve B−cell
plasmacytoid DC
myeloid DC
NK−cell
total PBMC
C
Group enriched
Cell type enriched
355
68
40
225
67
neutrophil
eosinophil
basophil
T-reg
naïve CD4 T-cell
naïve CD8 T-cell
MAIT T-cell
memory CD8 T-cell
memory CD4 T-cell
gdT-cell
NK-cell
plasma-cytoid DC
memory B-cell
naïve B-cell
classical monocyte
non-classical monocyte
intermediate monocyte
myeloid DC
266
97
106
48
114
51
D
eosinophil
basophil
neutrophil
myeloid DC
classical monocyte
non−classical monocyte
intermediate monocyte
plasmacytoid DC
memory CD8 T-cell
gdT-cell
naive CD8 T-cell
MAIT T-cell
naïve CD4 T-cell
memory CD4 T-cell
T−reg
NK−cell
naïve B-cell
memory B-cell
0.20 0.10 0.00
Distance (1- Spearman's rho)
0 200 400 600
Number of genes
Cell type enhanced
Group enriched
Cell type enriched
E
low blood cell specificity
blood cell elevated
blood elevated
elevated in other tissue
low tissue specificity
not detected in tissues
not detected in blood cells
6769
2764
406
1426
3519
689
3
3388
190
216
3388
2764
3519
190
6769
1426
216

Fig. 5. Cell type–specific classification of the human blood cells. (**A**) The number of genes classified according to cell type specificity. (**B**) A heatmap showing the expression of all the cell type–enriched genes across the 18 cell types. Heatmaps for the other specificity categories can be found in figs. S12 to S15. (**C**) Network plot showing the number of cell type– and group-enriched genes in the 18 cell types. The network is limited to nodes with a minimum number of seven genes. (**D**) (Left) A dendrogram based on the correlation of global expression profiles across the 18 cell types. (Right) Barplot displaying the number of elevated genes for each cell type. (**E**) The relationship of all human protein-coding genes with regard to single blood cell type specificity and whole-body tissue and organ specificity.

catalase (CAT), a gene encoding a key antioxidant enzyme converting the toxic reactive oxygen species hydrogen peroxide to water and oxygen and believed to be expressed broadly in the peroxisome of most cells (*38*). Our data indicated a strongly enriched expression level of CAT in eosinophils, which is much higher than the expression in any other immune cell population. This finding warrants more mechanistic analyses of CAT in eosinophils. Another notable finding is the chemokine receptor CXCR6, which is more highly expressed by MAIT cells than any other cell population, suggesting a particular importance of this receptor and its ligand, the chemokine CXCL16, in regulating MAIT cell trafficking. MAIT cells are a population of T cells that has gained a lot of interest in recent years for its role in antibacterial defense, particularly on mucosal sites, through its recognition of molecules derived from the bacterial and fungal riboflavin biosynthesis pathway (*39*). These cells have been shown to express multiple trafficking receptors, and their circulation between blood and tissues has been debated.

Another example is the granzyme B (*GZMB*) gene, a well-known serine protease secreted in granules by cytotoxic T cells and NK cells and necessary for target cell apoptosis (*40*). We found that GZMB expression is strongly enriched in plasmacytoid dendritic cells (pDCs). GZMB expression in pDCs has been reported previously (*40*), but according to our data, GZMB expression in pDCs is about fivefold higher than in any other cell type, which suggests an important function of granzyme B in pDCs (*41*). It is of interest that the population of pDCs also exhibits elevated levels of several other genes (*AXL*, *PPP1R14A*, *SIGLEC6*, *ITM2C*, and *DAB2*) suggested to be specific for a low abundant subgroup of DCs called AS DC with negative GZMB expression, recently described by Villani *et al.* (*42*). Because GZMB variants have been associated with the autoimmune disease vitiligo (*43*), pDCs could potentially play an unappreciated role in the pathogenesis of this condition. To confirm this elevated expression in pDC at the protein level, blood immune cells were analyzed by mass cytometry, and the results confirm higher protein levels of granzyme B in the cytoplasm of pDCs as compared with NK cells and CD8^{+} T cells (Fig. 6C). The GZMB expression levels examined by mass cytometry could not distinguish the proposed AS DC subgroup within the pDC population.

We also complemented our classification strategy by performing a large number of differential expression analyses based on DESeq2 (*44*) to identify genes with variable expression when comparing two cell lineages or two cell populations (fig. S17). The comparison between cell lineages B and T cells show many genes with differential expression, including well-known B cell markers, such as CD19, CD22, and CD79, but also several genes not previously described as elevated in B cells, such as Ras associated domain family member 6 (RASSF6) and the zinc finger protein 860 (ZNF860). Similarly, genes identified as T cell markers include well-known genes, such as *CD3*, *CD6*, inducible T-cell costimulatory (*ICOS*), and thymocyte selection associated (*THEMIS*), but also other genes not yet identified as T cell elevated, such as Ras guanyl releasing protein 1 (*RASGRP1*) and fibroblast growth factor binding protein 2 (*FGFBP2*). All significantly differentially expressed genes for each DESeq2 analysis are available as a separate list (data S2).

Cellular expression of genes causing inborn errors of immunity

In a recent listing of primary immunodeficiency diseases (PID), 354 diseases were listed as consequences of monogenic defects in genes associated with the immune system (*45*) involving 224 known genes. The mechanism of disease is often incompletely understood, and we reasoned that an analysis of cellular expression of identified genes could help generate better hypotheses for further mechanistic investigation. We analyzed the NX levels of 224 PID genes across the 18 sorted immune cell populations, as well as some selected tissue profiles, and identified seven clusters with shared cellular and tissue distribution (Fig. 6D and figs. S18 and S19). A first group (cluster A) consists of 11 proteins restricted to T cells and NK cells, such as CD3 and the signaling intermediates ZAP70 and LCK (Fig. 6E). A second group (cluster B) consists of a subgroup of 15 genes present in all blood cells, but with much lower expression in the other tissues. Cluster C consists of genes ubiquitously expressed across all analyzed tissues and immune cell types. Cluster D consists of 34 proteins mainly originating from the liver and involves known plasma proteins such as complement factors C5, C8, and C9. Cluster E consists of proteins mainly expressed in particular cell lineages, such a B cell–restricted proteins, CD19, and CD79A. Cluster F consists of genes with elevated expression in monocytes and dendritic cells, and cluster G has relatively high expression in lymphoid tissues and bone marrow but low expression in the mature immune cell type in circulation. Several examples of interesting expression patterns can be observed, including the *CEBPE* gene (cluster E) causing specific granule deficiency 1 (SG1) (*46*) that has high expression in eosinophils. This condition has been considered a neutrophil-granule deficiency associated with recurrent pyogenic infections, but our cell type expression pattern indicates that CEBPE is mostly expressed by eosinophils and not at all by neutrophils. It is possible that during neutrophil development, or upon stimulation, CEBPE might also be expressed in neutrophils, but our results suggested that eosinophil deficiency should also be considered in SG1. This use case illustrates the usefulness of the updated human protein atlas as novel genes are identified as possible causes of immunodeficiencies and other diseases in human patients.

Discussion

Here, we present an atlas of the expression of all protein-coding genes in human blood cells, and this data has been integrated with an analysis of the tissue specificity of all genes covering all major tissues and organs in the human body. An interactive Blood Atlas resource is presented as part of the Human Protein Atlas, including expression data from other sources, such as blood cell transcriptomics from Monaco *et al.* (*21*) and Schmiedel *et al.* (*20*). The resource described here enables comparative analysis with other sources of data, such as single-cell genomics, proteomics, and antibody-based measurements, to allow comprehensive molecular profiles of the individual human blood cell types. In addition, the Tissue Atlas (*17*) was complemented with transcript expression data for brain and other normal tissue types from GTEx (*4*) and FANTOM5 (*5*). A normalization strategy has been introduced which has allowed integration of the various diverse datasets to produce a consensus classification across the cells, tissues, and organs. This has enabled the analysis of the cell type–specific expression across the blood immune cell types as well as the various tissues and organs. A revised classification of all protein-coding genes is presented with regard to both cell and tissue distribution.

The tissue expression profiles described earlier (*17*) are supported, but the inclusion of the comprehensive single cell type analysis of human blood, together with inclusion of more brain regions and specialized tissue, has changed some of the patterns of tissue specificity. The

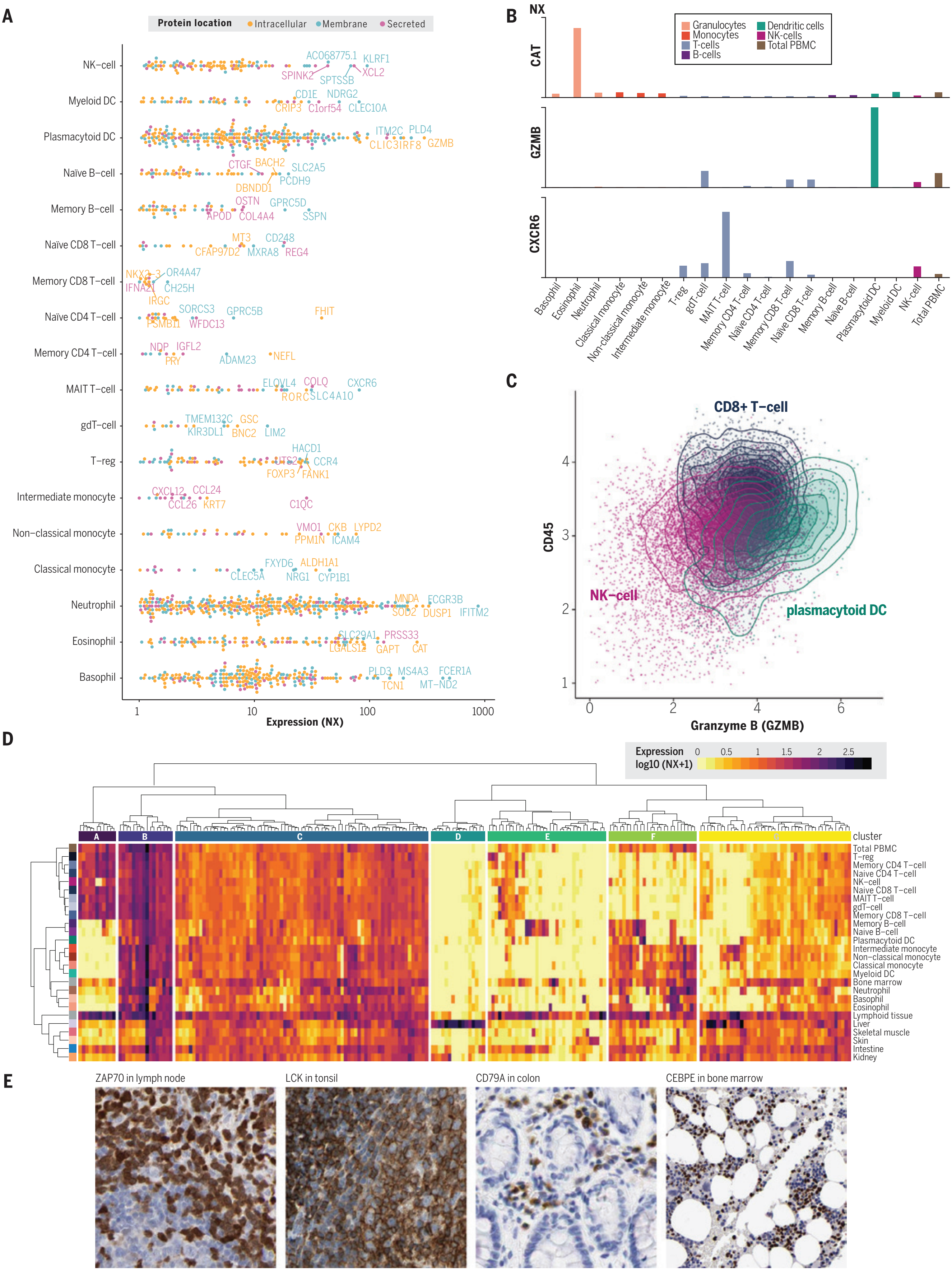

A
Protein location ● Intracellular ● Membrane ● Secreted
NK−cell
AC068775.1 KLRF1 SPINK2 XCL2 SPTSSB
Myeloid DC
CD1E NDRG2 CRIP3 C1orf54 CLEC10A
Plasmacytoid DC
ITM2C PLD4 CLIC3 IRF8 GZMB
Naïve B−cell
CTGF BACH2 SLC2A5 PCDH9 DBNDD1
Memory B−cell
OSTN GPRC5D APOD COL4A4 SSPN
Naïve CD8 T−cell
MT3 CD248 CFAP97D2 MXRA8 REG4
Memory CD8 T−cell
NKX2−3 OR4A47 IFNA21 CH25H IRGC
Naïve CD4 T−cell
SORCS3 GPRC5B FHIT PSMB11 WFDC13
Memory CD4 T−cell
NDP IGFL2 PRY ADAM23 NEFL
MAIT T−cell
ELOVL4 COLQ CXCR6 RORC SLC4A10
gdT−cell
TMEM132C GSC KIR3DL1 BNC2 LIM2
T−reg
HACD1 UTS2 CCR4 FOXP3 FANK1
Intermediate monocyte
CXCL12 CCL24 CCL26 KRT7 C1QC
Non−classical monocyte
VMO1 CKB LYPD2 PPM1N ICAM4
Classical monocyte
FXYD6 ALDH1A1 CLEC5A NRG1 CYP1B1
Neutrophil
MNDA FCGR3B SOD2 DUSP1 IFITM2
Eosinophil
SLC29A1 PRSS33 LGALS12 GAPT CAT
Basophil
PLD3 MS4A3 FCER1A TCN1 MT−ND2
1 10 100 1000
Expression (NX)
B
NX
Granulocytes
Monocytes
T-cells
B-cells
Dendritic cells
NK-cells
Total PBMC
CAT
GZMB
CXCR6
Basophil
Eosinophil
Neutrophil
Classical monocyte
Non-classical monocyte
Intermediate monocyte
T-reg
gdT-cell
MAIT T-cell
Memory CD4 T-cell
Naïve CD4 T-cell
Memory CD8 T-cell
Naïve CD8 T-cell
Memory B-cell
Naïve B-cell
Plasmacytoid DC
Myeloid DC
NK-cell
Total PBMC
C
CD8+ T−cell
NK−cell
plasmacytoid DC
CD45
1 2 3 4
0 2 4 6
Granzyme B (GZMB)
D
Expression log10 (NX+1)
0 0.5 1 1.5 2 2.5
A B C D E F G
cluster
Total PBMC
T−reg
Memory CD4 T−cell
Naive CD4 T−cell
NK−cell
Naive CD8 T−cell
MAIT T−cell
gdT−cell
Memory CD8 T−cell
Memory B−cell
Naive B−cell
Plasmacytoid DC
Intermediate monocyte
Non−classical monocyte
Classical monocyte
Myeloid DC
Bone marrow
Neutrophil
Basophil
Eosinophil
Lymphoid tissue
Liver
Skeletal muscle
Skin
Intestine
Kidney
E
ZAP70 in lymph node
LCK in tonsil
CD79A in colon
CEBPE in bone marrow

Fig. 6. The relationship between blood cell type–specific genes and tissue-specific genes and analysis of genes causing inborn errors of immunity. (**A**) The expression levels of all cell type–enriched genes, with the five most abundant genes named. (**B**) The expression profiles of some selected genes. (**C**) The results of flow sorting (CyTOF) using antibodies toward GZMB and CD45. (**D**) A heatmap showing the expression of 224 genes known to cause human inborn errors of immunity and their expression across all major tissues in the human body. A similar heatmap containing the gene names can be found in fig. S18, and separate heatmaps of each major disease type in all blood cells and tissues can be found in fig. S19. (**E**) IHC images from the Human Protein Atlas for four of the genes causing inborn errors.

brain now has the highest number of elevated genes, while testis still has most enriched genes, defined as an expression fourfold higher than that of any other tissue. The inclusion of more cells and tissues has also allowed us to provide evidence for many more genes, and the total number of missing genes with no protein or RNA evidence is now only ~200. For blood cells, a comprehensive list of all proteins showing an enriched expression in the various cell types is presented, confirming well-known protein markers but also identifying interesting targets for in-depth analysis both to study the basic biology of blood cells and to develop new targets for immune-based diagnostics and therapies. The examples presented here illustrate the potential of the Blood Atlas, and its determination of cell type gene enrichment, for the generation of hypotheses from previously unknown differences in cell population expression of important genes in the immune system.

This newly created resource elucidates the gene expression of individual immune cell populations to allow a better understanding of diseases involving the immune system. The emerging technology of single-cell genomics (*42*, *47*) will in the future be a good complement to such studies to identify low abundant cell subpopulations previously not described. Here, we also highlighted the cell type–specific expression of 224 genes associated with primary immunodeficiencies in humans, and we find cell type–specific expression patterns of relevance for their respective clinical phenotype. A large fraction of these genes is expressed in a large number of cell types, enforcing the need to take a holistic, body-wide approach to identify genes of importance for human biology and diseases. To facilitate such studies, we have launched an interactive, open-access Blood Atlas with all the data integrated as part of the Human Protein Atlas, allowing for genome-wide exploration of the protein-coding genes expressed across immune cell populations and in relation to spatial expression patterns in all major human tissues and organs.

REFERENCES AND NOTES

1. A. Regev *et al*., The Human Cell Atlas. *eLife* **6**, e27041 (2017). doi: 10.7554/eLife.27041; pmid: 29206104
2. J. M. Smith, R. M. Conroy, The NIH Common Fund Human Biomolecular Atlas Program, (HuBMAP): Building a Framework for Mapping the Human Body. *FASEB J.* **32**, 818.2 (2018).
3. J. Kaiser, Chan Zuckerberg Biohub funds first crop of 47 investigators. *Science* 10.1126/science.aal0719 (2017). doi: 10.1126/science.aal0719
4. J. Lonsdale *et al*., The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013). doi: 10.1038/ng.2653; pmid: 23715323
5. H. Kawaji, T. Kasukawa, A. Forrest, P. Carninci, Y. Hayashizaki, The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. *Sci. Data* **4**, 170113 (2017). doi: 10.1038/sdata.2017.113; pmid: 28850107
6. M. J. Hawrylycz *et al*., An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012). doi: 10.1038/nature11405; pmid: 22996553
7. T. Kalisky, S. R. Quake, Single-cell genomics. *Nat. Methods* **8**, 311–314 (2011). doi: 10.1038/nmeth0411-311; pmid: 21451520
8. P. L. Ståhl *et al*., Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016). doi: 10.1126/science.aaf2403; pmid: 27365449
9. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008). doi: 10.1038/nmeth.1226; pmid: 18516045
10. M. Wilhelm *et al*., Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014). doi: 10.1038/nature13319; pmid: 24870543
11. M. Uhlén *et al*., Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010). doi: 10.1038/nbt1210-1248; pmid: 21139605
12. A. Bairoch *et al*., The universal protein resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005). doi: 10.1093/nar/gki070; pmid: 15608167
13. L. C. Crosswell, J. M. Thornton, ELIXIR: A distributed infrastructure for European biological data. *Trends Biotechnol.* **30**, 241–242 (2012). doi: 10.1016/j.tibtech.2012.02.002; pmid: 22417641
14. A. Brazma *et al*., ArrayExpress—A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003). doi: 10.1093/nar/gkg091; pmid: 12519949
15. F. Desiere *et al*., The PeptideAtlas project. *Nucleic Acids Res.* **34**, D655–D658 (2006). doi: 10.1093/nar/gkj040; pmid: 16381952
16. S. Bhattacharya *et al*., ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* **5**, 180015 (2018). doi: 10.1038/sdata.2018.15; pmid: 29485622
17. M. Uhlén *et al*., Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). doi: 10.1126/science.1260419; pmid: 25613900
18. P. J. Thul *et al*., A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017). doi: 10.1126/science.aal3321; pmid: 28495876
19. M. Uhlén *et al*., A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017). doi: 10.1126/science.aan2507; pmid: 28818916
20. B. J. Schmiedel *et al*., Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715.e16 (2018). doi: 10.1016/j.cell.2018.10.022; pmid: 30449622
21. G. Monaco *et al*., RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Reports* **26**, 1627–1640.e7 (2019). doi: 10.1016/j.celrep.2019.01.041; pmid: 30726743
22. J. Lovén *et al*., Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012). doi: 10.1016/j.cell.2012.10.012; pmid: 23101621
23. J. E. Coate, J. J. Doyle, Variation in transcriptome size: Are we getting the message? *Chromosoma* **124**, 27–43 (2015). doi: 10.1007/s00412-014-0496-3; pmid: 25421950
24. A. S. Kopin, M. B. Wheeler, A. B. Leiter, Secretin: Structure of the precursor and tissue distribution of the mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2299–2303 (1990). doi: 10.1073/pnas.87.6.2299; pmid: 2315322
25. S. Kuttruff *et al*., NKp80 defines and stimulates a reactive subset of CD8 T cells. *Blood* **113**, 358–369 (2009). doi: 10.1182/blood-2008-03-145615; pmid: 18922855
26. P. Brodin *et al*., Variation in the human immune system is largely driven by non-heritable influences. *Cell* **160**, 37–47 (2015). doi: 10.1016/j.cell.2014.12.020; pmid: 25594173
27. H. Wold, "Estimation of principal components and related models by iterative least squares" in *Multivariate Analysis*, P. R. Krishnajah, Ed. (Academic Press, 1966), pp. 391–420.
28. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML] (9 February 2018).
29. M. Uhlén, Mapping the human proteome using antibodies. *Mol. Cell. Proteomics* **6**, 1455–1456 (2007). pmid: 17703056
30. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010). doi: 10.1186/gb-2010-11-3-r25; pmid: 20196867
31. R. A. van den Berg, H. C. J. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142 (2006). doi: 10.1186/1471-2164-7-142; pmid: 16762068
32. M. E. Ritchie *et al*., *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015). doi: 10.1093/nar/gkv007; pmid: 25605792
33. T. Hart *et al*., High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015). doi: 10.1016/j.cell.2015.11.015; pmid: 26627737
34. T. Wang *et al*., Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015). doi: 10.1126/science.aac7041; pmid: 26472758
35. L. Fagerberg *et al*., Contribution of antibody-based protein profiling to the human Chromosome-centric Proteome Project (C-HPP). *J. Proteome Res.* **12**, 2439–2448 (2013). doi: 10.1021/pr300924j; pmid: 23276153
36. M. Magrane; UniProt Consortium, UniProt Knowledgebase: A hub of integrated protein data. *Database* **2011**, bar009 (2011). pmid: 21447597
37. P. Gaudet *et al*., The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* **45**, D177–D182 (2017). doi: 10.1093/nar/gkw1062; pmid: 27899619
38. P. Chelikani, I. Fita, P. C. Loewen, Diversity of structures and properties among catalases. *Cell. Mol. Life Sci.* **61**, 192–208 (2004). doi: 10.1007/s00018-003-3206-5; pmid: 14745498
39. R. J. Napier, E. J. Adams, M. C. Gold, D. M. Lewinsohn, The role of mucosal associated invariant T cells in antimicrobial immunity. *Front. Immunol.* **6**, 344 (2015). doi: 10.3389/fimmu.2015.00344; pmid: 26217338
40. M.-C. Rissoan *et al*., Subtractive hybridization reveals the expression of immunoglobulin-like transcript 7, Eph-B1, granzyme B, and 3 novel transcripts in human plasmacytoid dendritic cells. *Blood* **100**, 3295–3303 (2002). doi: 10.1182/blood-2002-02-0638; pmid: 12384430
41. C. Chauvin, R. Josien, Dendritic cells as killers: Mechanistic aspects and potential roles. *J. Immunol.* **181**, 11–16 (2008). doi: 10.4049/jimmunol.181.1.11; pmid: 18566364
42. A.-C. Villani *et al*., Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017). doi: 10.1126/science.aah4573; pmid: 28428369
43. Y. Jin *et al*., Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat. Genet.* **48**, 1418–1424 (2016). doi: 10.1038/ng.3680; pmid: 27723757
44. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). doi: 10.1186/s13059-014-0550-8; pmid: 25516281
45. P. W. Sullivan, V. H. Ghushchyan, G. Globe, M. Schatz, Oral corticosteroid exposure and adverse effects in asthmatic patients. *J. Allergy Clin. Immunol.* **141**, 110–116.e7 (2018). doi: 10.1016/j.jaci.2017.04.009; pmid: 28456623
46. Online Mendelian Inheritance in Man, OMIM, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (2018); https://omim.org/.

47. V. S. Patil *et al*., Precursors of human CD4$^+$ cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci. Immunol.* **3**, eaan8664 (2018). doi: 10.1126/sciimmunol.aan8664; pmid: 29352091

ACKNOWLEDGMENTS

We thank the nurses at the Coagulation Unit, Karolinska University Hospital, for their assistance in handling donors and sampling. We acknowledge the entire staff of the Human Protein Atlas program and the Science for Life Laboratory for their valuable contributions. **Funding:** Funding was provided by the Knut and Alice Wallenberg Foundation (WCPR), the Erling Persson Foundation (KCAP), and the Novo Nordisk Foundation (CFB). Support from the National Genomics Infrastructure in Stockholm is acknowledged, with funding from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science. **Author contributions:** M.U. and L.F. conceived of and designed the study. C.P., J.Mi., T.L., and P.B. performed the single cell sorting and analysis. B.F., F.E., and J.O. collected the clinical samples. M.U., Å.S., A.M., C.L., E.S., J.Mi., B.F., F.E., F.P., A.H., W.Z., M.J.K., J.Mu., A.T., C.Z., and L.F. performed the data analysis. K.v.F., P.O., and M.Z. provided the infrastructure for the data. M.U. drafted the manuscript. M.U., L.F., P.B., M.J.K., and Å.S. revised the manuscript. All authors discussed the results and contributed to the final manuscript. **Competing interests:** No competing interests. **Data and materials availability:** All raw flow cytometry data are available at FlowRepository (http://flowrepository.org/) under ID FR-FCM-Z28R. Sequencing data used in the study are available without restriction at the Human Protein Atlas portal (www.proteinatlas.org/about/download). All custom code used for normalization and categorization can be downloaded from Github (https://github.com/human-protein-atlas/BloodAtlas).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/366/6472/eaax9198/suppl/DC1
Materials and Methods
Figs. S1 to S21
Tables S1 to S4
References (*48*–*61*)
Data S1 to S3

10 May 2019; accepted 6 November 2019
10.1126/science.aax9198

RESEARCH ARTICLE SUMMARY

INTEGRATIVE OMICS

An atlas of the protein-coding genes in the human, pig, and mouse brain

Evelina Sjöstedt, Wen Zhong, Linn Fagerberg, Max Karlsson, Nicholas Mitsios, Csaba Adori, Per Oksvold, Fredrik Edfors, Agnieszka Limiszewska, Feria Hikmet, Jinrong Huang, Yutao Du, Lin Lin, Zhanying Dong, Ling Yang, Xin Liu, Hui Jiang, Xun Xu, Jian Wang, Huanming Yang, Lars Bolund, Adil Mardinoglu, Cheng Zhang, Kalle von Feilitzen, Cecilia Lindskog, Fredrik Pontén, Yonglun Luo, Tomas Hökfelt, Mathias Uhlén*†, Jan Mulder*†

INTRODUCTION: The brain is the most complex organ of the mammalian body, boasting a diverse physiology combined with intricate cellular organization. In an effort to expand our basic understanding of the neurobiology of the brain and its diseases, we performed a comprehensive molecular dissection of the main regions of the human, pig, and mouse brain using transcriptomics and antibody-based mapping. With this approach, we have identified regional expression profiles and observed similarities and differences in expression levels between these three mammalian species.

RATIONALE: There is a need for a comprehensive overview of genes expressed in the mammalian brain categorized by organ, brain region, and species specificity. To address this need, a brain-centered knowledge resource of RNA and protein expression in the brain of three mammalian species has been created and used for cell topological analysis, systems modeling, and data integration. The regional expression of all protein-coding genes is reported, and this classification is integrated with results from the analysis of tissues and organs of the whole human body. All generated data, including high-resolution images and metadata, have been made publicly available in an open-access Human Protein Atlas (HPA) Brain Atlas.

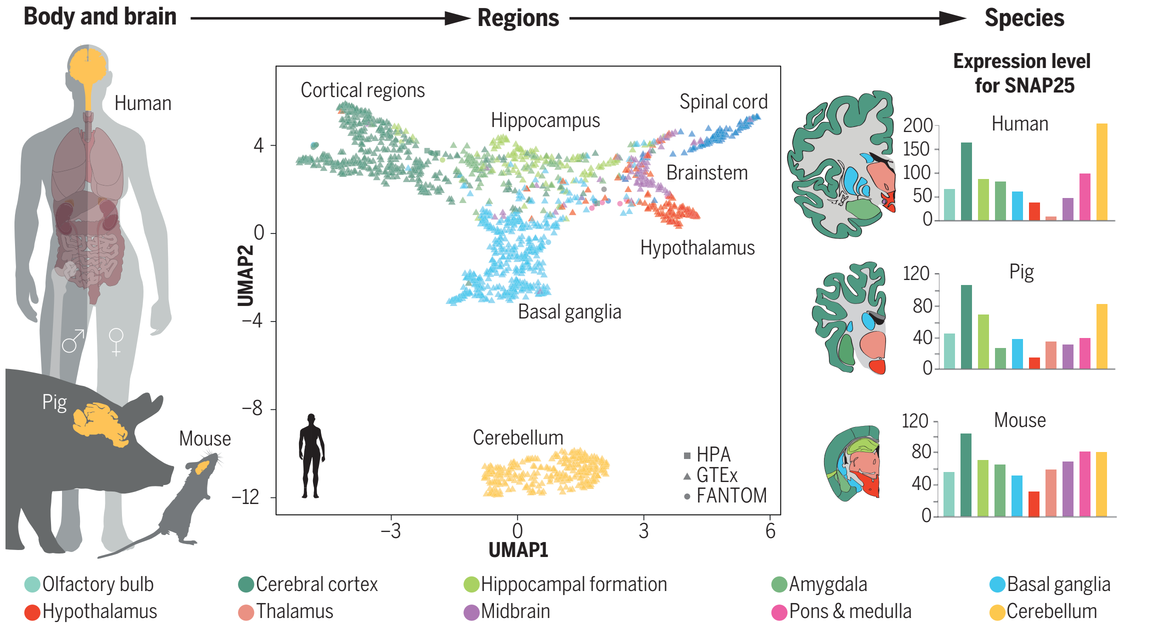
RESULTS: The global analysis suggests similar regional organization and expression patterns in the three mammalian species, consistent with the view that basic brain architecture is preserved during mammalian evolution. However, there is considerable variability between species for many neurotransmitter receptors, in particular between human and mouse. This calls for caution when using the mouse as a model system for the human brain, for example, in attempts to develop therapeutic strategies. For some of the brain regions, such as the cerebellum and hypothalamus, the human global expression profile is closer to that of the pig than it is to that of the mouse, suggesting that the pig might be considered a preferred animal model to study many brain processes. We show that many "signature genes" identified previously for specific brain cell types (such as astrocytes, microglia, oligodendrocytes, and neurons) are expressed at even higher levels in peripheral organs. In fact, our results support a view of shared functions between many genes in microglia and immune cells, and a large number of genes previously identified as signature genes for astrocytes are shown to be shared with liver or skeletal muscle. The cerebellum stands out as having a distinct molecular signature with many regionally enriched genes. Several genes suggested to be involved in neuropsychiatric diseases are selectively expressed in the cerebellum.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aay5947

CONCLUSION: The integration of data from several sources has allowed us to combine data from transcriptomics, single-cell genomics, in situ hybridization, and antibody-based protein profiling. This integrative approach for mapping the molecular profiles in the human, pig, and mouse brain has generated a detailed multilevel genome-wide view on the protein-coding genes of the mammalian brain, where we compared tissue specificity across the whole body, as classified in the HPA (www.proteinatlas.org). The open-access HPA Brain Atlas resource offers the opportunity to explore individual genes and classes of genes and their expression profiles in the various parts of the mammalian brain. ■

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: mathias.uhlen@scilifelab.se (M.U.); jan.mulder@ki.se (J.M.)
Cite this article as E. Sjöstedt *et al.*, *Science* **367**, eaay5947 (2020). DOI: 10.1126/science.aay5947



Genome-wide transcriptomics analysis of anatomically dissected regions in mammalian brains uncovers regional and species-specific expression. Multiple regions of the human, pig, and mouse brain were dissected and analyzed. A uniform manifold approximation and projection (UMAP) analysis (middle) shows the global expression patterns of 1710 samples in the human brain, with the cerebellum as the outlier. The HPA Brain Atlas (right) shows the expression of individual genes, for example, synaptosomal-associated protein 25 (*SNAP25*), in the different brain regions in the three mammalian species.

Originally published 6 March 2020

RESEARCH ARTICLE

INTEGRATIVE OMICS

An atlas of the protein-coding genes in the human, pig, and mouse brain

Evelina Sjöstedt[1,2], Wen Zhong[2], Linn Fagerberg[2], Max Karlsson[2], Nicholas Mitsios[1], Csaba Adori[1], Per Oksvold[2], Fredrik Edfors[2], Agnieszka Limiszewska[1], Feria Hikmet[3], Jinrong Huang[4,5,6,7], Yutao Du[5,8], Lin Lin[4,6], Zhanying Dong[4,5,6], Ling Yang[4,5,6], Xin Liu[5,8], Hui Jiang[9], Xun Xu[5,8], Jian Wang[5,8], Huanming Yang[5,8], Lars Bolund[4,5,6], Adil Mardinoglu[2], Cheng Zhang[2], Kalle von Feilitzen[2], Cecilia Lindskog[3], Fredrik Pontén[3], Yonglun Luo[4,5,6], Tomas Hökfelt[1], Mathias Uhlén[1,2]*†, Jan Mulder[1]*†

The brain, with its diverse physiology and intricate cellular organization, is the most complex organ of the mammalian body. To expand our basic understanding of the neurobiology of the brain and its diseases, we performed a comprehensive molecular dissection of 10 major brain regions and multiple subregions using a variety of transcriptomics methods and antibody-based mapping. This analysis was carried out in the human, pig, and mouse brain to allow the identification of regional expression profiles, as well as to study similarities and differences in expression levels between the three species. The resulting data have been made available in an open-access Brain Atlas resource, part of the Human Protein Atlas, to allow exploration and comparison of the expression of individual protein-coding genes in various parts of the mammalian brain.

The brain is an extraordinarily complex organ owing to its diverse physiology, complex cellular organization, and abundance of expressed genes. Identifying the molecular organization of the brain at regional, cellular, and subcellular levels will advance our understanding of its function under normal and diseased conditions. The Human Protein Atlas (HPA) program aims to combine antibody-based profiling with genome-wide transcriptomics analysis to explore the spatial expression levels of transcripts and proteins across cells, tissues, and organs (*1*). The Tissue Atlas (*1*, *2*)—a subsection of the HPA—includes only a limited number of human brain regions (the cerebral cortex, hippocampus, caudate nucleus, and cerebellum). Here, we describe genome-wide expression profiles for the protein-coding genes in 10 major well-defined mammalian brain regions to capture the complexity of the cellular organization. To identify differences and similarities of the brain in different phylogenetic orders, the expression profiles have been analyzed in three species: primates (human), Cetartiodactyla (pig), and Rodentia (mouse).

The effort described here is complementary to several brain mapping projects focused on basic organization and regional or cellular gene expression of the mammalian brain. The Allen Institute for Brain Science (https://alleninstitute.org) hosts several knowledge resources, including an in situ hybridization atlas of the adult (*3*) and developing (*4*) mouse brain; and a microarray-based atlas of the adult human brain (*5*) has been complemented with a map of the human brain during development (*6*). More recently, brain atlas strategies have been launched on the basis of different approaches: fluorescence-activated cell sorting in mouse (*7*), antibody-based cell sorting in human (*8*), single-cell gene expression in mouse (*9*) and human (*10*, *11*), and covariation analysis of transcriptomics expression (*12*). These efforts have been further complemented with several large-scale mapping programs, including the National Institutes of Health (NIH) BRAIN Initiative Cell Census Network (*13*), the European Human Brain Project (*14*), the NIH Human BioMolecular Atlas Program (*15*), and the Human Cell Atlas project (*16*).

Here, we present the HPA Brain Atlas (*17*), where the data collected have been used for cell topological analysis, systems modeling, and data integration, with the aim to create a knowledge resource of messenger RNA and protein expression in the mammalian brain. We complement the transcriptomics with antibody-based protein profiles of selected proteins in multiple regions of the mouse brain. In this open-access resource, transcriptomics data from three external sources—the Genotype-Tissue Expression (GTEx) portal (*18*), the Functional Annotation of Mammalian Genomes 5 (FANTOM5) project (*19*), and the Allen Mouse Brain Atlas (*3*)—are presented together with RNA profiles and protein stainings generated "in-house." The classification of all protein-coding genes with regard to brain regional specificity is reported, and this is integrated with the tissue and organ specificity across the human body.

Transcriptomics analysis of the human brain

Transcriptomics analysis was performed on anatomically dissected human, pig, and mouse brain regions (Fig. 1A and figs. S1 to S3). For the human brain, we integrated publicly available RNA sequencing (RNA-seq) data generated by the GTEx consortium (*18*) and cap analysis of gene expression (CAGE) data from the FANTOM consortium (*19*), with data from the HPA (*1*), for a total of 1710 samples from selected human brain regions (table S1). The combined dataset contains 23 human brain regions, including white matter (corpus callosum) and spinal cord, as outlined in Fig. 1B. Several issues complicate the combining of datasets. First, samples may not be homogeneous, especially for regions with a high level of cellular heterogeneity, such as the hypothalamus, midbrain, pons, and medulla oblongata. Furthermore, both HPA and GTEx data are based on RNA-seq protocols using polyadenylate [poly(A)] tail enrichment, whereas CAGE data are based on the selection and sequencing of the 5′ cap. As a result, genes lacking the poly(A) tail, such as canonical histone mRNA, are only detected by CAGE. Despite these complications, the large number of included samples and our gene classification approach enable us to generate a comprehensive overview of biologically relevant gene expression and regional and species variation.

We used normalization strategies to avoid batch effects caused by sampling, technology platforms, and differences in transcriptome size between different types of tissues and also to allow both within-sample and between-sample comparisons (*20*, *21*). The within-sample normalization was based on protein-coding transcripts per million (pTPM), while the between-sample normalization was based on trimmed means of M values (TMM) (*22*), Pareto scaling per gene (*23*), and *limma* (*24*), resulting in normalized expression (NX) values calculated for all genes across all tissue types, as outlined in Fig. 1C and described in detail in the supplementary information (figs. S4 to S6).

The uniform manifold approximation and projection (UMAP) (fig. S7) of all 1710 human brain samples shows the expected global expression patterns after normalization. Developmentally related anatomical regions cluster together, with the cerebellum being an outlier

[1]Department of Neuroscience, Karolinska Institutet, 171 77 Stockholm, Sweden. [2]Department of Protein Science, Science for Life Laboratory, KTH-Royal Institute of Technology, 17121 Stockholm, Sweden. [3]Department of Immunology, Genetics and Pathology, Uppsala University, 751 85 Uppsala, Sweden. [4]Lars Bolund Institute of Regenerative Medicine, BGI-Qingdao, Qingdao 266555, China. [5]BGI-Shenzhen, Shenzhen 518083, China. [6]Department of Biomedicine, Aarhus University, 80000 Aarhus, Denmark. [7]Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark. [8]China National GeneBank, BGI-Shenzhen, Shenzhen 518083, China. [9]MGI, BGI-Shenzhen, Shenzhen 518083, China.
*These authors contributed equally to this work.
†Corresponding author. Email: mathias.uhlen@scilifelab.se (M.U.); jan.mulder@ki.se (J.M.)

Originally published 6 March 2020

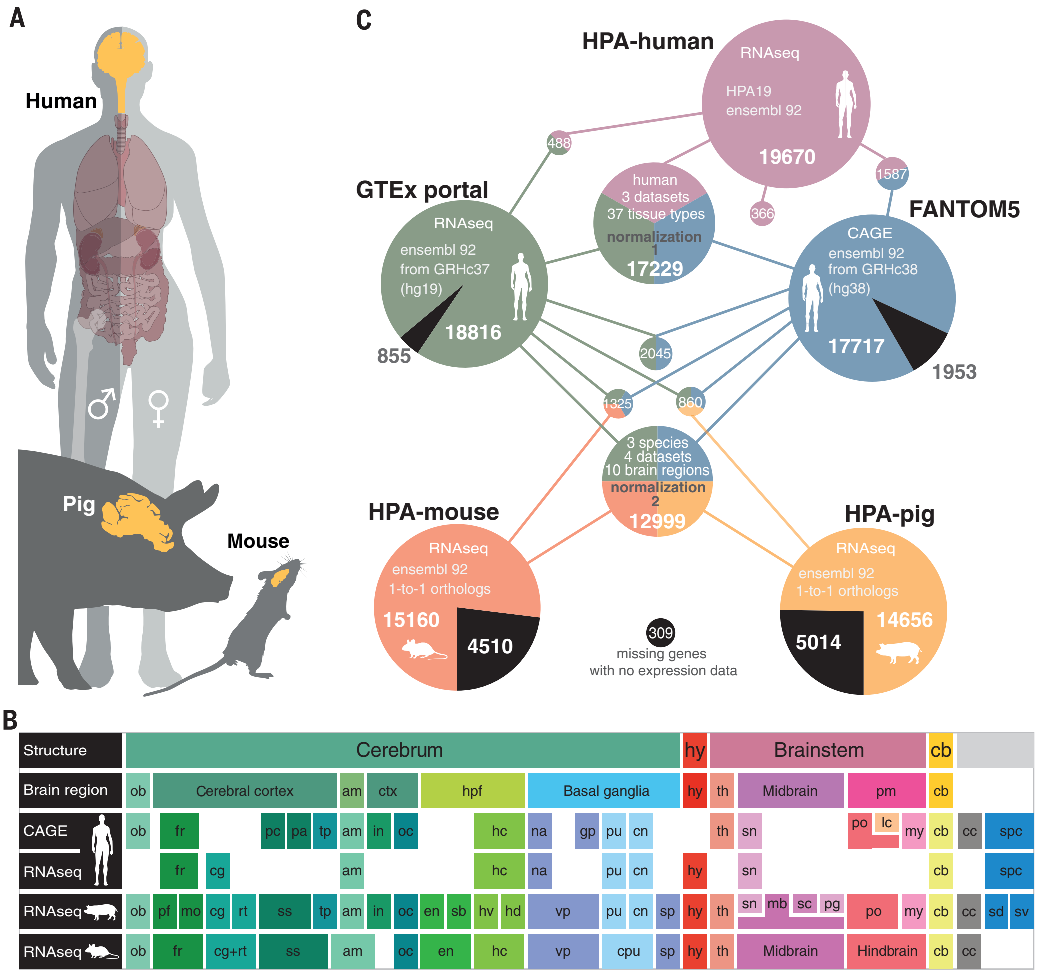

A
Human
♂
♀
Pig
Mouse
C
HPA-human
RNAseq
HPA19
ensembl 92
19670
488
1587
366
GTEx portal
RNAseq
ensembl 92
from GRHc37
(hg19)
18816
855
human
3 datasets
37 tissue types
normalization
1
17229
FANTOM5
CAGE
ensembl 92
from GRHc38
(hg38)
17717
1953
2045
1325
860
3 species
4 datasets
10 brain regions
normalization
2
12999
HPA-mouse
RNAseq
ensembl 92
1-to-1 orthologs
15160
4510
309
missing genes
with no expression data
HPA-pig
RNAseq
ensembl 92
1-to-1 orthologs
14656
5014
B
Structure: Cerebrum | hy | Brainstem | cb
Brain region: ob | Cerebral cortex | am | ctx | hpf | Basal ganglia | hy | th | Midbrain | pm | cb
CAGE: ob | fr | pc | pa | tp | am | in | oc | hc | na | gp | pu | cn | th | sn | po | lc | my | cb | cc | spc
RNAseq: fr | cg | am | hc | na | pu | cn | hy | sn | cb | spc
RNAseq: ob | pf | mo | cg | rt | ss | tp | am | in | oc | en | sb | hv | hd | vp | pu | cn | sp | hy | th | sn | mb | sc | pg | po | my | cb | cc | sd | sv
RNAseq: ob | fr | cg+rt | ss | am | oc | en | hc | vp | cpu | sp | hy | th | Midbrain | Hindbrain | cb | cc

Fig. 1. Genome-wide transcriptomics analysis of anatomically dissected regions in mammalian brains. (**A**) Multiple regions of the human, pig, and mouse brain were dissected and analyzed using transcriptomics methods. (**B**) A summary of the included brain subregions, with 23 human, 30 pig, and 17 mouse samples, in 10 main brain regions (for an anatomical overview, see figs. S1 to S3). The subregions are as follows: olfactory bulb, ob; prefrontal cortex, pf; frontal lobe, fr; motor cortex, mo; cingulate cortex, cg; retrosplenial cortex, rt; somatosensory cortex, ss; paracentral gyrus, pa; postcentral gyrus, pc; temporal lobe, tp; insula cortex, in; occipital lobe, oc; entorhinal cortex, en; subiculum, sb; amygdala, am; hippocampus, hc (ventral, hv, and dorsal, hd); nucleus accumbens, na; ventral pallidum, vp; globus pallidus, gp; putamen, pu; caudate nucleus, cn; caudate putamen, cpu; septum, sp; hypothalamus, hy; thalamus, th; substantia nigra, sn; midbrain, mb; superior colliculus, sc; periaqueductal gray, pg; pons, po; locus coeruleus, lc; medulla oblongata, my; cerebellum, cb; corpus callosum, cc; spinal cord, spc (dorsal, sd, and ventral, sv). (**C**) Overview of the data normalization approach, combining five separate datasets. Total gene numbers for respective datasets are shown, as well as genes overlapping and nonoverlapping between datasets (see fig. S5 for extended version).

compared with other brain regions. No bias between the different platforms for transcriptomics analysis (HPA, GTEx, and FANTOM) was observed. The expression data for all analyzed human brain regions covering 19,670 human protein-coding genes are presented in a gene-specific manner in the HPA Brain Atlas (see below). The regional expression data in the human brain include 15,157 protein-coding genes detected in at least one region of the brain, ranging from 13,068 to 14,332 expressed genes per brain region (fig. S8).

Transcriptomics analysis of the pig brain

Brain transcriptome analysis of two male and two female adult pigs (Bama minipig, aged 1 year) was performed for anatomically dissected brain regions covering the whole brain, as outlined in table S2. The pig brain was divided into 30 anatomically defined brain regions (Fig. 1B). A normalization protocol using TMM and Pareto scaling was used, as outlined in fig. S4. A UMAP analysis of the transcript expression profiles of all samples (fig. S9) indicates the overall similarity between subregions and illustrates the expression variation between the cerebrum regions and regions of the brainstem. On the basis of the pig gene build Ensembl 92 and a detection cutoff at NX = 1, a total of 18,686 genes were detected in the pig brain, with 15,601 to 17,394 genes detected in individual brain regions (fig. S10). Expression data for 14,656 protein-coding genes with a one-to-one pig ortholog can be found in the gene-specific pages of the HPA Brain Atlas (*17*).

Transcriptomics analysis of the mouse brain

A genome-wide transcriptomics analysis was performed on multiple regions of two male and two female adult mice (C57bl/6n, aged 2 months). The mouse brain was divided into 17 anatomically defined brain regions (table S3). A normalization protocol using TMM and Pareto scaling was used, as outlined in fig. S4. The UMAP plot of the global expression patterns shows the expected pattern with developmentally related anatomical regions clustering together (fig. S11). On the basis of a cutoff for detection at NX = 1, a total of 15,823 brain-expressed mouse genes were detected, with 12,977 to 14,402 genes per brain region (fig. S12). Data for 15,160 protein-coding genes with a mouse one-to-one ortholog are presented in the gene-specific pages of the HPA Brain Atlas (*17*).

Genome-wide classification of all protein-coding genes based on regional brain expression

Expression data for the various brain regions of the three species were summarized into 10 main regions (Fig. 2, A to C). On the basis of the maximum expression in any of the analyzed subregions, a consensus result of the 10 regions for three species was generated. These regions are the olfactory bulb, all cerebral cortex regions, subfields of the hippocampus, the amygdala, regions of the basal ganglia, the hypothalamus, the thalamus, subfields of the midbrain, the pons and medulla oblongata, and the cerebellum (Fig. 1B). A hierarchical clustering of the 10 main regions was performed using the global expression profiles

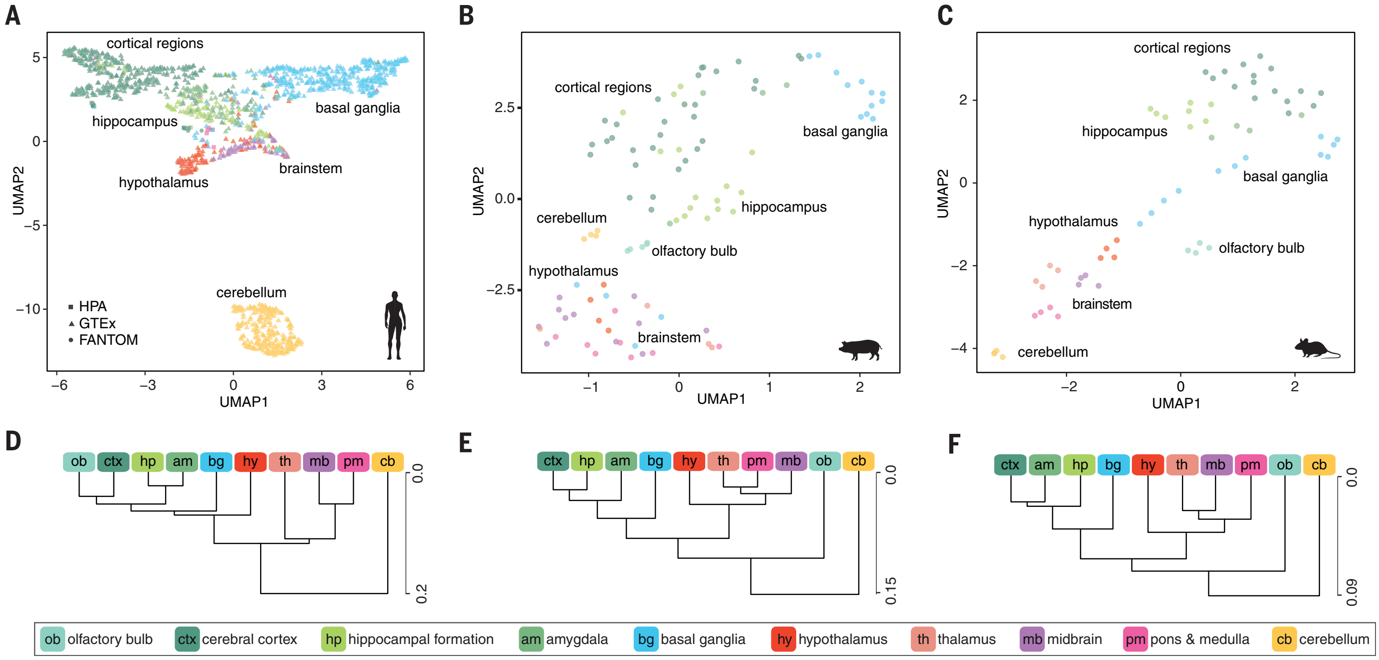


Fig. 2. Regional comparison based on global expression in three mammalian species. (**A**) Uniform manifold approximation and projection (UMAP) analysis showing the global expression patterns of all samples in 10 human brain regions (1616 total) from HPA, GTEx, and FANTOM. (**B**) UMAP plot of pig brain samples in 10 regions (107) used for mapping regional transcript expression in the pig brain. (**C**) UMAP plot of mouse brain samples in 10 regions (64) analyzed in this study from the mouse brain. (**D**) Hierarchical clustering based on pair-wise Spearman correlation of the transcript expression levels in 10 main brain regions is shown. (**E** and **F**) Same as (D), but for pig and mouse brain regions, respectively.

of all the protein-coding genes. The results for each of the three mammalian brains are shown in Fig. 2, D to F, with details in fig. S13. The hierarchical trees show a similar structure, with the cerebellum as an outlier in all three species and with the three cerebrum regions (cerebral cortex, hippocampus, and amygdala) close together, next to the basal ganglia. Similarly, the three brainstem regions (midbrain, thalamus, and pons and medulla) cluster together in all three species, next to the hypothalamus. The analysis confirms that the global expression patterns in the different regions of the three mammalian brains are preserved during mammalian evolution.

To identify regionally specific molecular features, regionally elevated genes were classified according to their expression across the 10 main brain regions. Elevated genes were further stratified into regionally enriched (fourfold higher expression compared with any other brain region), group enriched (several brain regions with fourfold higher values than all other regions), and regionally enhanced (fourfold higher expression than the average expression of the 10 regions). Genes not elevated in a single region or group of brain regions are classified as genes with low regional specificity (the classification is described in detail in table S4). This classification was performed across all protein-coding genes on the basis of NX values. The numbers of regionally enriched, group enriched, and regionally enhanced genes are shown in fig. S14 and in the HPA Brain Atlas resource (see below). Heatmaps show the distribution of regionally enriched, group-enriched, and regionally enhanced genes in the 10 regions of the human (fig. S15), pig (fig. S16), and mouse (fig. S17) brain. In all three species, the cerebellum contains the largest number of regionally enriched genes, while most group enriched genes are shared among the regions of the cerebrum and brainstem, respectively (fig. S18).

Comparative analysis of transcriptomics, in situ RNA hybridization, and immunofluorescence protein staining

A comprehensive and extensively used mouse brain gene expression atlas has been generated by the Allen Institute using probe-based in situ transcriptomics (*3*). In the HPA Brain Atlas, we have integrated expression profiles from the Allen Brain Atlas for all mouse genes (with a human one-to-one ortholog) with the HPA-generated RNA-seq and antibody-based protein distribution data. The two transcriptomics sets are highly complementary, because RNA-seq expression data provide sensitive quantitative transcript information, although these data have the disadvantage that mixtures of cell types are analyzed. The in situ hybridization data provide spatial expression data on a single-cell level, but this probe-based method is less quantitative than the count-based RNA-seq method. In addition, for selected proteins, an immunofluorescence protein distribution map was generated, allowing visualization of protein distribution on a cellular level, including neuronal processes, with high spatial resolution. An advantage of this protein staining is that anatomically stacked images can be generated, and this has allowed us to annotate more than 120 regions and subfields of the brain. Together, the three complementary datasets provide genome-wide regional profiles of the protein-coding genes and their expression in the different regions of the brain.

The results are displayed in the HPA Brain Atlas, and this resource allows for comparisons of the HPA data (RNA-seq), the probe-based in situ hybridization (ISH), and the antibody-based protein immunofluorescence (IF) staining for all 10 regions of the mouse brain, as exemplified for five genes in Fig. 3A. Insulin-like growth factor binding protein 5 (IGFBP5) is shown to be expressed in all analyzed regions of the mouse brain according to all datasets. However, both the mRNA location (ISH) and immunoreactivity (IF) reveal a distinct expression pattern in the mouse olfactory bulb with expression in mitral cells, localized both in soma and proximal dendrites. For NECAB1, an N-terminal EF-hand calcium binding protein with unknown function, brain-wide expression is also observed. The ISH and

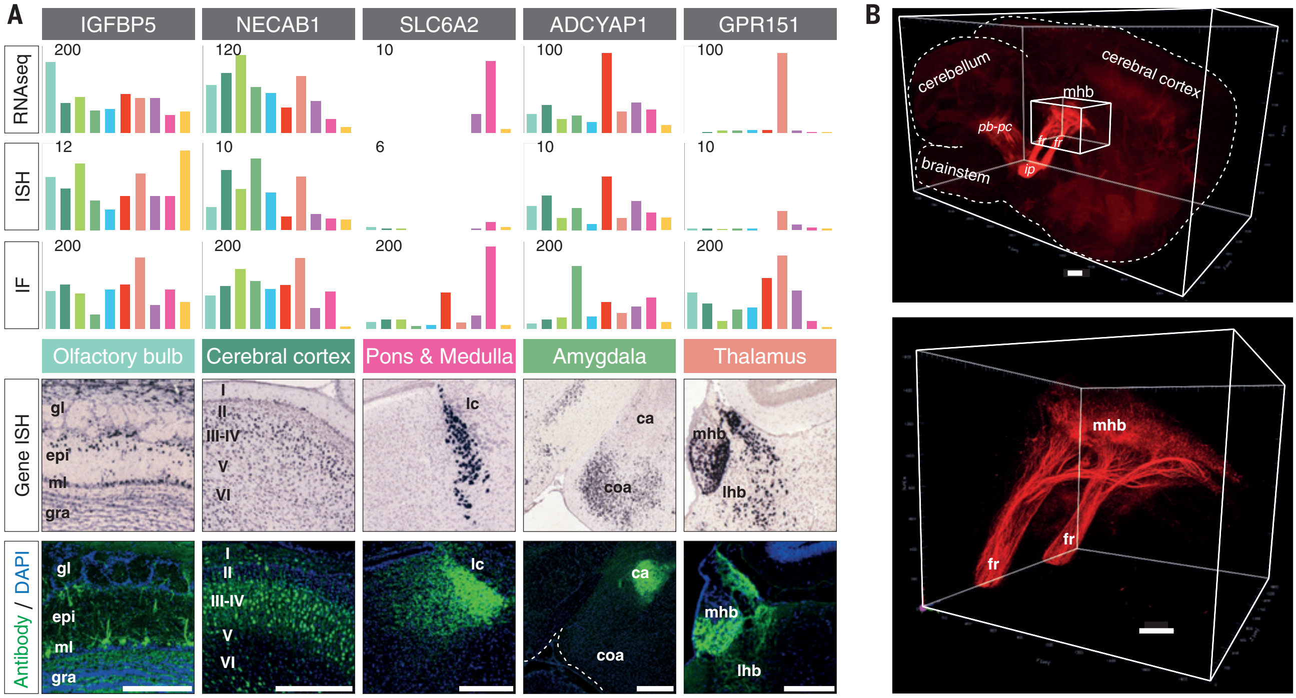


Fig. 3. Comparison of mouse regional data from transcriptomics, in situ hybridization, and immunofluorescence. (**A**) Examples of expression profiles from the HPA Brain Atlas, with mouse expression shown as data from RNA-seq (this study), in situ hybridization (ISH) data from the Allen Brain Atlas, and regional staining intensity based on antibody-based immunofluorescence (IF) profiling (this study). The color codes are the same as in Fig. 2. Below are examples of ISH and IF staining for each gene, from left to right: insulin-like growth factor binding protein 5 (IGFBP5); N-terminal EF-hand calcium binding protein 1 (NECAB1); norepinephrine transporter, also called solute carrier family 6 member 2 (NET1 or SLC6A2); adenylate cyclase–activating polypeptide 1 (ADCYAP1); and G protein–coupled receptor 151 (GPR151). Glomerular layer, gl; external plexiform layer, epi; mitral cell layer, ml; granule cell layer, gra; locus coeruleus, lc; central amygdala, ca; cortical amygdala, coa; lateral habenula, lhb. (**B**) iDISCO+ volume immunostaining of a whole mouse brain for GPR151 receptor. The medial habenula (mhb), fasciculus retroflexus (fr), interpeduncular nucleus (ip), and parabrachial-pericoerulear region (pb-pc) are strongly stained. The boxed region in the top image is enlarged in the bottom image. Scale bars, 250 μm.

the IF data indicate a distinct neuronal expression of NECAB1 in subsets of neurons in various regions of the thalamus and forebrain, including pyramidal neurons in the cerebral cortex. The (nor)epinephrine uptake transporter (SLC6A2), also called NET1, is an example of an apparent partial discrepancy between the RNA and the protein location: The RNA transcript is detected in cell bodies of the locus coeruleus in the pons with an expression pattern resembling that of the well-characterized (nor)epinephrinergic neurons. However, this transporter protein cannot be detected in the cell bodies with IF. This is because NET1 is rapidly transported into the extensive axonal network. Adenylate cyclase–activating polypeptide 1 (ADCYAP1) is known to stimulate the generation of cyclic adenosine monophosphate (cAMP), and all three datasets show widespread expression across the brain regions, with the highest levels in the hypothalamus and amygdala. In the latter region, ISH shows that this gene is expressed by cells located in the cortical amygdaloid nucleus, whereas protein labeling is found in nerve terminals in the central amygdaloid nucleus, which is known to receive an input from the cortical amygdala. The results suggest that this protein is primarily presynaptic and support a role in cAMP-mediated synaptic plasticity. Expression of the orphan G protein–coupled receptor 151 (GPR151) can be visualized both with ISH and IF in neuronal cell bodies in the habenular nucleus of the mouse thalamus. However, the protein staining also shows the projection from the thalamus to the interpeduncular nucleus in the midbrain. This orphan receptor is thus visualized in the neuronal soma, in axons running in the fasciculus retroflexus, and in the presynaptic terminals. This expression pattern can also be shown using three-dimensional imaging of solvent-cleared brain (iDISCO) encompassing a whole mouse hemisphere (Fig. 3B). In addition, this analysis revealed that a portion of the axons pass the interpeduncular nucleus and innervate the parabrachial-pericoerulear region. Movie S1 shows the three-dimensional location of this orphan receptor. Together, these examples illustrate how combining three different approaches for spatial transcriptomics and proteomics results in insights offering detailed information on cellular expression and protein location.

Species comparisons of regional brain expression

To compare the expression profiles in the three mammalian brains, all genes with one-to-one orthologs in human, mouse, and pig were identified, and a total of 12,999 protein-coding genes were analyzed (fig. S19A). Additional genes can be included in the analysis in the future, as additional one-to-one gene orthologs are identified. A combined hierarchical tree including all regions of the three species based on all regionally elevated genes across the 10 main regions is shown in Fig. 4A. The results again support a preserved brain architecture, where the hypothalamus and cerebellum of all three species cluster in proximity to each other. Similarly, the brainstem regions and the cerebrum regions cluster together. Neighboring regions from one species cluster together, but clustering is less tight for corresponding regions from different species. The olfactory bulbs from pig and mouse are clustered tightly together, and the outlier is the olfactory bulb from humans, which shows similarities with the cerebrum regions of humans. This might be due to sampling error, as discussed above, but could also reflect the

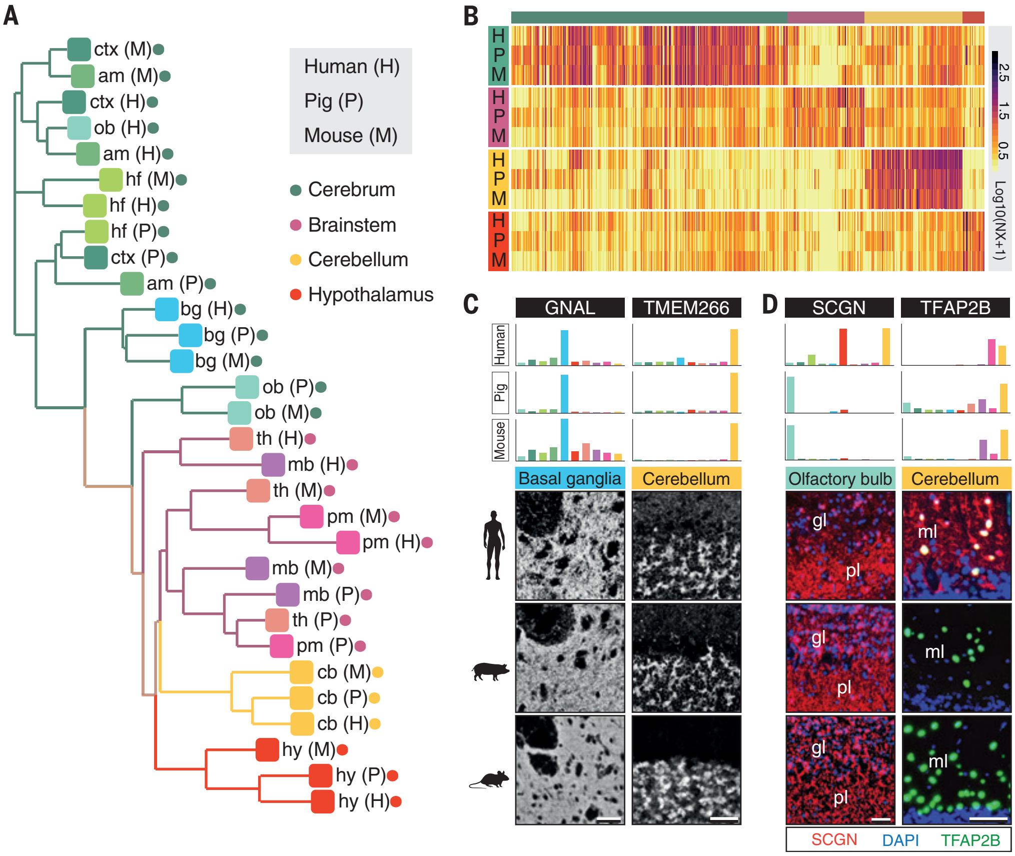


Fig. 4. Species comparison of regional expression in the mammalian brain. (**A**) The expression levels of 1422 genes classified as regionally elevated in either human, pig, or mouse were used for hierarchical clustering analysis, showing the relationship of the 10 main brain regions from the three species. (**B**) A heatmap showing the expression levels in the different brain structures in human (H), pig (P), and mouse (M) brain of enriched genes shared by the three species, based on brain structure comparison shown in Venn diagrams (fig. S19B). The same expression data, but visualized with the 10 regions of the brain, are shown in fig. S20. (**C**) Examples of regionally enriched genes in the mammalian brain and the protein location shown using immunofluorescence. GNAL is enriched in basal ganglia of human and pig brain, and this protein shows highest expression in basal ganglia (neuropil) in all three species. TMEM266 is cerebellum-enriched in all three species and located in synaptic glomeruli of the granular layer. (**D**) SCGN is expressed in the olfactory bulb in all three species with a higher expression in granule cells (gl) in the mouse and pig. In cerebellum, SCGN is only expressed in the molecular layer (ml) of the human and is not detected in pig or mouse cerebellum. In contrast, the transcription factor TFAP2B, coexpressed with SCGN in human, is expressed in the molecular layer of the cerebellum in all three species. External plexiform layer, pl. Scale bars, 50 μm.

more extended pig and rodent olfactory systems, which have olfactory bulbs that are larger than the more rudimentary human olfactory bulbs.

To analyze the differences in elevated genes in the three species, the regions were organized into four main brain structures (cerebrum, brainstem, cerebellum, and hypothalamus) (Fig. 4A), and the molecular features of each brain structure shared by all three species were identified. Genes with the highest expression in the same brain structure in all three species with enriched expression in at least one species were identified, and a list of 537 genes was obtained (fig. S19B). Heatmaps of the expression levels of these 537 brain structure–enriched genes are displayed in Fig. 4B and fig. S20, demonstrating the similarity of expression pattern for these genes across the regions of the brain in the three species. Many known genes are found among these 537 genes with brain structure–enriched expression, including the neuropeptides galanin, oxytocin, and vasopressin (hypothalamus); transcription factors such as T-box brain protein 1 (TBR1), special AT-rich sequence-binding protein 2 (SATB2), and neurogenic differentiation factor 6 (NEUROD6) (cerebrum); and hox genes (brainstem), but less well characterized genes are also found (table S5). The former include the G protein subunit alpha L (GNAL), which is highly expressed in the basal ganglia and known to couple to adenosine A2A and dopamine type 1 receptors (Fig. 4C) (25). The protein staining suggests synaptic location in the caudate nucleus or caudate putamen of all three species. Similarly, cerebellum-enriched transmembrane protein 266 (TMEM266, also known as HVRP1) is selectively detected in the synaptic glomeruli in the granular layer in all three species (Fig. 4C). This detection is in line with the reported role of this postsynaptic protein in the communication between mossy fibers and granule cells (26).

The normalized data also allowed us to identify genes with differential brain expression across the three species. Volcano plots show the overall fold difference in gene expression based on the 10 regions (figs. S21 to S23 and table S6), and these data were combined in scatterplots (figs. S24 to S26) showing species-specific molecular features (one versus two species). Many proteins show similar expression in the three species, as exemplified in Fig. 4D for transcription factor AP-2-beta (TFAP2B) expressed by γ-aminobutyric acid–releasing (GABAergic) interneurons (27), including stellate cells, in all three species. However, many differentially expressed genes associated to specific brain functions could also be identified, such as the low expression of the astrocytic genes glial fibrillary acidic protein (GFAP) and clusterin (CLU) in mouse compared with human and pig. For each of the 10 brain regions, a triangle plot indicates the relative expression of each gene in the three species (fig. S27). As an example, secretagogin (SCGN) is an EF-hand calcium binding protein expressed in the olfactory bulb (28) that is also seen in the stellate cells in the molecular layer of the human cerebellum. This contrasts with pig and mouse, where this protein cannot be detected in the cerebellum (Fig. 4D).

The neurochemical architecture of the mammalian brain

Brain functions are driven by complex circuits composed of different types of neurons with chemical phenotypes adapted to receive and generate signals. To identify species similarities and variations that characterize these types of neurons and their neurotransmitter systems, as well as other classes of cells, we analyzed the distribution of cell identity genes in all three mammalian species. These include (i) transcription factors (n = 1053 genes), which are essential for differentiation and maintenance of cell identities in the postmitotic brain (29); (ii) cell identity genes, including neuropeptides, proteins, and enzymes responsible for the production, transport, and clearance of neurotransmitters ($n = 63$); and (iii) all known neurotransmitter and neuropeptide receptors and receptor subunits ($n = 118$). Comparing the correlation values between species for these protein classes reveals a higher correlation for transcription factors ($P < 0.001$), cell identity genes ($P < 0.001$), and receptors ($P < 0.01$) relative to the gene expression of all other 11,765 genes with their one-to-one ortholog (Fig. 5A).

We found that expression of some transcription factors is highly conserved across the three species, while other transcription factors have a less-maintained expression profile, thus affecting the overall correlation (fig. S28). Examples of some of the transcription factors with conserved distribution across the brain regions are illustrated in Fig. 5B. Homeobox protein 1 (EMX1), known to be expressed by most of the neurons in the cerebral cortex and hippocampus (30), has a similar expression pattern and expression levels in all three species. Class E basic helix-loop-helix protein 22 (BHLHE22, or Bhlhb5), which regulates postmitotic differentiation of cortical neurons (31), has a highly similar expression pattern in the cerebrum regions. In contrast, expression of the transcription factor SIX6 is restricted to the hypothalamus of all three species, which is in line with earlier reports (32). Pituitary homeobox 3 (PITX3) is mainly expressed in the midbrain in all three species, with the highest expression in the substantia nigra in pig, also supporting earlier observations that PITX3 plays a role in the development of dopaminergic neurons of the substantia nigra in mice (33). Homeobox protein Hox-B5 (HOXB5) is specifically expressed in the pons–medulla oblongata region, supporting the observation that Hox genes are expressed in the hindbrain and known to be important for the segmental patterning of this part of the brain. Neurogenic differentiation factor 1 (NEUROD1) is essential for the development of the cerebellum in rats (34). We find that this transcription factor is restricted to the cerebellum in all three species

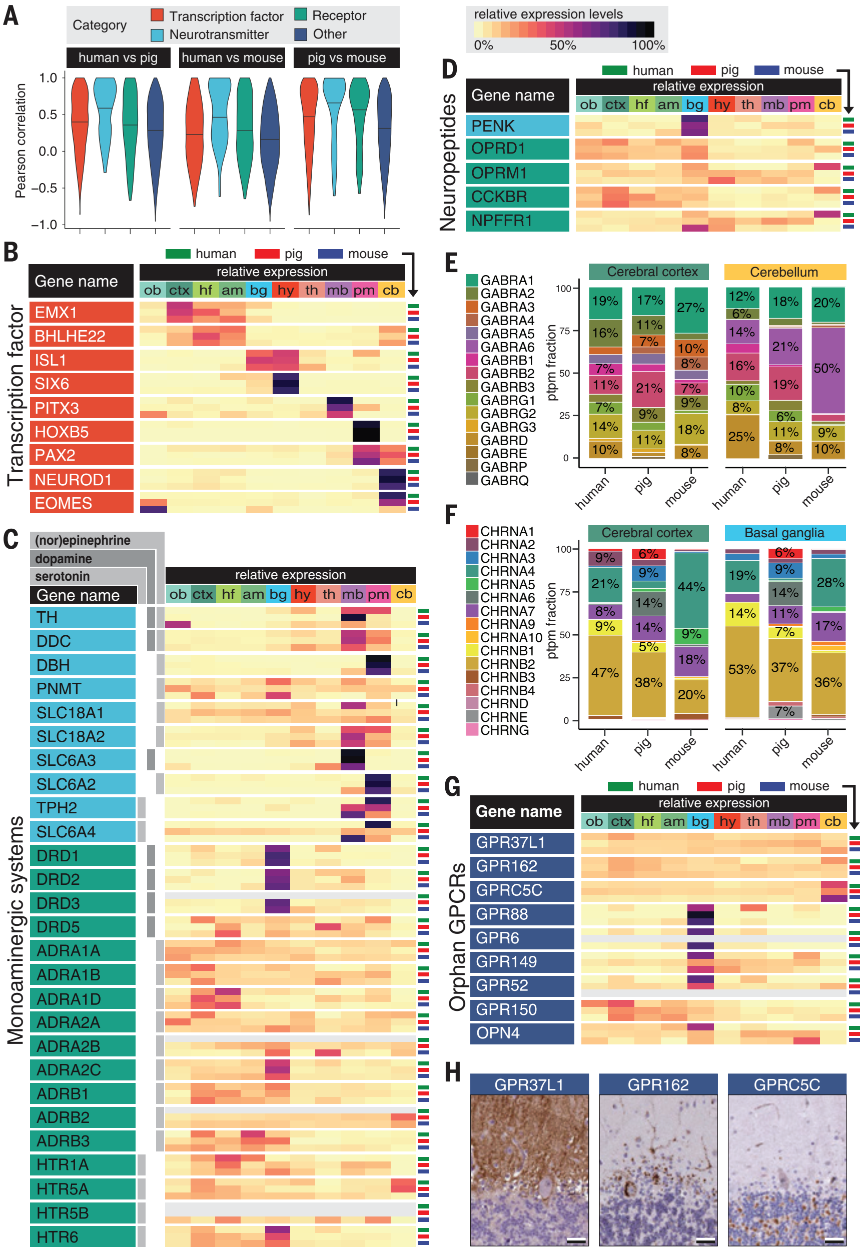


Fig. 5. Expression profiles of cell identity genes in the mammalian brain. (**A**) Overall Pearson correlation between species for transcription factors (red), genes involved in the production and processing of neurotransmitters and neuropeptides (blue), and metabotropic and ionotropic neurotransmitter and neuropeptide receptors (green) in contrast to all other genes (dark blue). (**B**) Examples showing the relative expression in the 10 regions of human, pig, and mouse brains display elevated expression in developmental and anatomical defined regions of the brain. (**C**) Comparing the relative distribution of the monoaminergic systems and (**D**) selected neuropeptide genes reveals a conserved pattern of expression, especially of the enzymes responsible for the production of dopamine (TH, DDC), noradrenaline (+DBH), adrenaline (+PNMT), and serotonin (TPH2) as well as the opioid peptide proenkephalin (PENK). Although many neurotransmitter receptors show a similar distribution profile [DDRs, ADRs, HTR1A, and 5-hydroxytryptamine receptor 6 (HTR6)], several exceptions with clear on/off differences between species could be observed, especially in the cerebellum (HTR5A, OPRM1, OPRD1, CCKBR, and NPFFR1). (**E**) Relative expression of all $GABA_A$ receptor subunits in the cerebral cortex and cerebellum and (**F**) nicotinic receptor subunits in all three species suggests alternative nicotinic subunit composition in different species, especially the pig. (**G**) Four brain-elevated orphan GPCRs (GPR6, GPR52, GPR88, and GPR149) are elevated in the caudate and putamen, whereas GPR150 is group-enriched in the forebrain regions. The nonvisual photoreceptor melanopsin (OPN4) is only expressed in the human basal ganglia. Missing values (gray bars) are due to missing one-to-one orthologs or genes not included in all used datasets. (**H**) The relative distribution of brain-elevated GPCRs with unknown function reveals widespread expression of the orphan GPR37L1 and GPR162, including the cerebellum, whereas GPRC5C expression is elevated in cerebellum. Scale bars, 40 um. Full overviews with NX for all transcription factors, neurotransmitters, and GPCRs are available in figs. S28 to S30.

(Fig. 5B). Notably, NEUROD1 is also expressed in retina of mouse and pig but not human (see gene-specific page of HPA Brain Atlas). An example of a transcription factor with a differential regional expression pattern is eomesodermin (EOMES, or TBR2), which has a high expression level in the human cerebellum but is expressed mainly in the olfactory bulb in the mouse brain. In pig, however, this transcription factor is expressed at equal levels in the cerebellum and olfactory bulb. Overall, these results show an evolutionarily preserved distribution and regulatory role of some of the transcription factors, which likely provide a foundation for basic brain architecture during evolution. However, these data also reveal substantial species variation in the distribution of transcription factors, including many uncharacterized transcription factors not yet linked to a cell type.

The neurotransmitters

The expression patterns of the proteins involved in brain signaling were analyzed in the various regions of the human, pig, and mouse brain, revealing a strong correlation between all species for genes essential for the production and transport of neurotransmitters and peptides (Fig. 5A). This confirms the known similarities between species with regard to the distribution of cell types and functions in the brain. In Fig. 5C, the relative distribution of molecular components of the neuromodulatory monoaminergic systems is shown, consisting of the (nor)adrenergic, dopaminergic, and serotonergic systems. In general, the enzymes responsible for the production of these neurotransmitters are distributed similarly in the three species, indicating a conserved organization in the mammalian brain, although there are some exceptions that might have neuropharmacological implications. For example, lysergic acid diethylamide (LSD) has a high affinity for several serotonin [5-hydroxytryptamine (5-HT)] receptors ([35]). Rodents have genes coding for serotonin receptors 5A and 5B (HTR5A and HTR5B) ([36]), whereas humans and pigs lost the gene for HTR5B during evolution. In mouse, HTR5A and HTR5B are expressed throughout the brain, with lowest expression in the cerebellum. Human and pig show a similar distribution to each other, with the highest expression of HTR5A in the cerebellum (Fig. 5C).

Another example of species differences is tyrosine hydroxylase (TH), the rate-limiting enzyme in the synthesis of the catecholamines dopamine, noradrenaline, and adrenaline. TH is expressed in the mouse olfactory bulb in a substantial number of mainly periglomerular, often GABAergic neurons, but this transcript is under the detection cutoff in human and pig. The human olfactory bulb does contain TH-positive neurons ([37]) and, as mentioned previously, the human results on the bulb may be compromised by dissection problems. Notably, the transcription factor PITX3 (Fig. 5B), known to bind with high affinity to the TH promotor ([38]), shows a similar species variation. The results also show a surprisingly strong and wide distribution of phenylethanolamine *N*-methyltransferase (PNMT) (Fig. 5C), the enzyme converting noradrenaline to adrenaline. As expected, PNMT is present in the pons and medulla oblongata of all three species ([39], [40]), but unexpectedly high expression is found in mouse and human basal ganglia, in mouse cortex, and in pig olfactory bulb. A transient and wide expression of PNMT transcript and protein has been observed in the rat brain, but only during the first postnatal month ([39]). A pronounced species difference is also seen for the expression of the sodium-dependent serotonin transporter (SLC6A4) expressed in the brain regions containing serotonergic neurons in mouse and human, while the pig has more widespread expression (Fig. 5C). It has earlier been reported ([41]) that the expression of this transporter in the human and mouse developing brain is widespread before its expression is restricted to serotonergic neurons in the adult brain. Thus, our data suggest that SLC6A4 retains its developmental distribution in adult pigs.

The analysis of the genes coding for neuropeptide systems also revealed similarities and differences in expression between the species; for example, in the opioid system there is a highly conserved expression of the ligand proenkephalin (PENK) combined with a dissimilar regional expression of the delta (OPRD1) and mu (OPRM1) opioid receptors, both G protein–coupled receptors (GPCRs) (Fig. 5D). Our data support earlier observations that OPRM1 is expressed in the human cerebellum ([42]) but not in the mouse cerebellum ([43]), and here we show that this receptor cannot be detected in the cerebellum of pig. Instead, the expression profiles show the presence of OPRD1 in the pig cerebellum. The neuropeptide receptor NPFFR1 is here shown to be predominantly expressed in the human cerebellum, although this receptor binds the opioid-modulating peptide NPFF expressed in the cerebellum of all three species. Similarly, the gastrin/cholecystokinin type B receptor (CCKBR) is expressed in the human cerebellum, whereas the transcripts for the ligands (cholecystokinin and/or gastrin) could not be detected in this part of the brain. The ligands may have an extracerebellar origin.

We subsequently analyzed the genes coding for the γ-aminobutyric acid type A ($GABA_A$) and nicotinic receptor subunits (nACHRs) (Fig. 5, E and F). Both receptor types can form ligand-gated ion channels with different physiological properties by variations in the subunit compositions. Many of the studies related to these receptors have been performed in mice and rats, and it is therefore also relevant to compare the expression of the various $GABA_A$ and nACHRs in humans and pigs. Variation in subunit composition between brain regions has been reported ([44]), and it is known that the α6 (GABRA6) subunit, which has the highest potency for GABA, is only expressed in the cerebellum, unlike receptors containing the α1 and α2 subunits, which are more widespread throughout the brain. Carriers of a variant allele of the *GABRA6* gene have an increased risk for suicide ([45]), suggesting that the cerebellum might be implicated in mental disorders.

Comparing the expression of the GABA receptors in the three species suggests a conserved subunit composition, as exemplified by the subunit distribution pattern in cerebral cortex and cerebellum (Fig. 5E). In contrast, comparing the relative and absolute expression of nACHR subunits reveals a high degree of phylogenetic differences between mouse, human, and pig (Fig. 5F). For example, pigs express low levels of neuronal acetylcholine receptor subunit α4 (CHRNA4) but higher levels of subunits α1 (CHRNA1), α3 (CHRNA3), and α6 (CHRNA6) in the cerebral cortex and basal ganglia compared with mouse and human. Furthermore, both human and pig express higher levels of the β1 subunit (CHRNB1) compared with mice, and pigs express the ε subunit (CHRNE) in the basal ganglia only. The β1, α1, and ε subunits are all considered to be exclusive mammalian muscular nicotinic receptor subunits. The results presented here therefore suggest alternative nicotinic receptor composition in different mammalian species and the possibility of "muscular" nicotinic receptor subunits involved in the formation of functional nicotinic receptors in the central nervous system (CNS). These examples confirm the shared basic architecture of the mammalian brain with regard to cell types, neurotransmitter systems, and physiological functions. However, we identified examples of clear species variation in the expression levels and distribution of receptors that suggest an important role for receptors in brain evolution and adaptation. This reinforces the need for caution when comparing the function of, for examples, serotonergic, opioid, and cholinergic receptors on the basis of animal experiments using rodents without considering the expression pattern of the ortholog receptor in humans. This is particularly important in the context of drug development, given that at least 30% of today's prescribed drugs act via GPCRs ([46]).

GPCRs with unknown functions

We next analyzed the expression profiles of all GPCRs to explore differences and similarities in expression pattern across the brain regions in the three mammalian species (fig. S30).

These GPCRs include many olfactory receptors with highest mRNA levels in the olfactory bulb, possibly located in the axons of the olfactory receptor cells. In addition to the GPCRs associated with olfaction and neurotransmission, we identified several human orphan GPCRs with brain-elevated expression (n = 30 genes). These include the GPR37L1 and GPR162 expressed in many brain regions (Fig. 5G), such as the cerebellum (Fig. 5H), but also several brain-elevated orphan GPCRs with regionally elevated expression, including GPRC5C, mainly expressed in the cerebellum in all three species. We confirm the exclusive expression of GPR88 in the basal ganglia (*47*) of human, pig, and mouse, but we were also able to identify four brain orphan GPCRs with elevated expression in the basal ganglia (Fig. 5G). The nonvisual photoreceptor melanopsin (OPN4) shows high expression in the human basal ganglia but very low expression in the brains of pig and mouse. OPN4 is expressed by ganglion cells in the retina and plays a role in the regulation of circadian rhythms (*48*), although the function of this photoreceptor deep in the "dark" core of the brain is unclear. It was recently reported that melanopsin also acts as a thermoreceptor mediating heat-activated expression of clock genes (*49*). Our data, based on the expression pattern found here, suggest a possible temperature-sensing role of OPN4 in the human basal ganglia that is not shared with mice or pigs.

Whole-body versus brain regional tissue specificity classification

The data presented here have made it possible to compare the brain enrichment of all genes with the whole-body tissue specificity using the Tissue Atlas resource (*1*). The NX data across all samples were used to classify all protein-coding genes according to organ and tissue expression, where the brain was classified as a single organ, and 36 additional organs and tissues were scored across the human body. These tissue types include, for example, liver, pancreas, intestine, lung, reproductive organs, and lymphoid tissues, as well as a group of cell types summarized as "blood," including 18 single blood cell types and peripheral blood mononuclear cells (*50*). For the brain, the maximum NX value for a given gene in one of the brain regions was used as the brain expression value. We previously reported 1113 genes with elevated expression in the brain on the basis of the comparison of the cerebral cortex with 26 peripheral tissue types (*1*, *51*). Here, we analyzed many more brain regions as well as spinal cord and corpus callosum, and we identified many more genes (n = 2587) with elevated expression in at least one region of the brain compared with peripheral tissues (fig. S31). In addition, 5298 genes were found to be expressed in the brain but had elevated expression levels in a peripheral organ. A total of 8342 genes showed low tissue specificity across all 37 tissues and organs (fig. S31 and table S7). Only 33 genes could be specifically defined as enriched in the brain and not detected in any of the peripheral tissues. Many of these "specific" genes were transcription factors, such as neurogenic differentiaton factors 2 and 6 (NEUROD2 and NEUROD6), BarH-like 1 homeobox protein (BARHL1), and GPCRs such as GPR101 and GPR26.

We analyzed the expression levels of the 2587 human genes classified as brain-elevated across all analyzed human peripheral tissues (fig. S32A). The analysis demonstrates two major clusters: the first with relatively restricted expression across the peripheral tissues, and the second containing genes with a more tissue-wide expression. An analysis of the first cluster shows smaller and more specific expression clusters, such as a subcluster of genes with expression in testis and fallopian tube, in addition to the brain. Most of these genes encode proteins specifically expressed in ciliated cells, including ependymal cells lining the ventricular wall in the brain (fig. S32B). Other notable subclusters harbor genes with elevated expression in the brain but also high expression in peripheral tissues, such as cardiac and skeletal muscles (fig. S32C) and liver (fig. S32D). The large cluster of genes with elevated expression in the brain, but with a more general tissue-wide expression pattern, includes a small cluster of genes encoding immune tissue–associated proteins (fig. S32, E and F).

The relationship between the whole-body specificity and the regional brain specificity was then analyzed for all protein-coding genes. In Fig. 6, all genes detected in any of the human tissues and organs are included with a gene-to-gene comparison to the regional brain specificity. Only 520 genes that are classified as brain-elevated have regional brain specificity expression, while a large fraction of the brain-elevated genes (n = 1776) have low regional specificity within the brain. The latter include (i) several well-known astrocyte markers, including GFAP and aquaporin 4 (AQP4); (ii) oligodendrocyte genes involved in myelination, including myelin basic protein (MBP) and proteolipid protein 1 (PLP1); and (iii) pan-neuronal genes expressed by most neurons, for example, sodium/potassium-transporting adenosine triphosphatase subunit alpha-3 (ATP1A3). The 520 genes classified as both regionally and brain-elevated include genes known to be expressed by different neuronal populations and genes involved in inter- and intracellular signaling cascades, such as (i) receptors, e.g., adenosine receptor 2A (ADORA2A) enriched in the basal ganglia; (ii) ion channels, e.g., calcium voltage-gated channel auxiliary subunit gamma 3 (CALCNG3) elevated in regions of the cerebrum; and (iii) components of intracellular signaling pathways, such as GTPases, e.g., Rho guanine nucleotide exchange factor 33 (ARHGEF33) enriched in the cerebellum.

Many of the genes that have regional brain specificity expression are not brain-elevated from a whole-body perspective but instead have elevated expression in one or a group of peripheral tissue types. For example, ankyrin-1 (ANK1), with expression enriched in skeletal muscle and tongue, is selectively expressed in the cerebellum and, on the protein level, associated with the membrane of Purkinje cells (Fig. 6). However, most of the genes classified as elevated in tissue types other than brain are classified as having low regional specificity within the brain, such as a number of proteins detected in astrocytes or oligodendrocytes. These proteins include crystallin alpha B (CRYAB) and aldehyde dehydrogenase 6 family member A1 (ALDH6A1), as well as many of the microglia proteins, such as the well-characterized allograft inflammatory factor 1 (AIF1) and the neuropil-associated protein regulatory factor X2 (RFX2) elevated in testis. The group of genes (n = 8027) classified as both low tissue specific and low regional specific in the brain include many housekeeping proteins but also proteins with a more selective location to certain cell types, such as Acyl–coenzyme A (CoA) synthetase long chain family member 4 (ACSL4) mainly detected in neuronal cell bodies, and A-kinase anchoring protein 17A (AKAP17A) detected in the nucleus of glial cells and neurons in the cerebellar granular layer (Fig. 6).

Global and regional expression landscape of cortical cell type signature genes

A large number of differentially expressed genes have previously been identified using various approaches, including single-cell genomics and coexpression analysis. Here, we have analyzed the whole-body expression pattern of a consensus set of signature genes for cortical neurons, astrocytes, oligodendrocytes, and microglia using an immunopanning approach (*8*) and a coexpression analysis of publicly available expression data (*12*). From these two datasets, 420 genes were identified as putative human cerebral cortex cell type signature genes (listed in table S9), with 196 neuron-specific, 180 astrocyte-specific, 65 oligodendrocyte-specific, and 51 microglia-specific genes. Analyzing the expression variance of these cell type signature genes across the different regions of the brain showed multiple outliers differentially expressed in different parts of the brain for both neuronal and astrocyte genes (Fig. 7A). Oligodendrocytes and microglia signature genes are less variable across the different brain regions. A large fraction of neuronal and oligodendrocyte genes are classified as brain-enriched (Fig. 7B and

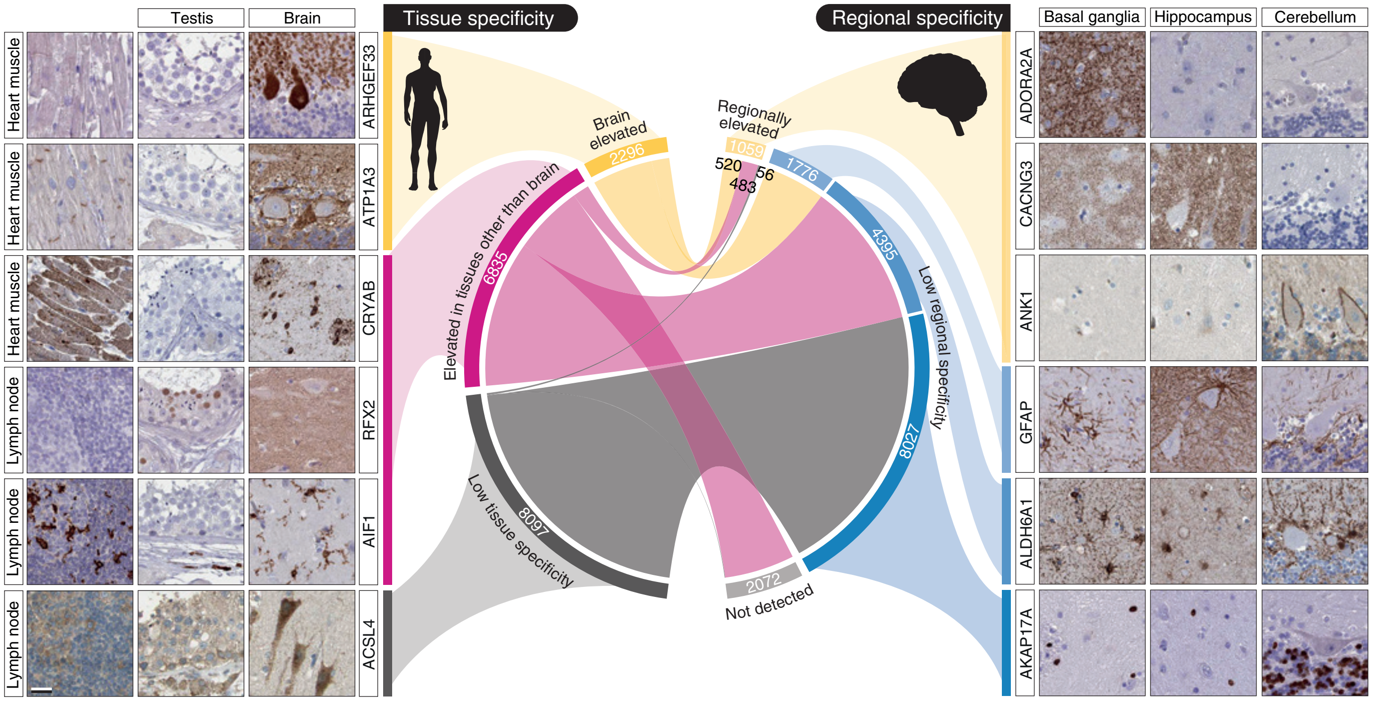


Fig. 6. The regional expression in brain compared with whole-body expression. Gene classification based on tissue specificity using transcript expression data in 37 different tissue types enables separation of genes into brain-elevated, elevated in tissues other than brain, and low tissue specificity. This is compared to the classification based on regional expression in the brain. Of 1059 genes, 520 with regionally elevated expression were also classified as brain-elevated. Genes classified as elevated in tissues other than the brain are often detected in brain but expressed with low regional specificity. Rho guanine nucleotide exchange factor 33 (ARHGEF33) is brain-enriched, including cerebellum-enriched, and here detected in Purkinje cells (HPA041051). ATPase Na^+/K^+ transporting subunit alpha 3 (ATP1A3) is group-enriched in brain and heart muscle and detected in all brain regions with low regional specificity. Immunohistochemistry using HPA056446 displayed the intercalated discs in heart muscle and had a synaptic location in brain. Crystallin alpha B (CRYAB) is tissue-enhanced in striated muscle, found in all brain regions with low tissue specificity, and detected in oligodendrocytes (HPA057100). Regulatory factor X2 (RFX2) is elevated in testis and detected in all brain regions with low regional specificity, and with a nuclear localization in testis and neuropil in brain (HPA048969). AIF1 is enhanced in blood and lymphoid tissues, while also detected in microglia in all brain regions with low regional specificity (HPA049234). ACSL4 is classified as low specificity both in tissues as well as brain regions (HPA005552). ADORA2A is group-enriched in lymphoid tissues and brain, including basal ganglia–enriched (HPA075997). Calcium voltage-gated channel auxiliary subunit gamma 3 (CACNG3) is brain-enriched and group-enriched in cerebrum regions but was not detected in cerebellum, also verified at the protein level (HPA077238). ANK1 is group-enriched in skeletal muscle and tongue as well as cerebellum-enhanced in brain, where the protein is selectively associated with the Purkinje cell membrane (HPA004842). GFAP is brain-enriched and detected in all brain regions with low regional specificity (HPA056030). ALDH6A1 is group-enriched in kidney and liver, detected in all brain regions with low regional specificity, and, as GFAP, localized to astrocytes (HPA029074). AKAP17A is expressed in all tissue types with low tissue specificity as well as low regional specificity in the brain; the protein is detected in subsets of cell nuclei and in brain in glial cells as well as granular cells in cerebellum (HPA043247). Scale bar, 25 μm.

fig. S35), while many astrocyte and microglia genes are classified as elevated in other tissues. In fact, several astrocyte signature genes are highly expressed in liver and/or muscle tissue, whereas many microglia signature genes are enriched in lymphoid tissue, bone marrow, and/or blood. In summary, the global analysis reveals that most of the putative astrocyte and microglia signature genes are highly expressed in selective peripheral tissues, often exceeding the expression levels in the brain.

We superimposed the putative cell type signature genes on the global tissue expression landscape using the data from this study (Fig. 7C). Expression of only 180 of the human cerebral cortex cell type signature genes (43%) was classified as brain-elevated with regard to expression, and 158 genes (38%) were classified as elevated in expression in nonbrain tissues and the expression of the remaining genes (n = 82) classified as low tissue specific. To further explore this lack of "brain specificity" of many of the putative brain signature genes, we analyzed some of these genes further, as shown in Fig. 7D. The microglia signature genes arachidonate 5-lipoxygenase (ALOX5) and integrin subunit beta 2 (ITGB2) both showed elevated expression in blood (lymphoid tissues), while other microglia signature genes showed elevated expression in specific blood cells. For example, the gene for PYD and CARD domain containing protein (PYCARD) is expressed by granulocytes, monocytes, and dendritic cells. These results confirm the notion of shared origin and functions between microglia and immune cells. In contrast, several astrocyte signature genes are classified as genes with elevated expression in liver or muscle tissue. Out of the 19 genes with liver-elevated expression, six are transport-related genes, including solute carrier family 13 member 5 (SLC13A5), detected in the membrane of hepatocytes and end feet of astrocytes in brain, and nine genes code for metabolic enzymes such as aldehyde dehydrogenase 1 family member L1 (ALDH1L1), detected in cytoplasm of hepatocytes and astrocytes. Genes classified as having muscle-elevated expression include several proteins with structural function, such as syntrophin alpha 1 (SNTA1), detected both in astrocytes and skeletal muscle. All three are thus classified as showing elevated expression in tissues other than brain on the tissue level. These results highlight the shared function of astrocytes

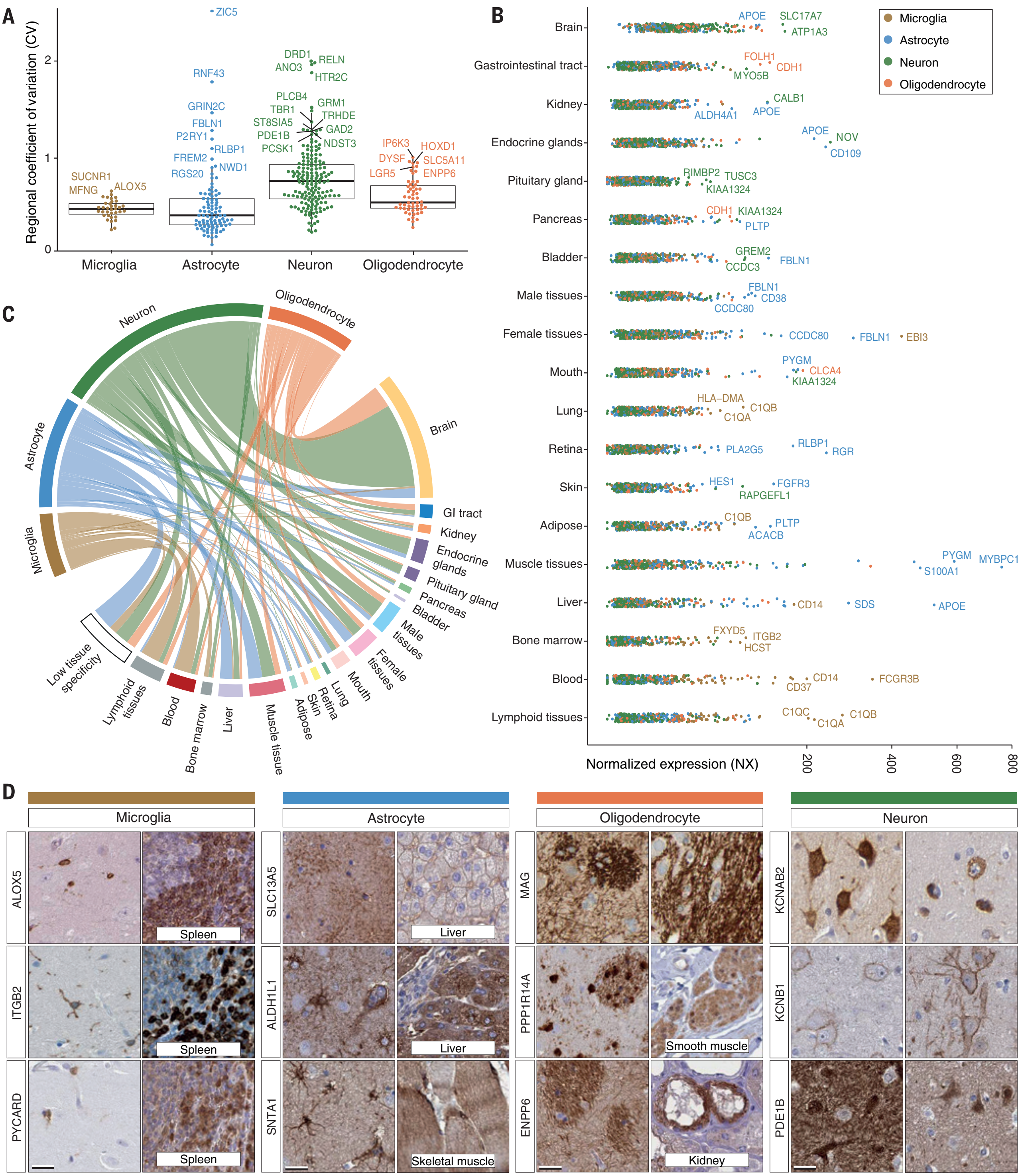


Fig. 7. The gene expression landscape of 420 putative brain cell type signature genes in cerebral cortex. (**A**) The expression variation for the 420 signature genes in the different brain regions was analyzed on the basis of the coefficient of variation (CV) of NX values across the 10 regions (see figs. S33 and S34 for expression heatmap and scatterplot across the 10 regions). (**B**) The expression levels for the individual signature genes in the different tissue types. The most abundant genes are indicated by gene names (see fig. S32 for expression heatmap of the signature genes across all tissues). (**C**) Chord diagram shows the tissue specificity of brain cell type signature genes. Each link represents the number of cell type signature genes that are elevated at a tissue level in the analyzed organ and tissue types. (**D**) Examples of signature genes and their protein localization in the different cell types. Microglia signature genes (ALOX5, ITGB2, and PYCARD) are highly expressed in spleen. Several astrocyte signature genes, including SLC13A5 and ALDH1L1, have elevated expression in liver, whereas SNTA1 is highly expressed in skeletal muscle. Neuronal examples (KCNAB2, KCNB1, and PDE1B) are located in neuronal cell bodies in the brain, while oligodendrocyte proteins, including MAG, are detected in white matter of the brain as well as in smooth muscle (PPP1R14A) and renal tubule of kidney (ENPP6). More detailed expression profiles for all cell signature genes are shown in figs. S33 to S35. Scale bars, 25 μm.

and hepatocytes in the metabolism of secreted substrates.

We subsequently analyzed the signature genes for oligodendrocytes. Expression of some putative oligodendrocyte signature genes is here classified to have low tissue specificity, such as protein phosphatase 1 regulatory inhibitor subunit 14A (PPP1R14A). Others are classified as having elevated expression in peripheral tissues, such as ectonucleotide pyrophosphatase 6 (ENPP6), which is elevated in kidney. The expression of putative neuronal signature genes is here mainly classified as brain-elevated, such as potassium voltage-gated channel subfamily A regulatory beta subunit 2 (KCNAB2), potassium voltage-gated channel subfamily B member 1 (KCNB1), and phosphodiesterase 1B (PDE1B), all detected on the protein level, using antibody-based localization, in a subset of neurons and with different subcellular locations (Fig. 7D). However, several neuronal signature genes also showed elevated expression in endocrine tissues, the pituitary gland, or male tissues. The analysis shows that caution should be taken regarding genes identified as signature genes, because the identification of these is context-dependent, and many of the genes previously identified as signature genes for specific cell types in the brain are in fact also highly expressed in peripheral tissues.

The HPA Brain Atlas

As part of this work, a brain atlas database has been launched to present and integrate all the data reported here, and this is an extension of the HPA portal, with a brain-centric summary page for each gene with expression data in human, pig, and mouse brain regions. The resource is presented (*17*) with an expression summary for all protein-coding genes. For selected genes, the distribution of the corresponding protein is visualized by antibody-based protein detection. Gene expression in the CNS is visually summarized on the basis of the 10 brain regions as well as spinal cord, corpus callosum, retina, and pituitary gland, with underlying data for many more subregions (21 subregions in human, 27 in pig, and 16 in mouse). Each of the 10 regions can also be reviewed in individual pages, which provide a classification overview, interactive lists, and figures, as well as highlighted examples of regionally specialized cells and proteins. A selection of 815 proteins is shown at the protein level in human tissues, and 271 genes include a complete mouse brain profile through a high-resolution virtual microscope. These 271 genes were analyzed with immunofluorescence-based imaging and include protein expression levels for 120 subregions of the brain.

Outlook

The expression profiles for the protein-coding genes in all major brain regions have been determined to capture the complexity of the cellular organization of the brain and to enable comparison between species. The integration of data from several sources has allowed us to combine data from transcriptomics, single-cell genomics, in situ hybridization, and antibody-based protein profiling. The rapid technological improvements in the field of spatial transcriptomics and single-cell genomics will in the future allow for an even higher degree of molecular granularity. The analysis presented here, relying on anatomical dissection of the different regions of the brain, allowed us to classify all of the individual protein-coding genes on a genome-wide level, where each gene is scored for its regional distribution. The resource provides detailed molecular transcriptomics maps of the mouse, pig, and human brains, and these maps are combined with immunofluorescence-based imaging of single cells using antibodies toward proteins identified as being of neurological and neuropsychiatric interest. In this manner, genes could be identified that are shown to be differentially expressed between organs and within the brain. By including more brain regions, the number of transcripts detected in the human brain has increased compared with previous studies. The number of regionally elevated genes in all three species is relatively small, with ~1000 genes identified in each species with an elevated expression (regional enriched, group enriched, or regional enhanced) across the 10 brain regions.

Analysis of the regionally elevated genes in the three species presented here supports the concept of a similar basic molecular brain architecture during mammalian evolution. The genes involved in production, vesicular transport, uptake, and degradation of the main neurotransmitter systems show overall high similarity among the three species, although notable differences have been identified. Thus, for example, some of the catecholamine-synthesizing enzymes show distinct species differences with regard to localization and expression levels, and several metabotropic and ionotropic receptors also exhibit species differences. Many neurotransmitter receptors, in particular the nicotinic and opioid receptors, show high variability in the different species, in particular between human and mouse. These types of gene differences between species highlight the fact that mouse models may not provide data that can be used to understand and treat human mental disorders. For some of the brain regions, such as cerebellum and hypothalamus, the global expression profile of pig is closer to that of human, suggesting that pig might be an attractive animal model to study many neurological and mental processes.

Many "signature genes" identified previously for specific brain cell types (such as astrocytes, microglia, oligodendrocytes, and neurons) are expressed at higher levels in peripheral organs, demonstrating that caution should be taken when using such genes as markers of specific brain cell types. In fact, our results support a view of shared functions between microglia and immune cells, with many genes elevated in both types of cells. Similarly, many genes previously identified as signature genes for astrocytes have a functional role in transport, and the elevated expression of these genes in astrocytes is often shared with liver or skeletal muscle. Cerebellum stands out with regard to the number of regional enriched genes and genes differentially expressed between species. This is also the brain region with the most distinct pattern of active cis regulatory elements compared with cortical and subcortical structures, and the cerebellum also has the highest degree of alteration within predicted enhancers among primates (*52*). Several genes suggested to be involved in neuropsychiatric diseases are found to be selectively expressed in the human cerebellum, which might be surprising for a brain region traditionally linked to fine-tuning motor behaviors. However, these data support the emerging notion that this part of the brain is associated with many neurological and psychiatric conditions.

We describe an integrative approach for mapping the molecular profiles in human, pig, and mouse brain that generates a detailed multilevel view on the protein-coding genes of the mammalian brain. We also compare the regional differences of the human brain with a genome-wide, whole-body tissue-specificity classification. An open-access Human Brain Atlas knowledge-based resource is presented as part of the HPA to allow the exploration of individual genes and classes of genes and their expression profiles in the various parts of the mammalian brain as well as all other major parts of the human body.

Material and methods

Animal procedures

The animal experiments conformed to the European Communities Council Directive (86/609/EEC), and all efforts were made to minimize the suffering and the number of animals used. Mouse brain tissue samples used for transcriptomic and proteomic analyses were collected and handled in accordance with Swedish laws and regulations, and all experiments were approved by the local ethical committee (Stockholms Norra Djurförsöksetiska Nämd N183/14). The experimental minipigs (Chinese Bama Minipig) were provided by the Peral Lab Animal Sci & Tech Co., Ltd (Permit number SYXK2017-0123). Brain tissue samples used for analysis were collected and handled in accordance with national guidance for large experimental animals and under permission of the local ethical committee (ethical permission numbers 44410500000078 and BGI-IRB18135) and experiments were conducted in line with European directives and regulations.

Wild-type male (n = 2) and female (n = 2) C57BL/6J mice (2 months old) were obtained from Charles River Laboratories and maintained under standard conditions on a 12-hour day/night cycle, with water and food ad libitum. Mice were deeply anesthetized and transcardially perfused with 0.9% saline solution. Brains were quickly removed from the skull and dissected on a glass plate on ice. The entire brain was carefully dissected into 16 subregions, and corpus callosum, pituitary gland, and retina were also collected. A complete list of samples and subregions is provided in table S2. Tissue samples were collected into tubes, snap frozen in dry ice, and stored at −80°C until further processing.

For immunofluorescence and iDISCO analysis, mice were anesthetized and transcardially perfused using balanced Tyrode's solution followed by fixation with modified Zamboni fixative (4% paraformaldehyde, 0.2% picric acid in 0.1 M phosphate buffer). For cryosectioning, brains were post-fixed for 90 min and transferred to PBS containing 30% sucrose and 0.1% sodium azide. After cryopreservation, brains were snap frozen using CO_2 and 16-μm-thick coronal sections were cut on a cryostat (Leica, CM1950) and thaw-mounted on SuperFrost Plus glass slides (VWR). For iDISCO experiments, samples were placed in PBS containing 0.1% sodium azide until further processing.

Male (n = 2) and female (n = 2) Chinese Bama minipigs (1 year old), were obtained from the Pearl Lab Animal Sci & Tech Co., Ltd. All animals were housed in a specific pathogen-free stable facility under standard conditions. Pigs were deeply anesthetized and slaughtered by terminal bleeding. The entire pig brain was quickly removed from the scull and submerged into ice-cold PBS buffer for 2 min to remove excess blood and stiffen the tissue. The brain was cut in coronal slabs at the level of (i) frontal lobe/olfactory tract, (ii) optic chiasm, and (iii) between hypothalamus and cerebral peduncle. Slaps were divided in two hemispheres, exposing all main brain structures. Sample blocks of one hemisphere were immersion-fixed in phosphate-buffered saline containing 4% paraformaldehyde for 1 week followed by storage in phosphate-buffered saline containing 0.1% sodium azide at 4°C. For mRNA analysis, pieces of cerebral cortex and cerebellum were collected on the basis of a sampling strategy collecting a representative sample containing all cell layers. All other regions were dissected and collected in their entirety, subregional samples are listed in table S3. Two samples (somatosensory cortex and periaqueductal gray) are missing from female 1, as these two regions could not be identified with 100% certainty and thus were excluded. Duplicate samples were taken from olfactory bulb from female 2, resulting in 119 brain samples and an additional 8 samples (retina and pituitary gland), for a total of 127 samples. All samples were stored at −80°C until RNA extraction took place, within one month.

For immunofluorescence analysis, samples were immersed in 70% ethanol before dehydration in absolute alcohol and xylene before paraffin embedding. Sections were cut (4 μm by Microm HM 355S, Thermo Fisher Scientific) and placed on SuperFrost Plus glass slides (VWR), baked, and then used for staining or stored in −20°C until stained.

RNA sequencing of pig and mouse brain samples

For mouse brain RNA extraction, the tissue was homogenized mechanically using a TissueLyser LT (Qiagen) and total RNA was prepared using the RNeasy Mini isolation kit (Qiagen) for each of the 19 samples. This generated high-quality RNA, with 84% of the samples having RNA integrity values >8.0, with only one sample removed owing to very low RIN value (<6.0). RNA integrity (RIN) was assessed using Agilent RNA 6000 Nano Kit (Agilent Technologies). In total, 75 samples were subsequently used for library construction with Illumina TruSeq Stranded mRNA reagents. The Illumina HiSeq2500 platform was used for sequencing at ~20 million reads depth. Detailed information about the samples and sequencing quality control is listed in table S11. The output analysis was performed using Kallisto v.0.43.1 and mapped to the mouse Ensembl v92 with 22,333 protein-coding genes, for the initial analysis. Human and mouse orthologs were defined as a one-to-one translation, resulting in a total of 15,160 genes.

For pig brain RNA extraction, the tissue was homogenized mechanically using a Dounce tissue grinder in liquid nitrogen. Total RNA was then extracted with a standardized protocol based on TRIzol reagent (Invitrogen). First, total mRNA and noncoding RNAs were enriched by removing ribosomal RNA (rRNA) using a MGIEasy rRNA depletion kit (MGI Tech, China). Enriched RNAs were then mixed with RNA fragmentation buffer resulting in short fragments (180 to 300 base pairs). Third, complementary DNA (cDNA) was synthesized from the fragmentated RNAs using N6 random primers, followed by end repair and ligation to BGISEQ sequencer compatible adapters. The quality and quantity of the cDNA libraries were assessed using Agilent 2100 BioAnalyzer (Agilent Technologies). Finally, the libraries were sequenced on the BGISEQ-500 with 100 paired-end read (PE100). A few randomly selected libraries were also resequenced and co-validated with MGI2000 sequencer. An average of 200 million reads were generated for each library. Sequencing reads that contained adapters and/or had low quality, aligned to rRNA were filtered before following bioinformatic analysis. An overview of the total reads, Q30 clean reads, and mapping ratio to the pig genome (Sscrofa11.1) is provided in table S12. More than 94% of the samples have <10% rRNA of total reads, indicating a highly efficient rRNA removal and RNA quality. One sample (pituitary gland from female 2) was excluded from final data analyzed because of high rRNA inclusion (table S12). The output analysis was performed using Kallisto v.0.43.1 and mapped to the pig Ensembl build 92 with 22,342 protein-coding genes, for the initial analysis. Human and pig orthologs were defined as a one-to-one translation, resulting in a total of 14,656 genes.

Human sequencing datasets

The Functional Annotation of Mammalian Genomes 5 (FANTOM5) project ([19]) provides transcriptomic profiles and functional annotation of mammalian cell types using cap analysis of gene expression (CAGE) ([53]), a method developed at RIKEN that is based on several full-length cDNA technologies. Expression data files with CAGE peaks and ontology for 77 samples (representing 30 different tissue types) were obtained from the version 4 FANTOM5 repository (https://fantom.gsc.riken.jp/5/datafiles/reprocessed/), which we mapped to Ensembl for calculation of the normalized tags per million for each gene. The Genotype-Tissue Expression (GTEx) ([18]) is an extensive project that has collected and analyzed thousands of human postmortem tissue samples. RNA-seq data from 26 tissue types (including more than 8000 patient samples) were mapped using RSEMv1.2.22 (v7,GTEx_Analysis_2016-01-15_v7_RSEMv1.2.22_transcript.tpm.txt.gz) and generated transcript per million (TPM) values that are included in the Human Protein Atlas. The in-house RNA-seq analysis on human tissue types includes 172 tissue samples covering 33 of the 37 tissue types representing the whole human body. The detailed protocol used for RNA-sequencing in the HPA has been described previously ([1], [51]).

Normalization of human data

To enable expression classification and mapping of all human protein-coding genes across all tissue types and samples, TPM expression values were obtained by mapping processed human reads to the human reference genome GRCh37/hg19 based on Ensembl build 92 ([54]) gene models using Kallisto (v.0.43.1) ([55]). Next, the gene expression levels were calculated by summing up the TPM values of all alternatively spliced protein coding transcripts for the corresponding gene for a total of 19,670 protein-coding genes. The average TPM value of all individual samples for each tissue, brain region, or cell type was used to estimate the gene expression level. Data analysis and visualization were performed using R (version

3.5.1 Feather Spray) (56). To allow the three datasets (HPA, GTEx, and FANTOM) to be combined (1, 18, 19), a pipeline was set up to normalize the data for all samples (fig. S4). In brief, we first scaled all TPM values per sample so that the sum was one million, to compensate for the noncoding transcripts that had been previously removed and to obtain pTPM values per sample. Next, all TPM values were TMM normalized (22) between all the samples in each data source (HPA tissues, HPA blood cells, GTEx, and FANTOM5, respectively), then each gene was Pareto scaled (23) within each data source. Tissue data from multiple sources were integrated using batch correction implemented as removeBatchEffect in the R package *limma* (24) using the data source as a batch parameter. The resulting transcript expression values, here called normalized expression (NX), are calculated for each gene in every sample. In the Human Protein Atlas, the NX value for every gene in every sample is calculated and visualized on the gene summary page together with the pTPM value. The expression classification across the 37 tissue types included four tissues with combined data: brain, intestine, lymphoid tissues, and blood cells, all represented by the maximum NX value within each group. In general, tissues, cells, or regions including multiple data sources or multiple subtissues were all represented by a consensus NX value, calculated for each gene as the maximum NX value in the subtissues/regions or cell types.

Normalization of pig and mouse data

All TPM values of pig and mouse datasets were TMM normalized (22) between all samples, respectively, and then each gene was Pareto scaled (23) within each species (fig. S4). NX for each gene was calculated in every sample as described for human, including calculation of pTPM values. In the HPA, the pTPM value for every gene in every sample is visualized on the gene summary page and the more detailed tissue pages. For regions containing multiple subregions, a consensus NX value was calculated for each gene as the maximum NX value of the subregions (Fig. 1B).

Comparisons of three species

Protein-coding genes with one-to-one orthologs in human, mouse, and pig were identified to compare the expression profiles in the three mammalian brains, and altogether 12,999 genes were analyzed (fig. S19A). All NX values of the 12,999 genes were then TMM normalized (22) between 10 brain regions in three species (figs. S4 to S6).

Classification based on RNA expression

All protein-coding genes were classified according to a new strategy based on categorization on both tissue specificity (expression abundance between tissues, table S4) and tissue distribution (detection level above cutoff NX =1, table S8). Tissue specificity highlights genes with elevated expression in one or a group of tissue types compared with the rest, with the three elevated categories being "enriched" (fourfold higher expression in one tissue compared with the second highest), "group enriched" (fourfold higher expression in a group of tissues compared with other tissues), and "enhanced" (fourfold higher expression in one or several tissues compared with the mean of all tissues) (table S4). These classification rules were applied to the expression profiles of the 37 tissue types representing the whole human body as well as the different brain regions in human, pig, and mouse (Fig. 2). The tissue distribution defines the number of tissues with expression levels above cutoff (NX = 1) (table S8). The combination of tissue specificity and distribution from a brain perspective (genes detected in brain distributed into the different categories) is shown in table S7. Tissue-based classification, highlighting the brain-elevated genes compared with peripheral tissues, is available for all human protein-coding genes, while the regional classification in human brain is limited by the availability of external expression data (GTEx and FANTOM) (Fig. 1C and fig. S5 for more details about the gene coverage and combinations of the datasets). A second step of normalization was introduced to enable comparison of the expression levels across species. All human protein-coding genes with one-to-one orthologs in both mouse and pig (12,999 genes) were adjusted by TMM normalization, as illustrated in the schematic overview of the normalization pipeline, fig. S4.

Hierarchical clustering and UMAP analysis

Clustering in heatmaps and dendrograms based on Spearman correlation were created by first calculating a correlation matrix of Spearman's ρ (57) between all brain regions. The correlation was converted to a distance metric (1 – ρ) and was clustered using unsupervised top-down hierarchical clustering, where, at each stage, the distances between clusters are recomputed by the Lance-Williams dissimilarity update formula according to average linkage. Dendrograms showing gene expression in heatmaps have been clustered using the Ward2 algorithm (58), an implementation of Ward's minimum variance method (59) implemented as "Ward.D2" in the hclust function in the R package stats, where clusters are chosen at each stage such that the increase in cluster variance is minimized after merging. The hierarchical clustering of brain regions in three species was conducted by using the neighbor-joining approach in the ape package (60) in R, based on pairwise Pearson correlational distances between samples. The reliability of branches was assessed using 100 bootstrap replicates. The phylogenetic tree was drawn using the plot.phylo function in ape. Uniform Manifold Approximation and Projection (UMAP) has been performed on NX values of brain samples by using the R packages UMAP (61) with default parameters.

Differential expression analysis of three species

Differential expression analysis was conducted by using normalized NX values of 10 regions of three species. The R package *limma*, which includes lmFit, eBayes, and topTable functions, was used for pairwise comparison of DEGs. False discovery rate (FDR) was calculated by using p.adjust() function in R, using the Benjamini-Hochberg method. Genes with FDRs less than 0.01 and absolute fold change larger than 2 were considered as differentially expressed genes.

Defining cell type signature genes

Human cerebral cortex signature genes for neurons, astrocytes, oligodendrocytes, and microglia were determined on the basis of the agreement between two independent (data source and approach) datasets. RNA-seq results of cells selected using immunopanning (8) were obtained from www.brainrnaseq.org, and results based on coexpression analysis (12) were obtained from http://oldhamlab.ctec.ucsf.edu/. By varying the inclusion criteria for RNA-seq data (fold-enrichment >2 to >5) and coexpression analysis (p-value 0.95 to 1) the optimal settings creating the maximum overlap between these datasets for each cell type were determined (Table 1). Human cerebral cortex cell type signature genes were defined as genes associated with the same cell type based on both datasets with an FPKM value of >1 in only one cell type based on RNA-seq. The list of 420 genes, here defined as cell type signature genes, are listed in table S9.

Antibody-based profiling of protein distribution

Protein profiling in human brain tissues was performed within the Human Protein Atlas pipeline, following previously published protocols (1). Formalin fixed paraffin embedded (FFPE) tissue samples were used for tissue microarray (TMA) construction, where 144 separate 1-mm cores were placed in a recipient paraffin block (62) representing 44 different tissue types. Sections were cut (4 μm by Microm HM 355S, Thermo Fisher Scientific) and placed on SuperFrost Plus glass slides (VWR). The sections were dewaxed, H_2O_2-incubated, and antigen retrieved by heat-induced epitope retrieval (HIER) in pH6 citric acid solution before commencing the staining procedure. The Leica Biosystems CV5030 immunostainer was used for pretreatment as well as in later steps of counterstaining and coverslipping. Staining protocols were standardized and executed in a

Table 1. Criteria used for best overlap in two independent datasets, defining the cell type signature genes. Percentages in parentheses represent maximal overlap between two datasets used for selection of inclusion criteria.

Cell type	RNA-seq enrichment (fold)	Coexpression analysis (*P* value)	Number of genes (two datasets)	Number of signature genes
Neurons	>2	>0.98	679 (30.0%)	196
Astrocytes	>3	>0.98	324 (40.2%)	180
Oligodendrocytes	>2	1	212 (32.3%)	65
Microglia	>5	>0.95	135 (19.4%)	51

LabVision Autostainer 480 using LabVision reagents and protocols with primary antibody incubation 30 min in room temperature and HRP-polymer secondary antibody and DAB (3,3-diaminobenzidine) chromogenic visualization. All slides were counterstained in HTXplus (Histolab) before coverslipping, and image digitalization was performed in Scanscope AT2 (Aperio Vista) using a 20× objective.

Protein profiling in mouse brain tissues was performed as described previously (*63*). Complete brain profiles were represented by 20 to 25 sections having a 400-μm interval, covering all major regions of the mouse brain. Briefly, sections were incubated with primary antibody (16 to 72 hours at 4°C), blocked in TNB buffer (0.1 M Tris-HCl, pH 7.5; 0.15 M NaCl; 0.5% blocking reagent) followed by HRP-conjugated secondary antibody (Dako), and immunoreactivity was visualized using the tyramide signal amplification system (TSA-Plus; NEN Life Science Products, Inc.). Fluorescent images were obtained using a “VSlide” slide scanning microscope (MetaSystems) equipped with a CoolCube 2 camera (12-bit grayscale), a 10× objective and filter sets for 4′,6-diamidino-2-phenylindole (DAPI) (EX350/50–EM470/40), Fluorescein isothiocyanate (FITC) (EX493/16–EM527/30), Cyanine (Cy) 3 (EX546/10–EM580/30), Cy3.5 (EX581/10–EM617/40), and Cy5 (EX630/20–647/long pass). Finally, the individual images were stitched together (VSlide) to generate a large image of the entire section, while the images (vsi-files) were additionally extracted to high quality .jpeg files for further analysis using the software Metaviewer (Metasystems). All images were manually evaluated and scored, always including verification by a second observer.

The human brain antibody-based chromogen stainings are all available on the Human Protein Atlas portal (www.proteinatlas.org). Antibody IDs and antibody dilutions are listed in table S13. More details about antibody validation and antigen design for respective antibodies are found on the respective antibody information page in the portal. Fluorescent mouse brain staining shown in Fig. 3 are examples from full mouse protein profiles available online, performed according to the protocol described above (list of antibodies in table S13). The examples shown for species comparison in human, pig, and mouse are all performed on FFPE tissue sections (4 μm), placed on SuperFrost Plus glass slides (VWR), baked before dewaxing and antigen retrieved according to standard procedure for FFPE sections as described previously, including H_2O_2-incubated and HIER in pH 6 citric acid solution. The sections were then incubated overnight at 4°C with primary antibody (list of antibodies and dilution in table S13), followed by blocking and secondary HRP-conjugated secondary antibody. The staining protocol followed was according to the mouse profiling standard using TSA amplification.

iDISCO+ volume immunostaining

iDISCO+ volume immunostaining and clearing process were performed as earlier described by Renier and colleagues (*64*). Briefly, whole mouse brains (one hemisphere was laterally slightly trimmed) were washed in 0.01 M PBS three times in 5-ml Eppendorf tubes and then dehydrated in a series of methanol/water solutions for 1 hour each. The samples were then bleached with 5% hydrogen peroxide in 100% methanol overnight at +4°C. Then, they were rehydrated, incubated in permeabilization solution for 2 days, and followed by blocking solution for an additional 2 days, both at 37°C (0.2% Triton-X100/20% DMSO/0.3 M glycine in 0.01 M PBS + 0.02% sodium azide, 0.2% Triton-X100/10% DMSO/6% normal donkey serum in 0.01 M PBS + 0.02% sodium azide, respectively). The samples were then incubated with a rabbit polyclonal primary antibody raised against human GPR151 (HPA065728, 1:150) solution for 7 days at 37°C (antibody diluent: 0.2% Tween-20/10 μg/ml heparin/5% DMSO/3% normal donkey serum in 0.01 M PBS + 0.02% sodium azide). After extensive washing, the blocks were incubated in secondary antibody (1:150; goat anti-rabbit, conjugated with Alexa Fluor 647; Molecular Probes, Oregon, USA) solution (0.2% Tween-20/10 μg/ml heparin/3% normal donkey serum in 0.01 M PBS + 0.02% sodium azide). The blocks were dehydrated in methanol/water series, incubated in 66% dichloromethane/33% methanol for 3 hours and in 100% dichloromethane for 2 × 15 min. Finally, the blocks were moved to tubes and stored in 100% dibenzyl ether for the long term.

A light sheet microscope (Ultramicroscope II, Lavision Biotec, Bielefeld, Germany) and the Imspector software were used for image acquisition of the whole mouse brains. The microscope was equipped with an sCMOS camera (Andor Neo) and a 2×/0.5 objective lens (MVPLAPO 23), with a 6.5-mm working distance spherical aberration corrected dipping cap.

The trimmed full mouse brain (cut at lateral ca −2.5 mm) was fixed on the sample holder with the surface of the trimmed hemisphere side down and acquired in sagittal position from ca. lateral −2.00 until the top of the other side cortex (ca 6.5 mm altogether). To obtain the required X/Y/Z resolution, homogeneous illumination within the entire focal plane and minimal photobleaching, the following microscope parameters were applied: 2× objective, 1.6× zoom body, and additional magnification of the dipping cap lens (altogether 3.6× effective magnification), 70% laser power (OBIS 647 laser), bilateral illumination (blend merging algorithm), 100-ms exposure time, max sheet numerical aperture (0.149), dynamic horizontal focus process with 13 steps (blend merging algorithm after precalibration process), 70% sheet width, 2-μm Z-step thickness, mosaic acquisition mode (six tiles with 15% overlap). Stitching was achieved by the Terastitcher-Imspector python script (LaVision Biotec, 2017), where the serials of 16-bit uncompressed stitched tiff images (ca. 3500 z-levels, ca 100 GB) were then converted to IMS file, and the 3D vision of acquisitions was reconstructed in the Imaris 9.1.2 (Bitplane, UK) software for inspection and quality control.

REFERENCES AND NOTES

1. M. Uhlén *et al*., Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015). doi: 10.1126/science.1260419; pmid: 25613900
2. The HPA Tissue Atlas; www.proteinatlas.org/tissue.
3. E. S. Lein *et al*., Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007). doi: 10.1038/nature05453; pmid: 17151600
4. C. L. Thompson *et al*., A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. *Neuron* **83**,

309–323 (2014). doi: 10.1016/j.neuron.2014.05.033; pmid: 24952961

5. M. J. Hawrylycz *et al*., An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012). doi: 10.1038/nature11405; pmid: 22996553
6. J. A. Miller *et al*., Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014). doi: 10.1038/nature13185; pmid: 24695229
7. Y. Zhang *et al*., An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* **34**, 11929–11947 (2014). doi: 10.1523/JNEUROSCI.1860-14.2014; pmid: 25186741
8. Y. Zhang *et al*., Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016). doi: 10.1016/j.neuron.2015.11.013; pmid: 26687838
9. A. Zeisel *et al*., Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018). doi: 10.1016/j.cell.2018.06.021; pmid: 30096314
10. B. B. Lake *et al*., Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018). doi: 10.1038/nbt.4038; pmid: 29227469
11. R. D. Hodge *et al*., Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019). doi: 10.1038/s41586-019-1506-7; pmid: 31435019
12. K. W. Kelley, H. Nakao-Inoue, A. V. Molofsky, M. C. Oldham, Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**, 1171–1184 (2018). doi: 10.1038/s41593-018-0216-z; pmid: 30154505
13. J. R. Ecker *et al*., The BRAIN Initiative Cell Census Consortium: Lessons learned toward generating a comprehensive brain cell atlas. *Neuron* **96**, 542–557 (2017). doi: 10.1016/j.neuron.2017.10.007; pmid: 29096072
14. H. Markram, The human brain project. *Sci. Am.* **306**, 50–55 (2012). doi: 10.1038/scientificamerican0612-50; pmid: 22649994
15. HuBMAP Consortium, The human body at cellular resolution: The NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019). doi: 10.1038/s41586-019-1629-x; pmid: 31597973
16. A. Regev *et al*., The Human Cell Atlas. *eLife* **6**, e27041 (2017). doi: 10.7554/eLife.27041; pmid: 29206104
17. The HPA Brain Atlas; www.proteinatlas.org/brain.
18. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). doi: 10.1038/nature24277; pmid: 29022597
19. S. Noguchi *et al*., FANTOM5 CAGE profiles of human and mouse samples. *Sci. Data* **4**, 170112 (2017). doi: 10.1038/sdata.2017.112; pmid: 28850106
20. J. E. Coate, J. J. Doyle, Variation in transcriptome size: Are we getting the message? *Chromosoma* **124**, 27–43 (2015). doi: 10.1007/s00412-014-0496-3; pmid: 25421950
21. J. Lovén *et al*., Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012). doi: 10.1016/j.cell.2012.10.012; pmid: 23101621
22. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010). doi: 10.1186/gb-2010-11-3-r25; pmid: 20196867
23. R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* **7**, 142 (2006). doi: 10.1186/1471-2164-7-142; pmid: 16762068
24. M. E. Ritchie *et al*., *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015). doi: 10.1093/nar/gkv007; pmid: 25605792
25. J. C. Corvol, J. M. Studler, J. S. Schonn, J. A. Girault, D. Hervé, $G\alpha_{olf}$ is necessary for coupling D1 and A2a receptors to adenylyl cyclase in the striatum. *J. Neurochem.* **76**, 1585–1588 (2001). doi: 10.1046/j.1471-4159.2001.00201.x; pmid: 11238742
26. I. H. Kim *et al*., Evidence for functional diversity between the voltage-gated proton channel Hv1 and its closest related protein HVRP1. *PLOS ONE* **9**, e105926 (2014). doi: 10.1371/journal.pone.0105926; pmid: 25165868
27. N. Zainolabidin, S. P. Kamath, A. R. Thanawalla, A. I. Chen, Distinct activities of Tfap2A and Tfap2B in the specification of GABAergic interneurons in the developing cerebellum. *Front. Mol. Neurosci.* **10**, 281 (2017). doi: 10.3389/fnmol.2017.00281; pmid: 28912684
28. J. Mulder *et al*., Secretagogin is a Ca^{2+}-binding protein specifying subpopulations of telencephalic neurons. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 22492–22497 (2009). doi: 10.1073/pnas.0912484106; pmid: 20018755
29. E. S. Deneris, O. Hobert, Maintenance of postmitotic neuronal cell identity. *Nat. Neurosci.* **17**, 899–907 (2014). doi: 10.1038/nn.3731; pmid: 24929660
30. H. Guo *et al*., Specificity and efficiency of Cre-mediated recombination in Emx1-Cre knock-in mice. *Biochem. Biophys. Res. Commun.* **273**, 661–665 (2000). doi: 10.1006/bbrc.2000.2870; pmid: 10873661
31. P. S. Joshi *et al*., Bhlhb5 regulates the postmitotic acquisition of area identities in layers II-V of the developing neocortex. *Neuron* **60**, 258–272 (2008). doi: 10.1016/j.neuron.2008.08.006; pmid: 18957218
32. D. Jean, G. Bernier, P. Gruss, *Six6* (*Optx2*) is a novel murine *Six3*-related homeobox gene that demarcates the presumptive pituitary/hypothalamic axis and the ventral optic stalk. *Mech. Dev.* **84**, 31–40 (1999). doi: 10.1016/S0925-4773(99)00068-4; pmid: 10473118
33. I. Nunes, L. T. Tovmasian, R. M. Silva, R. E. Burke, S. P. Goff, Pitx3 is required for development of substantia nigra dopaminergic neurons. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4245–4250 (2003). doi: 10.1073/pnas.0230529100; pmid: 12655058
34. F. Chantoux, J. Francon, Thyroid hormone regulates the expression of NeuroD/BHF1 during the development of rat cerebellum. *Mol. Cell. Endocrinol.* **194**, 157–163 (2002). doi: 10.1016/S0303-7207(02)00133-8; pmid: 12242038
35. R. Grailhe *et al*., Increased exploratory activity and altered response to LSD in mice lacking the 5-HT(5A) receptor. *Neuron* **22**, 581–591 (1999). doi: 10.1016/S0896-6273(00)80712-6; pmid: 10197537
36. R. Grailhe, G. W. Grabtree, R. Hen, Human 5-HT(5) receptors: The 5-HT(5A) receptor is functional but the 5-HT(5B) receptor was lost during mammalian evolution. *Eur. J. Pharmacol.* **418**, 157–167 (2001). doi: 10.1016/S0014-2999(01)00933-5; pmid: 11343685
37. R. L. Smith, H. Baker, K. Kolstad, D. D. Spencer, C. A. Greer, Localization of tyrosine hydroxylase and olfactory marker protein immunoreactivities in the human and macaque olfactory bulb. *Brain Res.* **548**, 140–148 (1991). doi: 10.1016/0006-8993(91)91115-H; pmid: 1678294
38. M. Lebel, Y. Gauthier, A. Moreau, J. Drouin, Pitx3 activates mouse tyrosine hydroxylase promoter via a high-affinity binding site. *J. Neurochem.* **77**, 558–567 (2001). doi: 10.1046/j.1471-4159.2001.00257.x; pmid: 11299318
39. E. Mezey, Phenylethanolamine N-methyltransferase-containing neurons in the limbic system of the young rat. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 347–351 (1989). doi: 10.1073/pnas.86.1.347; pmid: 2563164
40. N. Puskás, R. S. Papp, K. Gallatz, M. Palkovits, Interactions between orexin-immunoreactive fibers and adrenaline or noradrenaline-expressing neurons of the lower brainstem in rats and mice. *Peptides* **31**, 1589–1597 (2010). doi: 10.1016/j.peptides.2010.04.020; pmid: 20434498
41. L. C. Daws, G. G. Gould, Ontogeny and regulation of the serotonin transporter: Providing insights into human disorders. *Pharmacol. Ther.* **131**, 61–79 (2011). doi: 10.1016/j.pharmthera.2011.03.013; pmid: 21447358
42. J. Peng, S. Sarkar, S. L. Chang, Opioid receptor expression in human brain and peripheral tissues using absolute quantitative real-time RT-PCR. *Drug Alcohol Depend.* **124**, 223–228 (2012). doi: 10.1016/j.drugalcdep.2012.01.013; pmid: 22356890
43. D. L. Kaufman *et al*., Characterization of the murine μ opioid receptor gene. *J. Biol. Chem.* **270**, 15877–15883 (1995). doi: 10.1074/jbc.270.26.15877; pmid: 7797593
44. M. Mortensen, B. Patel, T. G. Smart, GABA potency at GABA(A) receptors found in synaptic and extrasynaptic zones. *Front. Cell. Neurosci.* **6**, 1 (2012). pmid: 22319471
45. X. Gonda *et al*., A new stress sensor and risk factor for suicide: The T allele of the functional genetic variant in the GABRA6 gene. *Sci. Rep.* **7**, 12887 (2017). doi: 10.1038/s41598-017-12776-8; pmid: 29018204
46. A. S. Hauser, M. M. Attwood, M. Rask-Andersen, H. B. Schiöth, D. E. Gloriam, Trends in GPCR drug discovery: New agents, targets and indications. *Nat. Rev. Drug Discov.* **16**, 829–842 (2017). doi: 10.1038/nrd.2017.178; pmid: 29075003
47. K. Mizushima *et al*., A novel G-protein-coupled receptor gene expressed in striatum. *Genomics* **69**, 314–321 (2000). doi: 10.1006/geno.2000.6340; pmid: 11056049
48. D. M. Berson, F. A. Dunn, M. Takao, Phototransduction by retinal ganglion cells that set the circadian clock. *Science* **295**, 1070–1073 (2002). doi: 10.1126/science.1067262; pmid: 11834835
49. M. N. Moraes *et al*., Melanopsin, a canonical light receptor, mediates thermal activation of clock genes. *Sci. Rep.* **7**, 13977 (2017). doi: 10.1038/s41598-017-13939-3; pmid: 29070825
50. M. Uhlen *et al*., A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366**, eaax9198 (2019). doi: 10.1126/science.aax9198; pmid: 31857451
51. E. Sjöstedt *et al*., Defining the human brain proteome using transcriptomics and antibody-based profiling with a focus on the cerebral cortex. *PLOS ONE* **10**, e0130028 (2015). doi: 10.1371/journal.pone.0130028; pmid: 26076492
52. M. W. Vermunt *et al*., Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat. Neurosci.* **19**, 494–503 (2016). doi: 10.1038/nn.4229; pmid: 26807951
53. H. Takahashi, S. Kato, M. Murata, P. Carninci, CAGE (cap analysis of gene expression): A protocol for the detection of promoter and transcriptional networks. *Methods Mol. Biol.* **786**, 181–200 (2012). doi: 10.1007/978-1-61779-292-2_11; pmid: 21938627
54. D. R. Zerbino *et al*., Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018). doi: 10.1093/nar/gkx1098; pmid: 29155950
55. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016). doi: 10.1038/nbt.3519; pmid: 27043002
56. R Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing, 2018); www.R-project.org.
57. C. Spearman, The proof and measurement of association between two things. By C. Spearman, 1904. *Am. J. Psychol.* **100**, 441–471 (1987). doi: 10.2307/1422689; pmid: 3322052
58. L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster analysis* (Wiley Series in Probability and Statistics, Wiley, 1990).
59. F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J. Classif.* **31**, 274–295 (2014). doi: 10.1007/s00357-014-9161-z
60. S. Bhattacharya *et al*., ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci. Data* **5**, 180015 (2018). doi: 10.1038/sdata.2018.15; pmid: 29485622
61. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 [stat.ML] (9 February 2018).
62. C. Kampf, I. Olsson, U. Ryberg, E. Sjöstedt, F. Pontén, Production of tissue microarrays, immunohistochemistry staining and digitalization within the Human Protein Atlas. *J. Vis. Exp.* 10.3791/3620 (2012). doi: 10.3791/3620; pmid: 22688270
63. J. Mulder *et al*., Tissue profiling of the mammalian central nervous system using human antibody-based proteomics. *Mol. Cell. Proteomics* **8**, 1612–1622 (2009). doi: 10.1074/mcp.M800539-MCP200; pmid: 19351664
64. N. Renier *et al*., Mapping of brain activity by automated volume analysis of immediate early genes. *Cell* **165**, 1789–1802 (2016). doi: 10.1016/j.cell.2016.05.007; pmid: 27238021

ACKNOWLEDGMENTS

We thank W. Gu from the School of Biotechnology, Southern Medical University, Guangzhou, for providing the pigs. We acknowledge the entire staff of the Human Protein Atlas program and the Science for Life Laboratory for their valuable contributions. Support from the National Genomics Infrastructure in Stockholm is acknowledged, with funding from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. **Funding:** Main funding was provided by the Knut and Alice Wallenberg Foundation (WCPR) and the Erling Persson Foundation (KCAP). The pig atlas part is supported by Sanming Project of Medicine in Shenzhen (SZSM201612074). L.L. is supported by the Lundbeck Foundation (R219–2016-1375) and the DFF Sapere Aude Starting grant (8048-00072A). **Author contributions:** M.U. and J.M. conceived of and designed the study. J.H., Y.D., L.L., Z.D., X.L., H.J., and Y.L. performed pig brain experiments (sample processing, RNA extraction, quality control, NGS library constructions, sequencing, and pilot data analysis). W.Z., L.F., M.K., A.M., and C.Z. performed the bioinformatics analysis. K.v.F. and P.O. provided the infrastructure for the data. E.S., M.U., W.Z., T.H., and J.M. drafted the manuscript. All authors discussed the results and contributed to the final manuscript. **Competing interests:** T.H. has shares in Lundbeck and Bioarctic. No other authors declare any competing interests. **Data and materials availability:** Materials are available upon request from J.M. All classification information, data, and gene

information are available, open access, and downloadable from the Human Protein Atlas database. The external data from GTEx, FANTOM5, and the Allen Brain Atlas are available at their respective sources. The transcript expression values at the brain regional level for human, pig, and mouse, as well as the remapped TPM and pTPM tables for the external data, are available from the HPA download section (www.proteinatlas.org/about/download), in addition to the human data for all tissue and blood cell types. The original data for the pig brain transcriptome has been deposited to the public data depository CNGB Nucleotide Sequence Archive (CNSA; https://db.cngb.org/cnsa/) of the China National GeneBank DataBase (CNGBdb) with accession number CNP0000483.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/367/6482/eaay5947/suppl/DC1

Figs. S1 to S35
Tables S1 to S13
Movie S1

8 July 2019; accepted 6 February 2020
10.1126/science.aay5947

METABOLISM

An atlas of human metabolism

Jonathan L. Robinson[1,2]*, Pınar Kocabaş[1,2]*, Hao Wang[1,3,4]*, Pierre-Etienne Cholley[4]*, Daniel Cook[1], Avlant Nilsson[1], Mihail Anton[4], Raphael Ferreira[1], Iván Domenzain[1,2], Virinchi Billa[1], Angelo Limeta[1], Alex Hedin[1], Johan Gustafsson[1,2], Eduard J. Kerkhoven[1], L. Thomas Svensson[4], Bernhard O. Palsson[5,6,7], Adil Mardinoglu[8,9], Lena Hansson[4,10], Mathias Uhlén[5,8,11], Jens Nielsen[1,2,5,12]†

Genome-scale metabolic models (GEMs) are valuable tools to study metabolism and provide a scaffold for the integrative analysis of omics data. Researchers have developed increasingly comprehensive human GEMs, but the disconnect among different model sources and versions impedes further progress. We therefore integrated and extensively curated the most recent human metabolic models to construct a consensus GEM, Human1. We demonstrated the versatility of Human1 through the generation and analysis of cell- and tissue-specific models using transcriptomic, proteomic, and kinetic data. We also present an accompanying web portal, Metabolic Atlas (https://www.metabolicatlas.org/), which facilitates further exploration and visualization of Human1 content. Human1 was created using a version-controlled, open-source model development framework to enable community-driven curation and refinement. This framework allows Human1 to be an evolving shared resource for future studies of human health and disease.

INTRODUCTION

Human metabolism is an integral part of cellular function, and many health conditions such as obesity, diabetes, hypertension, heart disease, and cancer (*1*, *2*) are associated with abnormal metabolic states. Several of these conditions can be diagnosed by screening for metabolite biomarkers in a patient's blood or urine (*3*), and recent studies have explored targeting metabolic processes for disease treatment (*4*, *5*).

Despite the importance of metabolism and advances allowing for simultaneous measurement of thousands of metabolites (*6*), understanding metabolism in a holistic manner in human cells remains challenging. One reason for this difficulty is that the defining feature of metabolism is not the concentrations of biomolecules themselves (such as metabolites, mRNA, or proteins), but metabolic fluxes through reactions, for which concentrations can only be used as indirect proxies for biological activity (*7*). This challenge has been addressed by building genome-scale metabolic models (GEMs), which have been used, for instance, in industrial applications involving *Saccharomyces cerevisiae* and *Escherichia coli* to understand metabolism, engineer new cellular objectives (such as biofuel production), and increase product yield (*8*, *9*).

Over the past 15 years, researchers have devoted a concerted effort to develop and improve such GEMs for human metabolism. This effort began in earnest with the development of Recon1 (*10*) and the Edinburgh Human Metabolic Network (EHMN) (*11*), which served as the starting point for two parallel model series: the Recon series (Recon1, 2, and 3D) (*10*, *12*, *13*) and the Human Metabolic Reaction series (HMR1 and 2) (*14*, *15*). These two model lineages incorporate heavily from each other during updates (fig. S1) and have been used to investigate diseases that include dysbiosis, diabetes, fatty liver disease, and cancer (*16*–*19*). Nevertheless, several challenges remain in the development of a human GEM, including the use of nonstandard identifiers for genes, metabolites, and reactions; duplication of model components; propagation of errors from previous model iterations; effort divided among multiple model lineages; and model updates that are delayed, nontransparent, and difficult to coordinate among the scientific community.

Here, we present Human1, the first version of a unified human GEM lineage (Human-GEM), and Metabolic Atlas, its companion web portal. Human-GEM was developed by integrating and extensively curating the Recon and HMR model lineages. The entire development process was conducted systematically in a version-controlled Git repository to make all past and future changes publicly accessible and to facilitate collaboration with the larger research community. We demonstrate the versatility and predictive accuracy of Human1 through an integrative analysis of transcriptomic data from 33 tumors and 53 healthy tissues, a gene-essentiality investigation involving more than 620 different cell types, and the prediction of nutrient exchange and growth rates of NCI-60 cell lines using enzyme-constrained GEMs (ecGEMs) derived from Human1.

RESULTS

Human1 generation and curation

Our primary focus was to establish a systematically curated model of human metabolism that accurately represents the underlying

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-41258 Gothenburg, Sweden. [2] Wallenberg Center for Protein Research, Chalmers University of Technology, Kemivägen 10, SE-41258 Gothenburg, Sweden. [3] Wallenberg Center for Molecular and Translational Medicine, University of Gothenburg, Kemivägen 10, SE-41258 Gothenburg, Sweden. [4] Department of Biology and Biological Engineering, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Chalmers University of Technology, Kemivägen 10, SE-41258 Gothenburg, Sweden. [5] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark. [6] Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA. [7] Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA. [8] Department of Protein Science, Science for Life Laboratory, KTH–Royal Institute of Technology, SE-10044 Stockholm, Sweden. [9] Centre for Host-Microbiome Interactions, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London, London WC2R 2LS, UK. [10] Novo Nordisk Research Centre Oxford, Oxford OX3 7FZ, UK. [11] Wallenberg Center for Protein Research, KTH–Royal Institute of Technology, SE-10044 Stockholm, Sweden. [12] BioInnovation Institute, Ole Maaløes Vej 3, DK-2200 Copenhagen, Denmark.
*These authors contributed equally to this work.
†Corresponding author. Email: nielsenj@chalmers.se

Originally published 24 March 2020

biology. We therefore leveraged the collective knowledge contained within existing human GEMs by integrating their information into a single resource. Components and information from HMR2, iHsa (*20*), and Recon3D were integrated and reconciled to yield a unified GEM consisting of 13,417 reactions, 10,138 metabolites (4164 unique), and 3625 genes (Fig. 1 and table S1).

Curation of the integrated model to generate Human1 involved the removal of 8185 duplicated reactions and 3215 duplicated metabolites, revision of 2016 metabolite formulas, rebalancing of 3226 reaction equations, correction of reversibility for 83 reactions, and the inactivation or removal of 576 reactions that were inconsistent (violated mass or energy conservation) or deemed unnecessary (tables S1 to S3). We also constructed a new generic human biomass reaction based on various tissue and cell composition data sources to facilitate flux simulations and other analyses relying on such a reaction (data files S1 and S2). All model changes were documented to provide justification and to ensure reproducibility. Furthermore, to ensure that these changes remained consistent with previous human GEM simulation studies, we repeated the infant growth simulation presented by Nilsson *et al.* (*21*) and found excellent agreement between their HMR2-based results and our Human1-based simulations (fig. S2).

The quality of Human1 was evaluated using Memote, a community-maintained framework for assessing GEMs with a standardized set of tests and metrics (*22*). In terms of consistency, Human1 exhibited excellent performance with 100% stoichiometric consistency, 99.4% mass-balanced reactions, and 98.2% charge-balanced reactions (fig. S3). This is a considerable improvement over the most recent GEM, Recon3D, which had 19.8% stoichiometric consistency and 94.2% mass-balanced and 95.8% charge-balanced reactions. Although the "model" version of Recon3D is fully stoichiometrically consistent and has a similar charge balance percentage (98.7%) as Human1, it has a lower percentage of mass-balanced reactions (97.3%) and contains 20% fewer total reactions and 33% fewer metabolites compared to Human1. The average Memote annotation score for metabolites, reactions, genes, and SBO (systems biology ontology) terms in Human1 was 66%; although this is a substantial improvement over previous models (46% for HMR2 and 25% for Recon3D), it indicates an area requiring further attention. We also used Memote to evaluate all 27 Human-GEM releases (versions) preceding Human1 to resolve

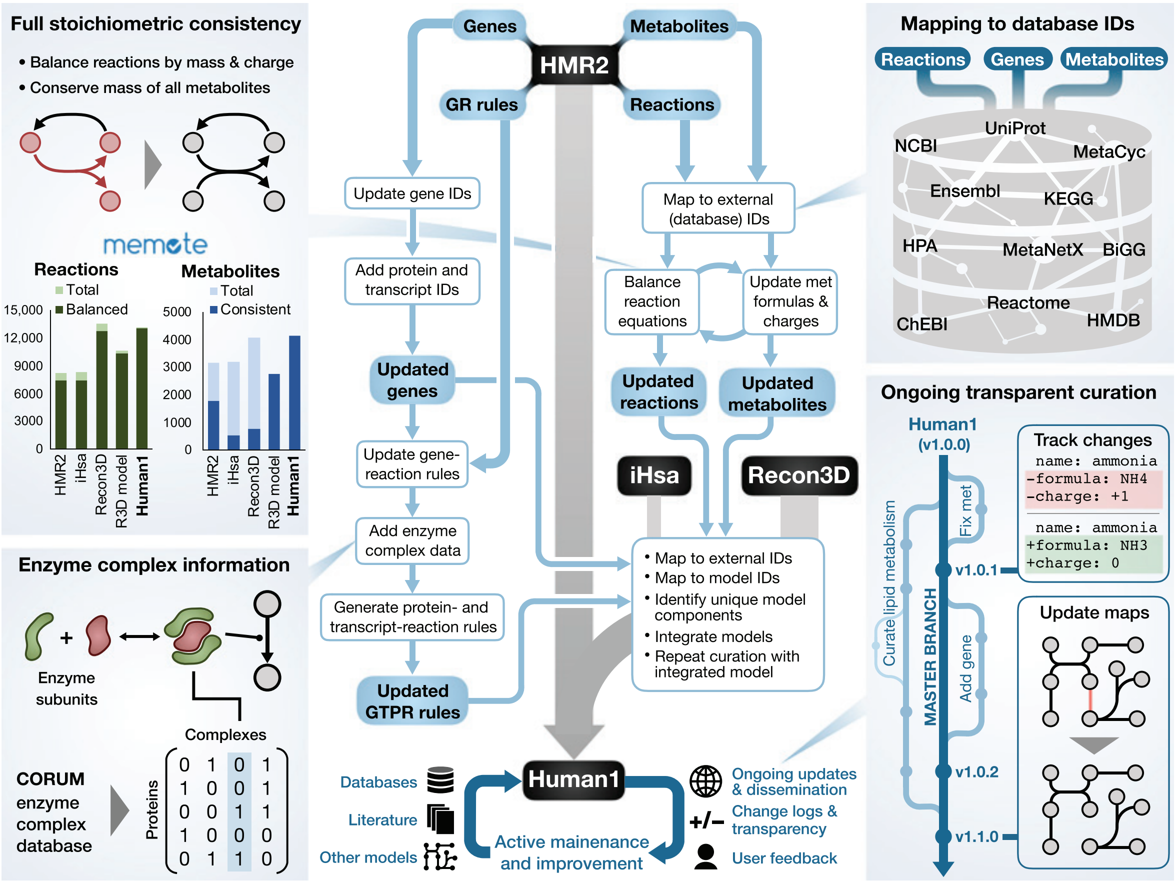


Fig. 1. Overview of Human1 generation and curation. A simplified illustration of the key steps involved in the generation of Human1 from HMR2, Recon3D, and iHsa. The bottom of the diagram represents the ongoing open-source curation of Human1 using input from databases, literature, other models, and the scientific community. The four side panels provide further detail into selected Human1 features: extensive reaction mass and charge balancing to achieve 100% stoichiometric consistency, incorporation of new enzyme complex information, mapping model components to standard database identifiers, and version-controlled and open-source model curation framework. In the bar graphs in the upper left panel, "Balanced" reactions represent the number of mass-balanced reactions, "Consistent" metabolites are the number of stoichiometrically consistent metabolites, and "R3D model" is the model version of Recon3D.

the effect of different curation processes on the various quality metrics (fig. S4, A to C).

A major advantage of GEMs is their ability to integrate different molecular datatypes to enable the interpretation of such data within the context of metabolism (*23*). We prioritized the curation and enhancement of gene-reaction associations for Human1 because such associations serve as an important link for the integration of multi-omics data. To this end, gene-reaction associations from HMR2, Recon3D, and iHsa were combined and integrated with enzyme complex information from Recon3D, iHsa, and the comprehensive resource of mammalian protein complexes database (CORUM) (*24*) to obtain gene-reaction rules for Human1. We also made available the transcript- and protein-reaction rules to facilitate direct integration of protein- or transcript-level data into the model, respectively (*25*). Furthermore, a key contribution of Recon3D was the association of protein structure information (such as 3D structure data) in a GEM-PRO data frame (*13*). We therefore regenerated the GEM-PRO data frame for Human1 to ensure that this same detailed protein information is also available for Human1.

An obstacle with existing human GEMs is their insufficient use of standard identifiers (such as KEGG, MetaCyc, and ChEBI) for many metabolites and reactions, thus impeding the retrieval of associated information from databases or the comparison of different models. To address this issue, we combined the available reaction and metabolite formulas, names, and identifiers in a semi-automated curation process using the MetaNetX reference database (*26*) to map 88.1% of reactions and 92.4% of metabolites to at least one standard identifier in Human1.

Other challenges facing human models are the ineffective communication and dissemination of their construction or revision. Traditionally, GEMs have been provided as a static object accompanying a publication, and thus, errors can remain without correction for years. On the basis of the approach applied for the Yeast8 GEM (*27*), we developed Human1 using a Git repository hosted on GitHub to establish a more systematic and community-driven development process. This configuration enables version control and tracking of all changes made to the model since its inception, accompanied by documentation such as commit messages and log files. The use of a public repository allows users to view or download the curation history of Human1 and submit issues to suggest changes or highlight errors. Thus, new knowledge can be efficiently integrated in future updates of the model using a community-wide effort.

Collectively, these improvements yield a standardized model enabling simple and accurate integration with different databases or omics datasets. We observed that the implementation of Human1 in a version-controlled framework such as Git is necessary to address many of the reproducibility and transparency concerns associated with computational research (*28*, *29*).

Metabolic Atlas

In parallel with the development of Human1, we developed Metabolic Atlas (www.metabolicatlas.org/), an online platform that enables interactive exploration of cell metabolism and convenient integration of omics data. Metabolic Atlas is an open-source reimplementation and complete redesign of its predecessor, the Human Metabolic Atlas (*30*).

Metabolic Atlas enables visualization of the complex metabolic network and interconnects model components (Fig. 2). It contains interactive two-dimensional (2D) maps at compartment and subsystem levels, allowing the use of smaller, more focused maps that pertain to metabolic areas of interest. The manually curated 2D maps cover 6793 nontransport/nonexchange reactions (90%), 4027 metabolites (97%), and 3316 genes (91%) present in Human1. These maps are integrated with transcriptomic data from the Human Protein Atlas (HPA) (*31*), upon which gene expression levels from 37 different tissue types can be overlaid. Users can also upload their own transcriptomic data to be visualized on the maps, and an expression comparison feature allows the overlaying of expression fold changes between two samples (such as different HPA tissues and/or user-uploaded data).

Selection of a component (gene, reaction, metabolite, subsystem) on any Metabolic Atlas map provides a descriptive summary on the sidebar, which includes a link to its complete information page with further details and links to external databases. Moreover, automatically generated 3D maps are available, which cover 100% of the Human1 network. In addition to maps, Metabolic Atlas dynamically generates graphs of interaction partners for any given enzyme or metabolite in Human1, which show the connectivity to other metabolites and enzymes based on their associated reactions. These graphs can be expanded to include more distant interaction partners and are also integrated with HPA transcriptomic data.

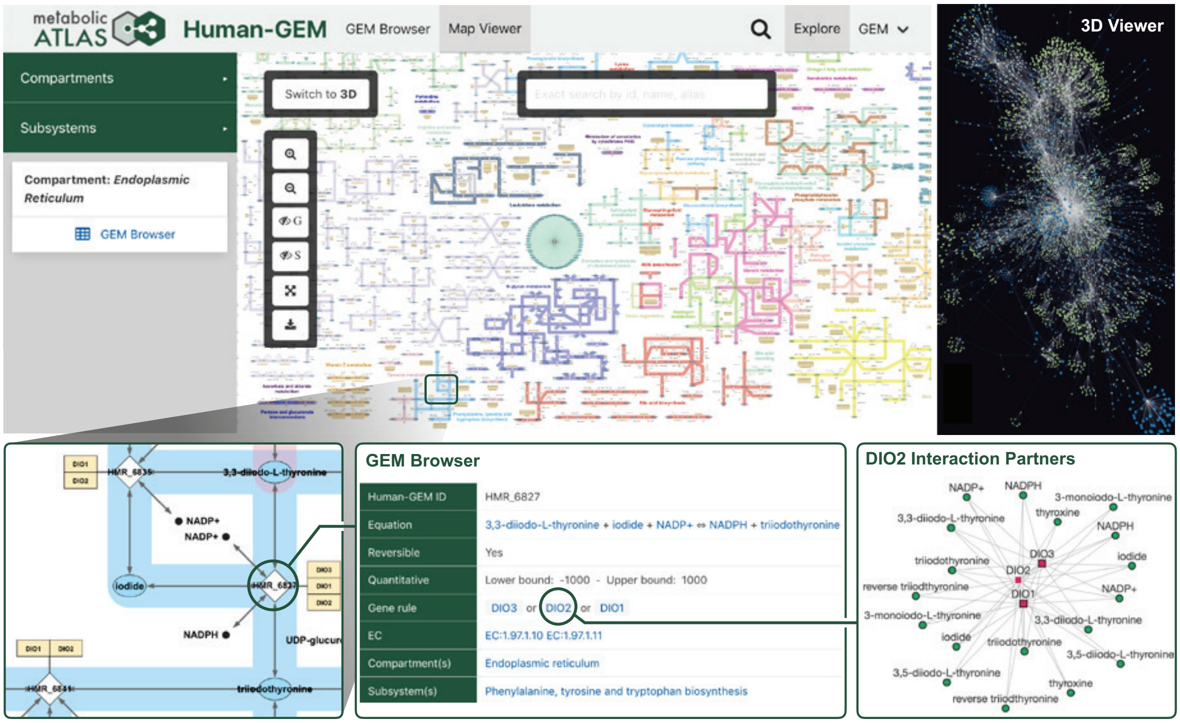
Metabolic Atlas continues to serve as a repository for an increasing number of GEMs (more than 350), ranging from those of individual human tissues and tumors to *S. cerevisiae* and other model organisms for fungi or bacteria. These models are summarized in a searchable table including information such as organism name, condition, year of publication, and number of reactions, metabolites, and enzymes. Furthermore, the content of Human1 can be accessed programmatically using the application programming interface (API) to retrieve, for example, all information associated with a given metabolite.

Metabolic Atlas provides a valuable resource and intuitive tool that complements the functionality of the Human1 model for studying metabolism. The coupling of Human1 and Metabolic Atlas enables valuable infrastructural support for future research in human health and disease.

Generation and comparison of healthy tissue– and tumor-specific models

To demonstrate the utility of Human1, we explored metabolic patterns across healthy tissues and primary cancers arising within those tissues. We performed GEM contextualization to construct tissue- and cancer-specific models because Human1 contains reactions across many human cell types and is thus not representative of any individual tissue or tumor type. The contextualization was performed using tINIT (*32*) based on gene expression levels from The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTEx) database (*33*) to construct 53 healthy tissue metabolic models and 33 cancer metabolic models.

We first investigated the global similarity in the structure of the metabolic models by comparing which reactions were included in each model. We visualized relationships across the reaction structures of the 86 models using a 2D t-distributed stochastic neighbor embedding (tSNE) projection, which showed that each cancer type's metabolic signature is more similar to the metabolism of its tissue of origin than to that of other cancer types (Fig. 3A and fig. S5). This phenomenon has also been observed when comparing gene expression data among different tissue and cancer types (*34*).



GEM Browser

Human-GEM ID	HMR_6827
Equation	3,3-diiodo-L-thyronine + iodide + NADP+ ⇔ NADPH + triiodothyronine
Reversible	Yes
Quantitative	Lower bound: -1000 - Upper bound: 1000
Gene rule	DIO3 or DIO2 or DIO1
EC	EC:1.97.1.10 EC:1.97.1.11
Compartment(s)	Endoplasmic reticulum
Subsystem(s)	Phenylalanine, tyrosine and tryptophan biosynthesis

Fig. 2. Highlighted features provided by the Metabolic Atlas web portal. A collection of screen captures from Metabolic Atlas, illustrating key features such as 2D and 3D metabolic network maps. A zoomed inset shows a subset of the endoplasmic reticulum compartment map, from which further information on components such as reactions, enzymes, or metabolites can be accessed in the GEM browser. Interaction partner graphs are dynamically generated for any given enzyme or metabolite in Human1, which show the connectivity to other metabolites and enzymes based on their associated reactions.

Several tissues and their associated tumors had markedly different metabolic capabilities than the other tissue models; these included the brain, liver, kidney, and tissues in the digestive system (stomach, colon, and rectum). This result highlights the role of these tissues as "metabolic specialists" as opposed to other human tissues.

We next focused on the GEMs of liver, liver cancer, blood, and blood cancer. A more detailed reaction structure comparison showed that liver and blood models (and their associated tumors) have distinct metabolic reaction structures and that, within liver models, cholangiocarcinoma (CHOL) was more distinct from healthy liver tissue, whereas hepatocellular carcinoma (LIHC) laid between the two states (Fig. 3B).

To further explore these differences, we investigated the metabolic subsystem coverage and functional differences between liver tissue and liver cancers. We found a distinct loss of metabolic functions in the CHOL GEM, including a deficiency in metabolic reactions associated with the urea cycle, bile acid recycling, metabolism of other amino acids, phenylalanine metabolism, and glucocorticoid biosynthesis (Fig. 3C), leading to a loss of function in urea production, ornithine degradation, arginine and creatine synthesis, ammonia import and degradation, and other metabolic tasks (Fig. 3D). The exception was proline de novo synthesis, which was the only metabolic task active in CHOL that was inactive in the other liver-related GEMs. This was supported at the mRNA level (visualized using Metabolic Atlas in fig. S6) and reflects previous studies that have shown increased proline synthesis and decreased proline degradation in other cancers in response to signaling through c-MYC and phosphatidylinositol 3-kinase (PI3K) oncogenes, where the disruption of such metabolic activity constitutes a potential therapeutic strategy (*35*, *36*). These and other approaches targeting metabolic functions such as ammonia buildup may constitute beneficial areas of research for developing CHOL treatments, which currently suffers from a lack of targeted therapies (*37*).

The construction of healthy and cancer-specific GEMs allowed us to compare cancer metabolism to healthy metabolism in systems for which paired normal tissue was not collected along with cancer tissue. An example is the comparison of the metabolism of acute myeloid leukemia (LAML) to that of healthy blood. The LAML GEM was characterized by a large increase in metabolic function over healthy blood (Fig. 3, E and F), including processes such as glucocorticoid biosynthesis, fatty acid oxidation (fig. S7), glycosphingolipid synthesis, and amino acid metabolism. This observation is consistent with previous studies showing that LAML relies on elevated fatty acid oxidation (*38*) and exhibits increased glycosphingolipid biosynthesis (*39*), which is associated with resistance to chemotherapeutics (*40*).

The large gain of metabolic function in LAML provides a rich number of pathways to target, such as heme biosynthesis, which constitutes a potential target for the treatment of LAML (*41*, *42*). Moreover, reduced coverage of a metabolic pathway in the disease-state GEM may indicate a less robust metabolic function that is more susceptible to therapeutic disruption. For example, the LAML GEM contained fewer reactions in the heme degradation subsystem compared to that of healthy blood, suggesting that targeting such activity could prove beneficial for treating LAML. Supporting this observation, inhibition of oxidative heme degradation

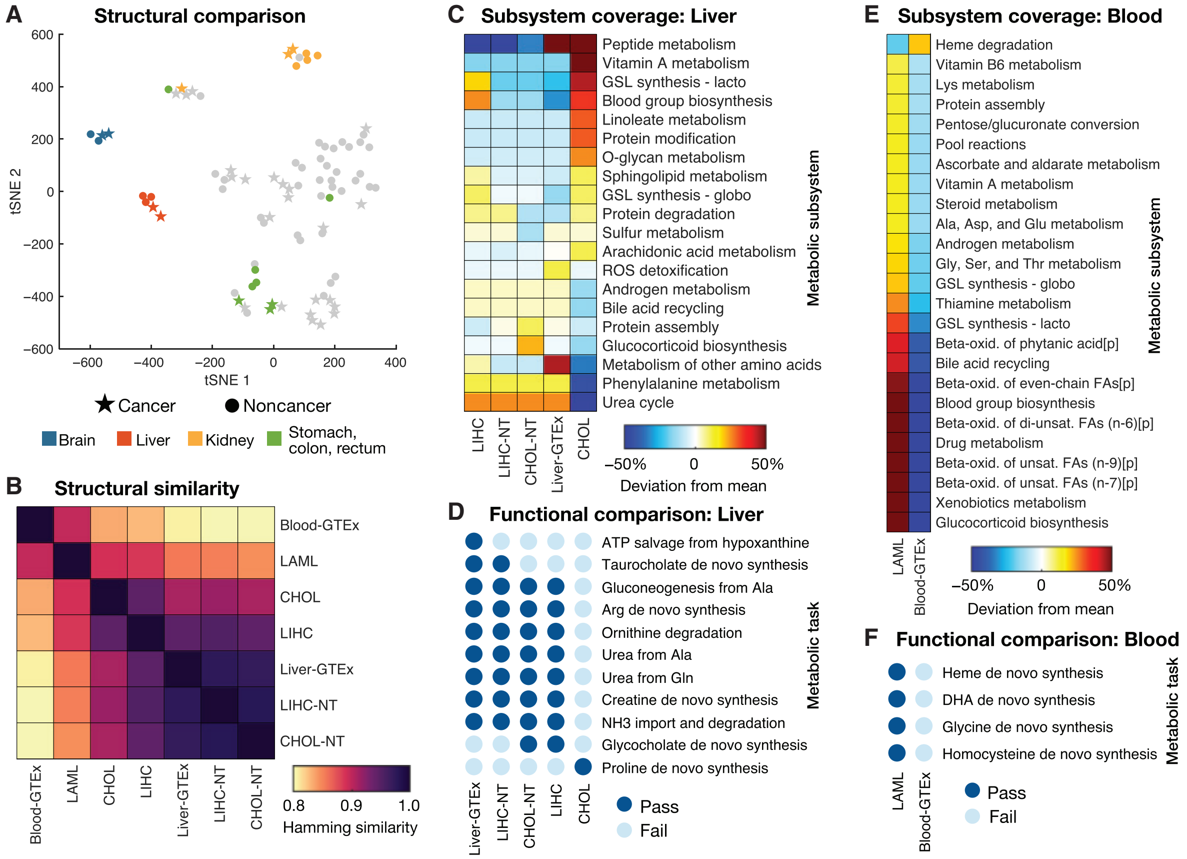


Fig. 3. Structural and functional comparison of cancer- and healthy tissue–specific GEMs. (**A**) Visualization of differences in models' reaction content using a tSNE projection to two dimensions based on the Hamming similarity. See fig. S5 for individual point labels. (**B**) Heat map showing pairwise comparisons of reaction content between GEMs specific to healthy liver (CHOL-NT, LIHC-NT, and Liver-GTEx), blood, and their corresponding cancers (CHOL, LIHC, and LAML). (**C**) Relative subsystem coverage (number of reactions present in a model that are associated with the given subsystem) compared among GEMs of liver and liver tumors. Only subsystems with at least a 10% deviation from mean subsystem coverage among the models are shown. (**D**) Summary of metabolic task performance by the healthy and cancerous liver models, showing only the tasks that differed in at least one of the models. (**E**) Comparison of relative subsystem coverage between LAML- and blood-specific GEMs, showing only subsystems with at least a 10% deviation between the two models. (**F**) Summary of the five metabolic tasks that could be completed by the LAML GEM but failed in the healthy blood GEM. ROS, reactive oxygen species; GSL, glycosphingolipid; FA, fatty acid; [p], peroxisomal compartment; DHA, docosahexaenoic acid.

has been demonstrated to be a promising treatment for myeloid leukemia (*43*).

Prediction of metabolic task-essential genes in human cell lines

Following the construction and analysis of context-specific GEMs derived from Human1, we performed additional analyses to validate the network topologies of such models. Gene-reaction associations encoded within GEMs enable predictions of how gene perturbations (such as deletions) affect metabolic functionality. A common approach involves the prediction of essential genes by determining which genes, when deleted in silico, sufficiently reduce or eliminate the function of a specified objective reaction, such as biomass production (*44*). This predicted set of essential genes can then be compared with experimental gene essentiality measurements to quantitatively evaluate model performance.

Genome-wide knockout screens have provided gene essentiality data to validate microbial GEMs, but these data have been unavailable for human cells due to challenges in genetically engineering these cells. Because the development of CRISPR technologies has enabled high-throughput genome-wide knockout screens in human cell lines, we leveraged this new data source to evaluate Human1 gene essentiality predictions. We retrieved gene essentiality data from a CRISPR knockout screen performed in five different human cell types: GBM, a patient-derived glioblastoma cell line; RPE1, retinal epithelial cells; HCT116 and DLD1, colorectal carcinoma cell lines; and HeLa, a cervical cancer cell line (*45*). Five cell line–specific GEMs were constructed from Human1 using tINIT and their respective gene expression [RNA sequencing (RNA-seq)] profiles (*45*), and in silico gene deletions were performed on each GEM (Fig. 4A). Rather than focusing solely on growth, essential genes were defined as those which, upon deletion, impaired any of the 57 basic metabolic tasks (including biomass production) that are required for human cell viability (data file S3) (*32*). This more general definition of gene essentiality reduces the extent to which predictions depend on the formulation of the biomass reaction and was hypothesized to increase sensitivity of the predictions by accounting for more functions of the metabolic network. We repeated this process using HMR2 and Recon3D as the template GEMs to enable comparison of Human1 performance with previous human model iterations.

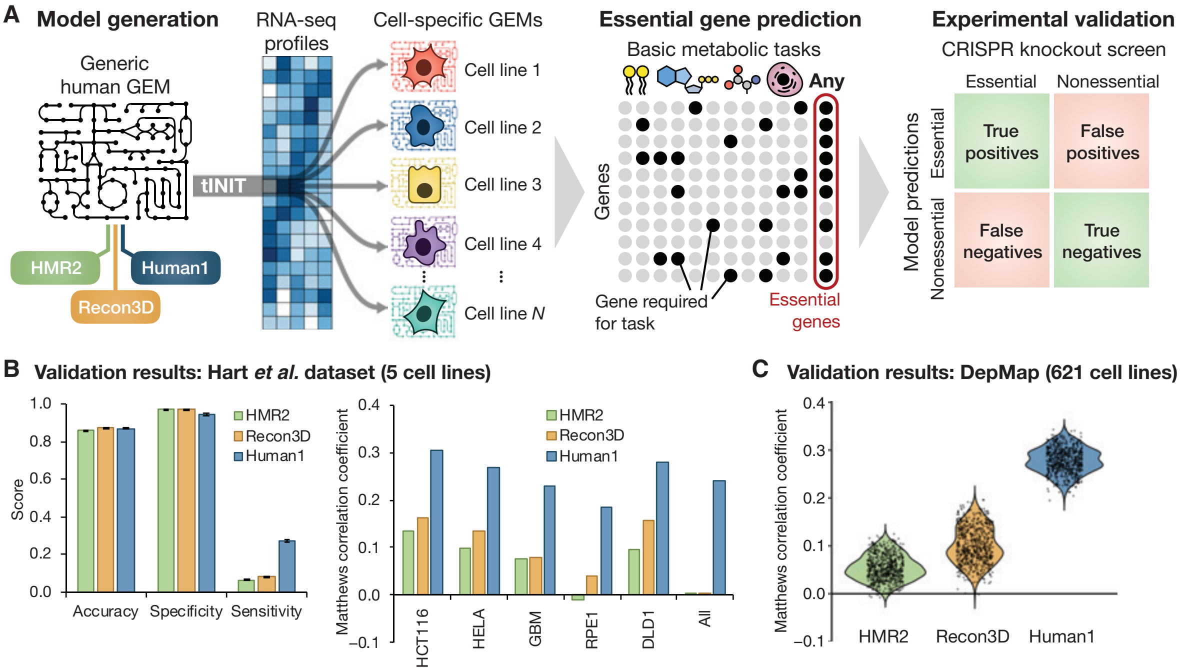


Fig. 4. Predicted gene essentiality among different cell lines and human GEMs. (**A**) Schematic illustration of the generation of cell line–specific GEMs from HMR2, Recon3D, and Human1 and subsequent prediction of gene essentiality based on the GEMs' ability to perform basic metabolic tasks. Genes predicted to be essential by the GEMs were compared to experimental measures of gene essentiality (*45*, *49*) obtained from CRISPR knockout screens. (**B**) Comparison of gene essentiality predictions among the three reference GEMs and their five derivative cell line models with CRISPR screen results from Hart *et al.* (*45*). Left: Average accuracy, specificity, and sensitivity of predictions across the five cell lines for each reference GEM, with error bars representing the SE of the mean. Right: Comparison of the Matthews correlation coefficient (MCC) of the predictions for each of the reference GEMs and cell lines. The "All" category indicates genes found to be essential in all five cell lines. (**C**) Comparison of gene essentiality predictions among the three reference GEMs and their 621 derivative cell line models with CRISPR screen results from the DepMap database (*49*).

We compared model-predicted essential genes for each individual cell line (as well as those essential in all five cell lines) to the set of essential genes identified in the corresponding CRISPR screen. The results were organized as confusion matrices quantifying the number of true and false positives and negatives (Fig. 4A), which were then used to evaluate prediction performance using several metrics (Fig. 4B). The general robustness of cells toward perturbations such as single-gene knockouts (*46*) yields a much smaller number of essential genes than nonessential genes, resulting in highly imbalanced group sizes. Accuracy is therefore an inappropriate metric for assessing the quality of gene essentiality predictions. For example, although all reference models (HMR2, Recon3D, and Human1) achieved similarly high accuracy across all cell types (mean accuracy of 86 to 88%), the same degree of accuracy is achieved if all genes are simply predicted as nonessential. This feature is reflected in the high specificity but low sensitivity exhibited by all three reference models. A more balanced prediction metric, the Matthews correlation coefficient (MCC) (*47*), was therefore calculated and compared among the different reference and cell-specific GEMs. Although the MCC values were relatively low overall, they showed a substantial increase (more than 2.5-fold) in prediction quality for Human1-derived GEMs compared to HMR2- and Recon3D-derived models. Moreover, a hypergeometric test for enrichment of true positives in each model's set of predicted essential genes showed significant enrichment for predictions from all Human1-derived GEMs (all $P < 10^{-20}$), whereas HMR2- and Recon3D-derived GEMs performed no better than random ($P > 0.05$) in predicting essential genes for the RPE1 cell line and/or those common to all five cell lines (fig. S8).

To further verify the improvement in Human1 gene essentiality predictions, we repeated the same pipeline (Fig. 4A) using RNA-seq profiles and CRISPR knockout screen data for 621 human cell lines retrieved from the DepMap database (*48*, *49*). The prediction performance of these 1863 cell-specific GEMs (621 models derived from each of the three reference GEMs) was again evaluated using several different metrics (fig. S9, A to D), including MCC (Fig. 4C). The analysis further confirmed the improvement in the performance of Human1, which exhibited a 2.8-fold mean increase in MCC over Recon3D. Because the CRISPR knockout screen scored genes on a continuous scale, it required the use of a threshold to categorize genes as essential or nonessential. We therefore repeated the analysis with a range of threshold values to confirm that our results were insensitive to this parameter (fig. S10). To ensure that the selection of metabolic tasks was not biasing the results, we repeated the analysis using only biomass production to define gene essentiality. Although the relative performance between the three reference models was not affected, the results demonstrated an increased sensitivity in all GEMs' predictions when using metabolic tasks instead of only biomass to define gene essentiality (fig. S11, A and B).

Collectively, these results demonstrated a marked improvement in Human1 over previous GEMs. However, the large number and diversity of curations involved in the development of Human1 make it difficult to resolve which changes contributed to the improved gene essentiality predictions. We therefore repeated the gene essentiality analysis pipeline (Fig. 4A) and comparison with the five cell lines from the Hart 2015 dataset (*45*) for all 27 versions preceding the current release of Human1 (v1.3.0). Although the largest increases in performance were the result of updates to model genes or gene-reaction rules (based on database information, other GEMs, or the literature), other curations such as mass-balancing reactions and correcting reversibility of reactions associated with the electron transport chain also contributed to increases in Human1 predictive performance (table S3 and fig. S4D).

An enzyme-constrained human model

Human GEMs are often poorly constrained because of the limited availability of measured flux data, as well as the reliance of human cells on essential amino acids and vitamins as nutrients in addition to a dominant carbon source such as glucose (*50*). The GECKO (enhancement of a Genome-scale model with Enzymatic Constraints using Kinetic and Omics data) modeling framework was developed to integrate enzyme abundance and kinetic data into GEMs to constrain the flux space to a more meaningful region without requiring extensive nutrient exchange data (*51*). We therefore applied the GECKO framework to Human1-derived GEMs to generate enzyme constrained ecGEMs. GECKO implements enzyme constraints by incorporating the enzymes into their catalyzed reactions as pseudo-metabolites with a stoichiometric coefficient inversely proportional to their turnover rate (k_{cat}). The explicit incorporation of enzymes allows the use of absolute proteomics datasets as constraints for each protein. If protein measurements are not available, the total protein content can be used as a global constraint for an additional pseudo-metabolite (protein pool) from which all enzymes are drawn.

To evaluate the improvement in flux predictions for ecGEMs derived from Human1, we used 11 NCI-60 cell line–specific GEMs generated during the gene essentiality analysis (part of the DepMap dataset) for which reliable nutrient exchange rate data (*52*, *53*) were available. Other NCI-60 cell lines were excluded as their metabolite exchange data were deemed unreliable due to early depletion of one or more nutrients (*53*, *54*). Enzyme constraints were incorporated into each of these cell-specific GEMs using the GECKO framework, yielding 11 cell-specific ecGEMs (Fig. 5A).

After generating the cell-specific ecGEMs, we sought to evaluate the impact of the enzyme constraints on the accessible (feasible) flux space. An approach often used to assess the feasible flux range for all reactions in a model is flux variability analysis (FVA) (*55*). We conducted FVA on each of the 11 cell-specific ecGEMs and compared the flux variabilities with the corresponding non-ecGEMs. The analysis revealed a substantial reduction in solution space, where the median decrease in flux variability across the 11 cell line models ranged from 3.5 to 7 orders of magnitude (Fig. 5B, fig. S12, and data file S4).

The integration of enzyme constraints substantially reduced the available flux space of Human1 but did not guarantee that this space was more accurate or biologically meaningful. We therefore sought to validate the ecGEMs by comparing predicted exchange fluxes with measured fluxes for 26 different metabolites and comparing growth rates (data file S5) (*52*). Fluxes were simulated by maximizing biomass production while specifying only which metabolites were present in the medium (Ham's medium)—no uptake or excretion rates were provided. Under these conditions, exchange fluxes cannot be predicted by non-ecGEMs because the solution is unbounded (the maximum growth rate is effectively infinite). However, all ecGEMs were able to predict finite exchange fluxes for each of the 26 metabolites as well as growth rates, where most (~78%) were in reasonably good agreement with experimental measurements (Fig. 5C). The largest disagreements involved the overprediction of fluxes for folate, α-ketoglutarate, and aspartate and an underprediction for pyruvate, carnitine, and ornithine.

To further explore the improvement in flux predictions upon incorporating enzyme constraints into Human1-derived GEMs, we investigated the effect of specifying one or more metabolite exchange rates in addition to the media composition. Comparison of predicted to measured growth rates for the 11 cell lines revealed that non-ecGEMs could only achieve bounded solutions with errors comparable to their enzyme-constrained counterparts if the exchange rates of glucose, lactate, and at least one essential amino acid (threonine, in this case) were specified (Fig. 5D). These results also highlight an important feature of the enzyme-constraint framework: The greatest advantages and improvement in flux predictions are achieved when experimental exchange rates are limited or unavailable, which is most often the case when modeling human systems. However, when such flux measurements are available, the potential improvement offered by enzyme constraints becomes limited, as illustrated in the most constrained simulation in Fig. 5D.

The ability to estimate metabolic fluxes and growth rates with reasonable accuracy through the integration of enzyme constraints with Human1 represents a substantial development in human metabolic modeling. Whereas previous applications of human GEMs have largely been restricted to network-based analyses, the enzyme constraint formulation enables simulation-based approaches in the absence of metabolite exchange information.

DISCUSSION

We developed Human1, a systematically curated and version-controlled human GEM. Human1 is the unification of the parallel HMR and Recon human GEM lineages and effectively represents HMR3 and Recon4 with the aim of consolidating scientific efforts into a more efficient and coordinated approach to modeling human metabolism. We used Human1 to compare metabolic network structure and function across different healthy tissue and tumor types and demonstrated improved reliability of gene essentiality predictions for human cells; Human1 furthermore enables accurate simulation of cell growth and metabolite exchange rates given limited flux information.

The value of the rigorous curation process that was applied to Human1 is exemplified in part by the improved performance in gene essentiality predictions compared to other human GEMs (Fig. 4, B and C). These improvements can be attributed to the integration of enzyme complex information from multiple models and databases into Human1 followed by careful curation of gene-reaction associations. The development of Human1 extended beyond gene-reaction associations and gene essentiality analyses, including an extensive mass and energy balancing process, yielding a 100% stoichiometrically consistent GEM with more than 99% mass-balanced reactions. Furthermore, the quantification of these metrics over the curation process (fig. S4, A to D) enabled us to link various operations to changes in model performance or quality. This can help others identify where to focus efforts when applying this procedure to another organism or system, particularly if they are interested in improving one or two specific metrics.

An important feature of GEM-based analyses is that GEMs allow for simulation of flux through a metabolic network, enabling prediction of growth rates and intracellular reaction fluxes. Traditional

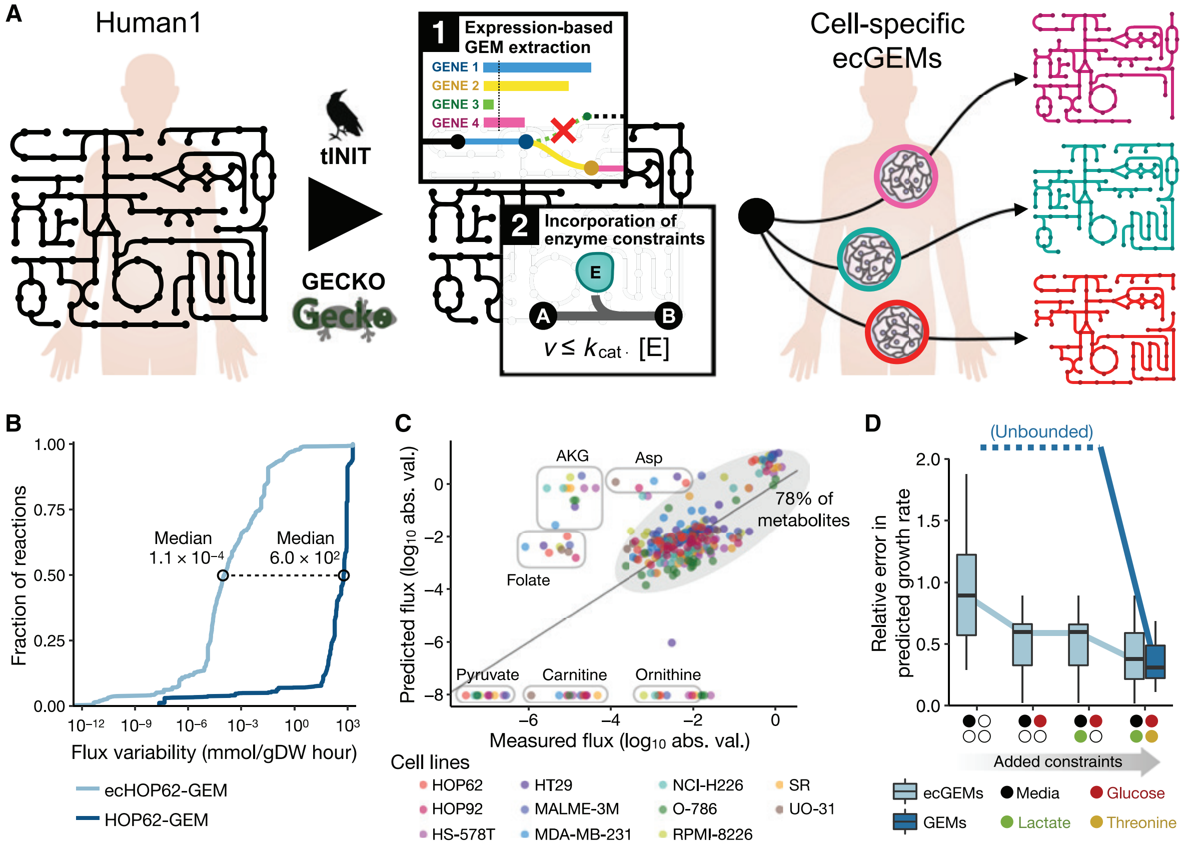


Fig. 5. Generation and analysis of human ecGEMs. (**A**) Graphical representation of the pipeline used to construct NCI-60 cell line–specific ecGEMs from Human1. (**B**) Cumulative distribution of flux variability among reactions in HOP62-GEM and ecHOP62-GEM. Only the ~3200 reactions that carried a flux of $>10^{-8}$ mmol/gDW hour when optimizing biomass production in HOP62-GEM were included in the plot. Distributions for all 11 cell lines are shown in fig. S12. (**C**) Comparison of predicted with measured exchange fluxes ($\log_{10}$-transformed absolute flux values) for the 11 cell-specific ecGEMs, where only the set of metabolites present in the growth medium (Ham's medium) was specified. Different colored markers represent the different cell lines. Metabolites whose fluxes were systematically under- or overpredicted among the different models are labeled in circles, whereas the other ~78% lie within the shaded oval. Note that metabolites along the bottom of the plot have a predicted flux of zero but are shown here as having the absolute minimum measured value to avoid logarithm of zero. (**D**) Boxplots showing the relative error in predicted growth rate for the 11 cell-specific ecGEMs and non-ecGEMs. "Unbounded" indicates that the solutions are effectively unbounded and therefore have unquantifiable (infinite) error. Colored markers on the *x* axis denote the exchange constraints that were cumulatively added to the models when making predictions. "Media" indicates that only the metabolites present in the growth medium were specified, without constraining their exchange rates. "Glucose," "Lactate," and "Threonine" indicate that the exchange flux for those metabolites in the model was constrained to the measured value.

simulations of human GEMs involve specifying external parameters (such as metabolite uptake rates) and internal parameters (such as specific growth rate and internal flux splits) to capture metabolic phenotypes, particularly in cancer (*52*). Measurements to determine these parameters in vivo are challenging or currently impossible, resulting in poorly constrained flux predictions and hindering the ability of GEMs to describe human metabolism where it matters most—within humans. In this work, we presented the construction and analysis of human ecGEMs, which integrate enzyme kinetics and optionally proteomic data to allow physiologically meaningful flux simulations given little or no metabolite exchange information (*51*). This formalism enables flux simulations by specifying internal model constraints using more readily available omics data rather than defining external model constraints based on metabolite exchange rates, greatly expanding the application potential of Human1, particularly for modeling metabolism of tissues and tumors in vivo.

As a complement to Human1, we developed the Metabolic Atlas web portal. This portal supplements and enriches the features of Human1 by providing users with deeper information on model components (for example, listing all reactions involving a given metabolite) and links to external databases (such as HPA, Ensembl, and MetaNetX). Metabolic Atlas also offers interactive compartment and subsystem maps to visualize and navigate Human1 content. By presenting the content in a more visual and connected format, Metabolic Atlas unlocks the information and potential of Human1 for those who are unfamiliar with GEMs but are interested in human metabolism.

Although GEMs provide versatile tools for the exploration of metabolism, their value is contingent upon their quality. Researchers rely on GEMs to be meticulously curated and frequently updated to ensure that they are consistent with current knowledge. Furthermore, this process should be done in a manner that is open, systematic, and reproducible. We therefore constructed Human1 in a version-controlled GitHub repository (https://github.com/SysBioChalmers/Human-GEM), where its latest iteration (v1.3.0 at the time of writing) and complete history are publicly available. This formulation allows the implementation of improvements and repairs to the model on the order of days to weeks, rather than several months to years as is the case with traditional GEM releases. We expect this or analogous approaches to become common practice in GEM development because the rapid progress of the field requires a model development framework that can keep pace while maintaining transparency and reproducibility.

SUPPLEMENTARY MATERIALS

stke.sciencemag.org/cgi/content/full/13/624/eaaz1482/DC1
Materials and Methods
Fig. S1. The evolution of generic human GEMs.
Fig. S2. Replication of infant growth simulation using Human1.
Fig. S3. Memote report screenshot for Human1.
Fig. S4. Human1 quality and performance over the curation process.
Fig. S5. Labeled 2D tSNE projection of tissue- and tumor-specific GEM reaction content comparison based on Hamming similarity.
Fig. S6. Visualization of altered proline metabolism in CHOL using Metabolic Atlas.
Fig. S7. Visualization of increased expression in fatty acid beta oxidation subsystems for LAML using Metabolic Atlas.
Fig. S8. Enrichment of true positives in model-predicted essential genes.
Fig. S9. Comparison of gene essentiality predictions among the three reference GEMs and their 621 derivative cell line models with CRISPR knockout screen results from the DepMap database.
Fig. S10. Impact of gene essentiality threshold on DepMap gene essentiality analysis results.
Fig. S11. Gene essentiality predictions when considering only biomass production compared to considering the activity of 57 different metabolic tasks.
Fig. S12. Effect of enzyme constraints on GEM flux variability.
Table S1. Comparison of generic human GEM statistics.
Table S2. Issue-guided model curation workflow implemented on the Human-GEM GitHub repository.
Table S3. Summary of model changes associated with each version of Human-GEM.
Data file S1. Composition of the generic human cell biomass reaction.
Data file S2. Average fatty acid composition for the curation of lipid metabolism.
Data file S3. Metabolic tasks required for cellular viability.
Data file S4. FVA of ecGEMs.
Data file S5. NCI-60 cell line experimental exchange fluxes.
References (*56*–*72*)

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. R. J. DeBerardinis, C. B. Thompson, Cellular metabolism and disease: What do metabolic outliers teach us? *Cell* **148**, 1132–1144 (2012).
2. B. Ghesquiere, B. W. Wong, A. Kuchnio, P. Carmeliet, Metabolism of stromal and immune cells in health and disease. *Nature* **511**, 167–176 (2014).
3. A.-H. M. Emwas, R. M. Salek, J. L. Griffin, J. Merzaban, NMR-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. *Metabolomics* **9**, 1048–1072 (2013).
4. E. A. Day, R. J. Ford, G. R. Steinberg, AMPK as a therapeutic target for treating metabolic diseases. *Trends Endocrinol. Metab.* **28**, 545–560 (2017).
5. P. Dey, J. Baddour, F. Muller, C. C. Wu, H. Wang, W. T. Liao, Z. Lan, A. Chen, T. Gutschner, Y. Kang, J. Fleming, N. Satani, D. Zhao, A. Achreja, L. Yang, J. Lee, E. Chang, G. Genovese, A. Viale, H. Ying, G. Draetta, A. Maitra, Y. A. Wang, D. Nagrath, R. A. DePinho, Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. *Nature* **542**, 119–123 (2017).
6. C. H. Johnson, J. Ivanisevic, G. Siuzdak, Metabolomics: Beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
7. U. Sauer, Metabolic networks in motion: 13C-based flux analysis. *Mol. Syst. Biol.* **2**, 62 (2006).
8. C. B. Milne, P. J. Kim, J. A. Eddy, N. D. Price, Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnol. J.* **4**, 1653–1670 (2009).
9. M. A. Oberhardt, B. Ø. Palsson, J. A. Papin, Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320 (2009).
10. N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, B. Ø. Palsson, Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1777–1782 (2007).
11. H. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, I. Goryanin, The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol. Syst. Biol.* **3**, 135 (2007).
12. I. Thiele, N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bolling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novere, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. G. M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, B. Ø. Palsson, A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **31**, 419–425 (2013).
13. E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Drager, N. Mih, F. Gatto, A. Nilsson, G. A. Preciat Gonzalez, M. K. Aurich, A. Prlić, A. Sastry, A. D. Danielsdottir, A. Heinken, A. Noronha, P. W. Rose, S. K. Burley, R. M. T. Fleming, J. Nielsen, I. Thiele, B. O. Palsson, Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat. Biotechnol.* **36**, 272–281 (2018).
14. A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, I. Nookaew, P. Jacobson, A. J. Walley, P. Froguel, L. M. Carlsson, M. Uhlen, J. Nielsen, Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol. Syst. Biol.* **9**, 649 (2013).
15. A. Mardinoglu, R. Agren, C. Kampf, A. Asplund, M. Uhlen, J. Nielsen, Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 3083 (2014).
16. L. Väremo, I. Nookaew, J. Nielsen, Novel insights into obesity and diabetes through genome-scale metabolic modeling. *Front. Physiol.* **4**, 92 (2013).
17. A. Mardinoglu, S. Shoaie, M. Bergentall, P. Ghaffari, C. Zhang, E. Larsson, F. Backhed, J. Nielsen, The gut microbiota modulates host amino acid and glutathione metabolism in mice. *Mol. Syst. Biol.* **11**, 834 (2015).
18. G. Bidkhori, R. Benfeitas, M. Klevstig, C. Zhang, J. Nielsen, M. Uhlen, J. Boren, A. Mardinoglu, Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11874–E11883 (2018).
19. S. Lee, C. Zhang, Z. Liu, M. Klevstig, B. Mukhopadhyay, M. Bergentall, R. Cinar, M. Ståhlman, N. Sikanic, J. K. Park, S. Deshmukh, A. M. Harzandi, T. Kuijpers, M. Grøtli, S. J. Elsässer, B. D. Piening, M. Snyder, U. Smith, J. Nielsen, F. Backhed, G. Kunos, M. Uhlen, J. Boren, A. Mardinoglu, Network analyses identify liver-specific targets for treating liver diseases. *Mol. Syst. Biol.* **13**, 938 (2017).
20. E. M. Blais, K. D. Rawls, B. V. Dougherty, Z. I. Li, G. L. Kolling, P. Ye, A. Wallqvist, J. A. Papin, Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. *Nat. Commun.* **8**, 14250 (2017).
21. A. Nilsson, A. Mardinoglu, J. Nielsen, Predicting growth of the healthy infant using a genome scale metabolic model. *Npj Syst. Biol. Appl.* **3**, 3 (2017).
22. C. Lieven, M. E. Beber, B. G. Olivier, F. T. Bergmann, M. Ataman, P. Babaei, J. A. Bartell, L. M. Blank, S. Chauhan, K. Correia, C. Diener, A. Dräger, B. E. Ebert, J. N. Edirisinghe, J. P. Faria, A. M. Feist, G. Fengos, R. M. T. Fleming, B. García-Jiménez, V. Hatzimanikatis, W. van Helvoirt, C. S. Henry, H. Hermjakob, M. J. Herrgard, A. Kaafarani, H. U. Kim, Z. King, S. Klamt, E. Klipp, J. J. Koehorst, M. König, M. Lakshmanan, D. Y. Lee, S. Y. Lee, S. Lee, N. E. Lewis, F. Liu, H. Ma, D. Machado, R. Mahadevan, P. Maia, A. Mardinoglu, G. L. Medlock, J. M. Monk, J. Nielsen, L. K. Nielsen, J. Nogales, I. Nookaew, B. O. Palsson, J. A. Papin, K. R. Patil, M. Poolman, N. D. Price, O. Resendis-Antonio, A. Richelle, I. Rocha, B. J. Sanchez, P. J. Schaap, R. S. M. Sheriff, S. Shoaie, N. Sonnenschein, B. Teusink, P. Vilaca, J. O. Vik, J. A. H. Wodke, J. C. Xavier, Q. Yuan, M. Zakhartsev, C. Zhang, MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* **38**, 372–276 (2020).
23. A. Bordbar, J. M. Monk, Z. A. King, B. O. Palsson, Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* **15**, 107–120 (2014).
24. M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, A. Ruepp, CORUM: The comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
25. J. Y. Ryu, H. U. Kim, S. Y. Lee, Framework and resource for more than 11,000 gene-transcript-protein-reaction associations in human metabolism. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E9740–E9749 (2017).
26. S. Moretti, O. Martin, T. V. Tran, A. Bridge, A. Morgat, M. Pagni, MetaNetX/MNXref—Reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44**, D523–D526 (2016).
27. H. Z. Lu, F. R. Li, B. J. Sanchez, Z. M. Zhu, G. Li, I. Domenzain, S. Marcisauskas, P. M. Anton, D. Lappa, C. Lieven, M. E. Beber, N. Sonnenschein, E. J. Kerkhoven, J. Nielsen, A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10**, 3586 (2019).
28. V. Stodden, M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. A. Ioannidis, M. Taufer, Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241 (2016).
29. H. Wang, S. Marcisauskas, B. J. Sánchez, I. Domenzain, D. Hermansson, R. Agren, J. Nielsen, E. J. Kerkhoven, RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Comput. Biol.* **14**, e1006541 (2018).
30. N. Pornputtapong, I. Nookaew, J. Nielsen, Human metabolic atlas: An online resource for human metabolism. *Database* **2015**, bav068 (2015).
31. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm,

P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Ponten, Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

32. R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, J. Nielsen, Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.* **10**, 721 (2014).

33. L. J. Carithers, K. Ardlie, M. Barcus, P. A. Branton, A. Britton, S. A. Buia, C. C. Compton, D. S. DeLuca, J. Peter-Demchok, E. T. Gelfand, P. Guan, G. E. Korzeniewski, N. C. Lockhart, C. A. Rabiner, A. K. Rao, K. L. Robinson, N. V. Roche, S. J. Sawyer, A. V. Segre, C. E. Shive, A. M. Smith, L. H. Sobin, A. H. Undale, K. M. Valentino, J. Vaught, T. R. Young, H. M. Moore; GTEx Consortium, A novel approach to high-quality postmortem tissue procurement: The GTEx project. *Biopreserv. Biobank.* **13**, 311–319 (2015).

34. J. Hu, J. W. Locasale, J. H. Bielas, J. O'Sullivan, K. Sheahan, L. C. Cantley, M. G. Vander Heiden, D. Vitkup, Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nat. Biotechnol.* **31**, 522–529 (2013).

35. W. Liu, A. Le, C. Hancock, A. N. Lane, C. V. Dang, T. W. Fan, J. M. Phang, Reprogramming of proline and glutamine metabolism contributes to the proliferative and metabolic responses regulated by oncogenic transcription factor c-MYC. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 8983–8988 (2012).

36. J. J. Tanner, S. M. Fendt, D. F. Becker, The proline cycle as a potential cancer therapy target. *Biochemistry* **57**, 3433–3444 (2018).

37. M. M. Simile, P. Bagella, G. Vidili, A. Spanu, R. Manetti, M. A. Seddaiu, S. Babudieri, G. Madeddu, P. A. Serra, M. Altana, P. Paliogiannis, Targeted therapies in cholangiocarcinoma: Emerging evidence from clinical trials. *Medicina* **55**, E42 (2019).

38. M. Maher, J. Diesch, R. Casquero, M. Buschbeck, Epigenetic-transcriptional regulation of fatty acid metabolism and its alterations in leukaemia. *Front. Genet.* **9**, 405 (2018).

39. Z. Wang, L. Wen, X. Ma, Z. Chen, Y. Yu, J. Zhu, Y. Wang, Z. Liu, H. Liu, D. Wu, D. Zhou, Y. Li, High expression of lactotriaosylceramide, a differentiation-associated glycosphingolipid, in the bone marrow of acute myeloid leukemia patients. *Glycobiology* **22**, 930–938 (2012).

40. V. Gouaze-Andersson, M. C. Cabot, Glycosphingolipids and drug resistance. *Biochim. Biophys. Acta* **1758**, 2096–2103 (2006).

41. K. H. Lin, A. Xie, J. C. Rutter, Y. R. Ahn, J. M. Lloyd-Cowden, A. G. Nichols, R. S. Soderquist, T. R. Koves, D. M. Muoio, N. J. MacIver, J. K. Lamba, T. S. Pardee, C. M. McCall, D. A. Rizzieri, K. C. Wood, Systematic dissection of the metabolic-apoptotic interface in AML reveals heme biosynthesis to be a regulator of drug sensitivity. *Cell Metab.* **29**, 1217–1231.e7 (2019).

42. Y. Fukuda, Y. Wang, S. L. Lian, J. Lynch, S. Nagai, B. Fanshawe, A. Kandilci, L. J. Janke, G. Neale, Y. P. Fan, B. P. Sorrentino, M. F. Roussel, G. Grosveld, J. D. Schuetz, Upregulated heme biosynthesis, an exploitable vulnerability in MYCN-driven leukemogenesis. *JCI Insight* **2**, 92409 (2017).

43. L. Salerno, G. Romeo, M. N. Modica, E. Amata, V. Sorrenti, I. Barbagallo, V. Pittalá, Heme oxygenase-1: A new druggable target in the management of chronic and acute myeloid leukemia. *Eur. J. Med. Chem.* **142**, 163–178 (2017).

44. A. R. Joyce, B. O. Palsson, Predicting gene essentiality using genome-scale in silico models. *Methods Mol. Biol.* **416**, 433–457 (2008).

45. T. Hart, M. Chandrashekhar, M. Aregger, Z. Steinhart, K. R. Brown, G. MacLeod, M. Mis, M. Zimmermann, A. Fradet-Turcotte, S. Sun, P. Mero, P. Dirks, S. Sidhu, F. P. Roth, O. S. Rissland, D. Durocher, S. Angers, J. Moffat, High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).

46. J. Masel, M. L. Siegal, Robustness: Mechanisms and consequences. *Trends Genet.* **25**, 395–403 (2009).

47. B. W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).

48. M. Ghandi, F. W. Huang, J. Jane-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald III, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, K. Hu, A. Y. Andreev-Drakhlin, J. Kim, J. M. Hess, B. J. Haas, F. Aguet, B. A. Weir, M. V. Rothberg, B. R. Paolella, M. S. Lawrence, R. Akbani, Y. Lu, H. L. Tiv, P. C. Gokhale, A. de Weck, A. A. Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J. M. Korn, D. A. Porter, M. D. Jones, J. Golji, G. Caponigro, J. E. Taylor, C. M. Dunning, A. L. Creech, A. C. Warren, J. M. McFarland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y. E. Maruvka, A. D. Cherniack, A. Tsherniak, F. Vazquez, J. D. Jaffe, A. A. Lane, D. M. Weinstock, C. M. Johannessen, M. P. Morrissey, F. Stegmeier, R. Schlegel, W. C. Hahn, G. Getz, G. B. Mills, J. S. Boehm, T. R. Golub, L. A. Garraway, W. R. Sellers, Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).

49. R. M. Meyers, J. G. Bryan, J. M. McFarland, B. A. Weir, A. E. Sizemore, H. Xu, N. V. Dharia, P. G. Montgomery, G. S. Cowley, S. Pantel, A. Goodale, Y. Lee, L. D. Ali, G. Jiang, R. Lubonja, W. F. Harrington, M. Strickland, T. Wu, D. C. Hawes, V. A. Zhivich, M. R. Wyatt, Z. Kalani, J. J. Chang, M. Okamoto, K. Stegmaier, T. R. Golub, J. S. Boehm, F. Vazquez, D. E. Root, W. C. Hahn, A. Tsherniak, Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).

50. D. J. Cook, J. Nielsen, Genome-scale metabolic models applied to human health and disease. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **9**, e1393 (2017).

51. B. J. Sanchez, C. Zhang, A. Nilsson, P. J. Lahtvee, E. J. Kerkhoven, J. Nielsen, Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).

52. D. C. Zielinski, N. Jamshidi, A. J. Corbett, A. Bordbar, A. Thomas, B. O. Palsson, Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism. *Sci. Rep.* **7**, 41241 (2017).

53. M. Jain, R. Nilsson, S. Sharma, N. Madhusudhan, T. Kitami, A. L. Souza, R. Kafri, M. W. Kirschner, C. B. Clish, V. K. Mootha, Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science* **336**, 1040–1044 (2012).

54. A. Nilsson, J. R. Haanstra, B. Teusink, J. Nielsen, Metabolite depletion affects flux profiling of cell lines. *Trends Biochem. Sci.* **43**, 395–397 (2018).

55. R. Mahadevan, C. H. Schilling, The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).

56. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

57. R. Caspi, R. Billington, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P. E. Midford, Q. Ong, W. K. Ong, S. Paley, P. Subhraveti, P. D. Karp, The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).

58. Z. A. King, J. Lu, A. Drager, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, N. E. Lewis, BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515–D522 (2016).

59. D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, P. Flicek, Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

60. L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdottir, J. Wachowiak, S. M. Keating, V. Vlasov, S. Magnusdottir, C. Y. Ng, G. Preciat, A. Zagare, S. H. J. Chan, M. K. Aurich, C. M. Clancy, J. Modamio, J. T. Sauls, A. Noronha, A. Bordbar, B. Cousins, D. C. El Assal, L. V. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. B. Guebila, A. Kostromins, N. Sompairac, H. M. Le, D. Ma, Y. K. Sun, L. Wang, J. T. Yurkovich, M. A. P. Oliveira, P. T. Vuong, L. P. El Assal, I. Kuperstein, A. Zinovyev, H. S. Hinton, W. A. Bryant, F. J. A. Artacho, F. J. Planes, E. Stalidzans, A. Maass, S. Vempala, M. Hucka, M. A. Saunders, C. D. Maranas, N. E. Lewis, T. Sauter, B. Ø. Palsson, I. Thiele, R. M. T. Fleming, Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).

61. B. J. Sanchez, F. Li, E. J. Kerkhoven, J. Nielsen, SLIMEr: Probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework. *BMC Syst. Biol.* **13**, 4 (2019).

62. P. R. Shorten, G. C. Upreti, A mathematical model of fatty acid metabolism and VLDL assembly in human liver. *Biochim. Biophys. Acta* **1736**, 94–108 (2005).

63. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular Biology of the Cell* (Garland Science, ed. 5, 2008).

64. D. X. Wang, B. Eraslan, T. Wieland, B. Hallstrom, T. Hopf, D. P. Zolg, J. Zecha, A. Asplund, L. H. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, B. Kuster, A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).

65. K. Sheikh, J. Forster, L. K. Nielsen, Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnol. Prog.* **21**, 112–121 (2005).

66. J. O. Park, S. A. Rubin, Y. F. Xu, D. Amador-Noguez, J. Fan, T. Shlomi, J. D. Rabinowitz, Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nat. Chem. Biol.* **12**, 482–489 (2016).

67. N. Mih, E. Brunk, K. Chen, E. Catoiu, A. Sastry, E. Kavvas, J. M. Monk, Z. Zhang, B. O. Palsson, ssbio: A Python framework for structural systems biology. *Bioinformatics* **34**, 2155–2157 (2018).

68. S. A. Becker, B. Ø. Palsson, Context-specific metabolic networks are consistent with experiments. *PLOS Comput. Biol.* **4**, e1000082 (2008).

69. R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, J. Nielsen, Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLOS Comput. Biol.* **8**, e1002518 (2012).

70. A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni, M. Ceccarelli, G. Bontempi, H. Noushmehr, TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).

71. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2018).
72. N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, K. K. Weitz, R. Eils, R. Konig, R. D. Smith, B. O. Palsson, Omic data from evolved E-coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**, 390 (2010).

Acknowledgments: We thank S. Huang and J. Boekel for contributions to the metabolicatlas.org web portal. The results published here are in whole or part based on data generated by the TCGA Research Network: www.cancer.gov/tcga. The GTEx Project was supported by the Common Fund of the Office of the Director of the NIH and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. **Funding:** Research reported in this publication was supported by funding from the Knut and Alice Wallenberg Foundation, the National Cancer Institute of the NIH under award number F32CA220848 (J.L.R.), the Novo Nordisk Foundation (grant no. NNF10CC1016517), and the European Union's Horizon 2020 research and innovation program under grant agreement no. 720824 (I.D.). **Author contributions:** Conceptualization: J.L.R., P.K., H.W., P.-E.C., E.J.K., and J.N.; methodology: J.L.R., P.K., H.W., P.-E.C., D.C., A.N., R.F., and I.D.; software: J.L.R., P.K., H.W., P.-E.C., D.C., A.N., M.A., R.F., I.D., V.B., A.L., A.H., L.H.; investigation: J.L.R., P.K., H.W., P.-E.C., D.C., A.N., M.A., R.F., I.D., and J.G.; writing (original draft): J.L.R., P.K., H.W., P.-E.C., D.C., M.A., R.F., and I.D.; writing (review and editing): E.J.K., L.T.S., B.O.P., A.M., L.H., M.U., and J.N.; supervision: A.M., L.T.S., B.O.P., M.U., and J.N.; project administration: J.N.; funding acquisition: M.U. and J.N. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Human1 and its associated functions, scripts, and data are available on GitHub (https://github.com/SysBioChalmers/Human-GEM), and Metabolic Atlas is available at www.metabolicatlas.org. The Human1 GEM-PRO, as well as all cell-specific GEMs and enzyme-constrained ecGEMs described in this study, and the datasets and scripts necessary to reproduce the models, analyses, and figures are available on Zenodo (https://doi.org/10.5281/zenodo.3577466). All other data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials.

Submitted 15 August 2019
Accepted 10 March 2020
Published 24 March 2020
10.1126/scisignal.aaz1482

Citation: J. L. Robinson, P. Kocabaş, H. Wang, P.-E. Cholley, D. Cook, A. Nilsson, M. Anton, R. Ferreira, I. Domenzain, V. Billa, A. Limeta, A. Hedin, J. Gustafsson, E. J. Kerkhoven, L. T. Svensson, B. O. Palsson, A. Mardinoglu, L. Hansson, M. Uhlén, J. Nielsen, An atlas of human metabolism. *Sci. Signal.* **13**, eaaz1482 (2020).

THE HPA community

The Human Protein Atlas is the result of contributions from many researchers, each providing data and input to the different parts of the database. Here is a partial list of more than 600 researchers who contributed to the creation of the atlas.

Annica Åbergh, Csaba Adori, Delaram Afshari, Lotta Agaton, Margret Agnarsdottir, Inger Åhlen, Lavina Ahmed, Matilda Ahnfelt, Emelie Ahnfelt, Lovisa Åkesson, Kalle Alanen, Groom Alemayehu, Cajsa Älgenäs, Nuzhat Ali, Eva Allerbring, Tove Alm, Ylva Almqvist, Ozlem Altay, Rose-Marie Amini, Bahram Amini, Elisabet Andersen, Ann-Catrin Andersson, Helene Andersson, Sandra Andersson, Per Henrik Andersson, Eni Andersson, Philip Andersson, Pia Angelidou, Hayrie Aptula, Angela Arokianathan, Maria Aronsson, Caroline Asplund, Anna Asplund, Roxana Astefanei, Ulrika Axelsson, Burcu Ayoglu, Rana Aziz, Julie Bachmann, Thomas Backlund, Max Backman, Carina Backman, Anna Bäckström, John Ballew, Piotr Banski, Sophie Barbaud, Laurent Barbe, Swapnali Barde, Galyna Bartish, Shaghayegh Bayati, Annika Bendes, Rui Benfeitas, Susanna Berg, Marie Berg, Ellinor Backlin Bergh, Sofia Berglind, Jacob Berglund, Julia Bergman, Hanna Bergman, Kristina Bergström, Sofia Bergström, Holger Berling, Maria Berling, Anna Berling, Bhavana Bharambe, Arivarasan Arasan Bharati, Faranak Bidad, Gholamreza Bidkhori, Elin Birgersson, Kaj Bjelkenkrantz, Sara Björk, Lars Björk, Marcus Gry Björklund, Maria Björklund, Erik Björling, Lisa Björling, Magnus Bjursell, Jonatan Blader, Hammou Ait Blal, Maria Arone Blanco, Jenny Blomqvist, Anna Bofin, Anna Bohlin, Paula Borg, Lisa Borggren, Jesper Borin, Tove Boström, Henning Boström, Johan Botling, Carl-Fredrik Bowin, Bela Bozoky, Sara Brännström, Johan Bredenberg, Lucas Bremer, Lisa Bremer, Christer Busch, Sanna Byström, Annelie Cajander, Malin Cammenberg, Tove Canerstam, Simon Cannava, Oana Carja, Sandra Silgård Casell, Karim Cassiminjee, Dijana Cerjan, Anthony Cesnik, Sushama Chandekar, Shuqi Chen, Barnik Choudhury, Birger Christensson, Maddie Ciszewska, Anna Maria Clementz, Parag Dabir, Leo Dahl, Lars-Göran Dahlgren, Matilda Dale, Pontus Danforth, Frida Danielsson, Angelika Danielsson, Hanna Danielsson, Melanie Dannemeyer, Spyros Darmanis, Issra Dawi, Anthony Decay, Anna-Maria Denes, Atul Deshmukh, Leyla Ali Dholey, Isabella Diaz, Andreas Digre, Soraya Djerbi, Miroslav Djokic, Dijana Djureinovic, Tea Dodig-Crnkovic, Lela Ali Doley, Anca Dragomir, Sascha Drews, Kimi Drobin, Naila Durrani, Jens Durruthy-Durruthy, Philip Dusart, Malin Ebbinge, Fredrik Edfors, Elsa Edlund, Karolina Edlund, Per-Henrik Edqvist, Maria Edvardsson, Åsa Edvinsson, Åsa Ehlén, Sara Ek, Siri Ekblad, Karin Elmén, Adila Elobeid, Hanna Emanuelsson, Linnea Enge, Sara Engström, Henric Enstedt, Ronny Falk, Katharina Ericson, Robin Eriksson, Cecilia Eriksson, Amanda Eriksson, Karin Ernberg, Henrik Everberg, Linn Fagerberg, Ronny Falk, August Jernbom Falk, Jenny Fall, Crystal Marian Farhat, Erik Fasterius, Kalle von Feilitzen, Olivia Feldt, Siri Flemming, Mattias Forsberg, Björn Forsström, Claudia Fredolini, Mikaela Friedman, Priti Fulgaonkar, Jesper Gantelius, Emil Gillberg, Christian Gnann, Bharat Godhke, Sanjay Gohil, Lili Gong, Leonardo Gottlob, Charles Goussu, Susanne Gräslunf, Torbjörn Gräslund, Gabriela Gremel, Lars Grimelius, Adrian Gronowski, Nathalie Hou Grün, Emma Grundell, Albin Grundstrom, Jeanette Grundström, Karolin Guldevall, Kristoffer Gumbel, Anna Gundberg, Sara Andersson Gunnerås, Julian Gur, Sofie Gustafsson, Jonas Gustafsson, Anna Häggmark, Asif Halimi, Anneli Halldin, Inga Hallin, Max Hallqvist, Hans Hamberg, Frank Hammar, Marica Hamsten, Carl Hamsten, Carl Hamsten, Marianne Hansson, Christofer Harris, Hanna Hassan, Ragna Häussler, Ida Hedberg, Geeta Hegde, Neda Hekmati, Cecilia Hellström, Christine Hemming, Frauke Henjes, Görel Hercules, Basia Hjelm, Martin Hjelmare, Peter Hlavcak, Sophia Hober, Andreas Hober, Jonathan Hober, Anna Höfges, Alexander Holm, Linnea Holmdahl, Tomas Hökfelt, Eckart Holtz, Mun-Gwan Hong, Ida Hossar, Francis Jingxin Hu, Paul Hudson, Emma Hultman, Fabiana Hyle, Maria Jesus Iglesias, Abhijeet Ingle, Tayyebeh Jafari, Karin Jakobsson, Liv Jakobsson, Louise Jansson, Annlouise Jansson, Josefine Jaxby, Gabriella Jensen, Gabriella Jensen, Tony Jiménez-Beristain, Karin Jirström, Henrik Johannesson, Anna Johansson, Sebastian Johansson, Fredric Johansson, Alex Johansson, Marica Merkel Johansson, Frida Henningson Johnson, Martina Jones, Pallavi Jonnalagadda, Maansi Joshi, Cecilia Juhlin, David Just, John Juter, Samuel Kääriä, Jay M. Kaimal, Lillemor Källström, Caroline Kampf, Sara Kanje, Maximilian Karlander, Josefine Karlsson, Max Karlsson, Borbala Katona, Dennis Kesti, Naila Khan, Wasif Ali Khan, Sania Kheder, Irina Kim, Daniel Klevebring, Anna Konrad, Suresh Kothari, David Kotol, Malin Kronqvist, Madeleine Kronqvist, Laura Kugel, Eléne Kunze, Eugenia Kuteeva, Nathalie Lager, Karin Larsson, Erik Larsson, Albin Holmberg Larsson, Ida Larsson, Louise Larsson, Magnus Larsson, Julius Lautenbach, Trang Le, Sunjae Lee, Heidi Lemström, Magnus Lennartsson, Isabelle Lilja, Agnieszka Limiszewska, Erik Lindahl, Sarah Lindbo, Klas Linderbäck, Anders Lindgren, Matilda Lindgren, Johan Lindholm, Anton Lindqvist, Isa Lindqvist, Mats Lindskog, Cecilia Lindskog, Hanna Lindberg, Robert Lindström, Emma Lindström, Sara Lindström, Jessica Lindström, Emil Lindström, Anna-Karin Lindström, Jerker Linné, Zhengtao Liu, Oscar Ljungqvist, John Löfbom, Lucia Lourido, Alen Lovric, Jan Lund, Erika Lundahl, Emma Lundberg, Joakim Lundeberg, Katarzyna Lundmark, Magnus Lundqvist, Veronica Lundström, Aruna Madan, Gianluca Maddalo, Kristina Magnusson, Diana Mahdessian, Fakhry Mahrous, Vikas Maindad, Åsa Makower, Swapna Mali, Erik Malm, Katarina Malm, Magdalena Malm, Mikael Malmqvist, Anna Månberg, Manjusha, Vidya Manohar, Adil Mardinoglu, Marta Martin, Anna Martinez, Kristina Mate, Cecilia Mattson, Emma Mattsson, Jonna Mattsson, Steffen Matz, Loren Mear, Edith van der Meijden, Lisa Meijer, Andreas Metousis, Artur Mezheyeuski, Patrick Micke, Cecilia Mikaelsson, Maria Mikus, Nicholas Mitsios, Sofia Moberg, Sara Moberg, Archita Mohanty, Chitralekha Mohanty, Atefeh Mohsenchian, Kristian Moller, Johan Mölne, Avin Monazzami, Mona Moradi, Nasim Moradi, Katie Morse, Jan Mulder, Jenny Mullen, Muna Muse, Khayrun Nahar, Åsa Nässtrӧm, Sanjay Navani, Maja Neiman, Hero Nikdin, Jens Nielsen, Markella Nikolopoulou, Peter Nilsson, Kenneth Nilsson, Anders Nilsson, Anders Nilsson, Mia Nilsson, Feria Hikmet Noraddin, Leo Nore, Isabella Norell, Hans Norlinder, Etienne-Nicholas Nyaiesh, Linda Nyberg, Emma Nygren, Malte Nygren, Gillian O'Hurley, Katja Obieglo, Jacob Odeberg, Jenny Ödling, Camilla Ohlsson, Per Oksvold, Sofie Olander, Tommie Olofsson, Jennie Olofsson, Markus Olofsson, Ingmarie Olsson, Natalia Orlowicz, Carolina Oses, Linda Oskarsson, Martin Östberg, Judie Östling, Wei Ouyang, Paulina Ozimek, Linda Paavilainen, Nerea Pajares, Alexandra Palmqvist, Nitin Pardule, Ida Parisi, Tushar Patil, Anna Perols, Anja Persson, Johanna Persson, Julia Persson, Lukas Persson, Nadezhda Petseva, Jennie Petterson, Erik Pettersson, Linnea Pettersson, Philippa Pettingill, Elisa Pin, Charles Pineau, Fredrik Pontén, Felix Pontén, Victor Pontén, Anna Porwit-MacDonald, Nusa Pristovsek, Ulrika Qundos, Mammar Rahmani, Sonika Rai, Margareta Ramström, Marc Rassy, Shailendra Rathod, Ronia Razavi, Venkatesh Chandra Reddy, Julia Remnestål, Philippa Reuterswärd, Elton Rexhepaj, Axel Riese, Rebecca Rimini, Erik Ringström, Johan Rockberg, Dorines Rosario, Sixten Rosenfeldt, Marcus Runeson, Urban Ryberg, Mehri Salahi, Selama Salh, Laura Sanchez-Rivera, Mats Sandberg, Birgitta Sander, Sonali Saraf, Sanem Sariyar, Aishe Sarshad, Julia Scheffel, Rutger Schutten, Jochen Schwenk, Johan Seijsing, Olof Serrander, Suzan Shalan, Jaekyung Shin, Shahrzad Shirazifard, Natasa Sikanic, Tom Simpson, Inna Sitnik, Åsa Sivertsson, Ronald Sjöberg, Anna Sjöberg, Anders Sjöland, Evelina Sjöstedt, Lovisa Skoglund, Marie Skogs, Anna Sköllermo, Cecilia Smedberg, Philip Smith, Ester Soderling, Charlotte Soläng, David Solomon, Ebba Gideon Sörman, Charlotte Stadler, Stefan Ståhl, Petter Stanley, Johanna Steen, Anna Stenius, Maria Stenvall, Fredrik Sterky, Kristin Stirm, Rebecka Stockgard, Ann-Sofi Strand, Linnea Strandberg, Sara Strömberg, Devin Sullivan, Mårten Sundberg, Christer Sundström, Stina Sundström, Andreas Svanström, Torbjorn Sveds, Anne-Sophie Svensson, Jenny Svensson, Per-Olof Syrén, Cristina Al-Khalili Szigyarto, Sofie Taberman, Jenny O Takanen, Helena Tåquist, Abdellah Tebani, Hanna Tegel, Gabriella Tekin, Niklas Thalen, Lina Thelander, Josefin Thelander, Cecilia Engel Thomas, Peter Thul, Manuel de la Torre, Samuel Tourle, Veronika Treiber, Surya Tripathi, Christian Tryggvason, Beste Turanli, Mathias Uhlén, Marie Utterback, Andrea Varadi, Shailesh Vartak, Ellen Vartia, Josefin Viking, Raghvendra Viswakarma, Anna-Luisa Volk, Helian Vunk, Jimmy Vuu, Eva Wahlund, Jacob Wakter, Pia Waldenbäck, Björn Wållberg, Jinghong Wan, Alkwin Wanders, Daniella Wei, Beata Werne, Henrik Wernérus, Joakim Westberg, Kenneth Wester, Tommy Wester, Cornelia Westerberg, Malin Westin, Anna Westring, Allison Whalen, Jens Widehammar, Mikaela Wiking, Viktoria Wiking, Sandra Wikström, Helena Willen, Björn Winckler, Casper Winsnes, Valtteri Wirta, Meike de Wit, Lisa Wolff, Ulla Wrethagen, Lan Lan Xu, Hao Xu, Cane Yaka, Anna Ybo, Louise Yderland, Vijay Yelane, Jamil Yousef, Feifan Yu, Jiaying Yu, Aisha Yusuf, Pawel Zajac, Arash Zandian, Cheng Zhang, Tianyu Zheng, Wen Zhong, Agata Zieba, Martin Zwahlen.

The HPA annual meeting in 2019. Held on the island of Gotland in Sweden.

Acknowledgments

We acknowledge all of the HPA teams for their hard work during the 20 years since the start of the HPA effort. More than 1,000 person-years have been invested to create this open-access public database, which has over 300,000 visitors per month from more than 150 countries. Special thanks also to the contributors to the Digital Collection, including Mattias Karlén (illustrations), Mathias Uhlén (project management, text and editing), Inger Åhlén (administration), Åsa Sivertsson (text and editing), Evelina Sjöstedt (text and editing), Emma Lundberg (text), Stefan Ståhl (text), Marianne Hansson (text), Fredrik Pontén (text), Kalle von Feilitzen (text), Caroline Kampf (text), Peter Nilsson (text), Cristina Al-Khalili (text), Jochen Schwenk (text), Fredrik Edfors (text), Cecilia Lindskog (text), Andreas Digre (text), Sophia Hober (text), Adil Mardinoglu (text), Max Karlsson (text), Jan Mulder (text), and Jens Nielsen (text).

We are grateful to the Knut and Alice Wallenberg Foundation for being the main funder of the HPA program and to the HPA Research Foundation for funding the Digital Collection.

About the Knut and Alice Wallenberg Foundation

The Knut and Alice Wallenberg Foundation is the largest private financier of research in Sweden and also one of the largest in Europe. The foundation's aim is to benefit Sweden by supporting basic research and education, mainly in medicine, technology, and the natural sciences. The foundation can also initiate grants for strategic projects and scholarship programs. For more information, see kaw.wallenberg.org.

About the Science for Life Laboratory

Science for Life Laboratory (SciLifeLab) is a research institution for the advancement of molecular biosciences in Sweden. SciLifeLab started out in 2010 as a joint effort between four universities: Karolinska Institutet, KTH Royal Institute of Technology, Stockholm University, and Uppsala University. The center provides access to a variety of advanced infrastructures in life science for thousands of researchers, creating a unique environment for health and environmental research at the highest level. For more information, see www.scilifelab.se.

About Karolinska Institutet

The vision of Karolinska Institutet (KI) is to advance knowledge about life and strive toward better health for all. As a university, KI is Sweden's largest center of medical academic research and offers the country's widest range of medical courses and programs. Since 1901, the Nobel Assembly at Karolinska Institutet has selected the Nobel laureates in Physiology or Medicine. For more information, see ki.se/en.

About KTH Royal Institute of Technology

Since its founding in 1827, KTH Royal Institute of Technology in Stockholm has grown to become Sweden's largest technical research and learning institution. For more information, see www.kth.se/en.

About Uppsala University

Uppsala University is the Nordic region's oldest university—founded in 1477—and is divided into three disciplinary domains: humanities and social sciences, medicine and pharmacy, and science and technology. These in turn comprise nine faculties and nearly 50 departments in total. For more information, see www.uu.se/en.

FOLLOW US!

@ProteinAtlas

@humanproteinatlas

@human-protein-atlas

@humanproteinatlas

VISIT US!

HPA

www.proteinatlas.org