ARTICLE

doi:10.1038/nature13302

A draft map of the human proteome

Min-Sik Kim[1,2], Sneha M. Pinto[3], Derese Getnet[1,4], Raja Sekhar Nirujogi[3], Srikanth S. Manda[3], Raghothama Chaerkady[1,2], Anil K. Madugundu[3], Dhanashree S. Kelkar[3], Ruth Isserlin[5], Shobhit Jain[5], Joji K. Thomas[3], Babylakshmi Muthusamy[3], Pamela Leal-Rojas[1,6], Praveen Kumar[3], Nandini A. Sahasrabuddhe[3], Lavanya Balakrishnan[3], Jayshree Advani[3], Bijesh George[3], Santosh Renuse[3], Lakshmi Dhevi N. Selvan[3], Arun H. Patil[3], Vishalakshi Nanjappa[3], Aneesha Radhakrishnan[3], Samarjeet Prasad[1], Tejaswini Subbannayya[3], Rajesh Raju[3], Manish Kumar[3], Sreelakshmi K. Sreenivasamurthy[3], Arivusudar Marimuthu[3], Gajanan J. Sathe[3], Sandip Chavan[3], Keshava K. Datta[3], Yashwanth Subbannayya[3], Apeksha Sahu[3], Soujanya D. Yelamanchi[3], Savita Jayaram[3], Pavithra Rajagopalan[3], Jyoti Sharma[3], Krishna R. Murthy[3], Nazia Syed[3], Renu Goel[3], Aafaque A. Khan[3], Sartaj Ahmad[3], Gourav Dey[3], Keshav Mudgal[7], Aditi Chatterjee[3], Tai-Chung Huang[1], Jun Zhong[1], Xinyan Wu[1,2], Patrick G. Shaw[1], Donald Freed[1], Muhammad S. Zahari[2], Kanchan K. Mukherjee[8], Subramanian Shankar[9], Anita Mahadevan[10,11], Henry Lam[12], Christopher J. Mitchell[1], Susarla Krishna Shankar[10,11], Parthasarathy Satishchandra[13], John T. Schroeder[14], Ravi Sirdeshmukh[3], Anirban Maitra[15,16], Steven D. Leach[1,17], Charles G. Drake[16,18], Marc K. Halushka[15], T. S. Keshava Prasad[3], Ralph H. Hruban[15,16], Candace L. Kerr[19]†, Gary D. Bader[5], Christine A. Iacobuzio-Donahue[15,16,17], Harsha Gowda[3] & Akhilesh Pandey[1,2,3,4,15,16,20]

The availability of human genome sequence has transformed biomedical research over the past decade. However, an equivalent map for the human proteome with direct measurements of proteins and peptides does not exist yet. Here we present a draft map of the human proteome using high-resolution Fourier-transform mass spectrometry. In-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, resulted in identification of proteins encoded by 17,294 genes accounting for approximately 84% of the total annotated protein-coding genes in humans. A unique and comprehensive strategy for proteogenomic analysis enabled us to discover a number of novel protein-coding regions, which includes translated pseudogenes, non-coding RNAs and upstream open reading frames. This large human proteome catalogue (available as an interactive web-based resource at http://www.humanproteomemap.org) will complement available human genome and transcriptome data to accelerate biomedical research in health and disease.

Analysis of the complete human genome sequence has thus far led to the identification of approximately 20,687 protein-coding genes[1], although the annotation still continues to be refined. Mass spectrometry has revolutionized proteomics studies in a manner analogous to the impact of next-generation sequencing on genomics and transcriptomics[2–4]. Several groups, including ours, have used mass spectrometry to catalogue complete proteomes of unicellular organisms[5–7] and to explore proteomes of higher organisms, including mouse[8] and human[9,10]. To develop a draft map of the human proteome by systematically identifying and annotating protein-coding genes in the human genome, we carried out proteomic profiling of 30 histologically normal human tissues and primary cells using high-resolution mass spectrometry. We generated tandem mass spectra corresponding to proteins encoded by 17,294 genes, accounting for approximately 84% of the annotated protein-coding genes in the human genome—to our knowledge the largest coverage of the human proteome reported so far. This includes mass spectrometric evidence for proteins encoded by 2,535 genes that have not been previously observed as evidenced by their absence in large community-based proteomic data sets—PeptideAtlas[11], GPMDB[12] and neXtProt[13] (which includes annotations from the Human Protein Atlas[14]).

A general limitation of current proteomics methods is their dependence on predefined protein sequence databases for identifying proteins. To overcome this, we also used a comprehensive proteogenomic analysis strategy to identify novel peptides/proteins that are currently not part of annotated protein databases. This approach revealed novel protein-coding genes in the human genome that are missing from current genome annotations in addition to evidence of translation of several annotated pseudogenes as well as non-coding RNAs. As discussed below, we provide evidence for revising hundreds of entries in protein databases based on our data. This includes novel translation start sites, gene/exon extensions and novel coding exons for annotated genes in the human genome.

Generating a high-quality mass spectrometry data set

To generate a baseline proteomic profile in humans, we studied 30 histologically normal human cell and tissue types, including 17 adult tissues,

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [2]Department of Biological Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. [3]Institute of Bioinformatics, International Tech Park, Bangalore 560066, India. [4]Adrienne Helis Malvin Medical Research Foundation, New Orleans, Louisiana 70130, USA. [5]The Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada. [6]Department of Pathology, Universidad de La Frontera, Center of Genetic and Immunological Studies-Scientific and Technological Bioresource Nucleus, Temuco 4811230, Chile. [7]School of Medicine, Imperial College London, South Kensington Campus, London SW7 2AZ, UK. [8]Department of Neurosurgery, Postgraduate Institute of Medical Education & Research, Chandigarh 160012, India. [9]Department of Internal Medicine Armed Forces Medical College, Pune 411040, India. [10]Department of Neuropathology, National Institute of Mental Health and Neurosciences, Bangalore 560029, India. [11]Human Brain Tissue Repository, Neurobiology Research Centre, National Institute of Mental Health and Neurosciences, Bangalore 560029, India. [12]Department of Chemical and Biomolecular Engineering and Division of Biomedical Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong. [13]Department of Neurology, National Institute of Mental Health and Neurosciences, Bangalore 560029, India. [14]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21224, USA. [15]The Sol Goldman Pancreatic Cancer Research Center, Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [16]Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [17]Department of Surgery, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [18]Departments of Immunology and Urology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland 21231, USA. [19]Department of Obstetrics and Gynecology, Johns Hopkins University School of Medicine Baltimore, Maryland 21205, USA. [20]Diana Helis Henry Medical Research Foundation, New Orleans, Louisiana 70130, USA. †Present address: Department of Biochemistry and Molecular Biology, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA.

7 fetal tissues, and 6 haematopoietic cell types (Fig. 1a). Pooled samples from three individuals per tissue type were processed and fractionated at the protein level by SDS–polyacrylamide gel electrophoresis (SDS–PAGE) and at the peptide level by basic reversed-phase liquid chromatography (RPLC) and analysed on high-resolution Fourier-transform mass spectrometers (LTQ-Orbitrap Elite and LTQ-Orbitrap Velos) (Fig. 1b). To generate a high-quality data set, both precursor ions and higher-energy collisional dissociation (HCD)-derived fragment ions were measured using the high-resolution and high-accuracy Orbitrap mass spectrometer. Approximately 25 million high-resolution tandem mass spectra, acquired from more than 2,000 LC-MS/MS (liquid chromatography followed by tandem mass spectrometry) runs, were searched against NCBI's RefSeq[15] human protein sequence database using the MASCOT[16] and SEQUEST[17] search engines. The search results were rescored using the Percolator[18] algorithm and a total of approximately 293,000 non-redundant peptides were identified at a q value < 0.01 with a median mass measurement error of approximately 260 parts per billion (Extended Data Fig. 1a). The median number of peptides and corresponding tandem mass spectra identified per gene are 10 and 37, respectively, whereas the median protein sequence coverage was approximately 28% (Extended Data Fig. 1b, c). It should be noted, however, that false-positive rates for subgroups of peptide-spectrum matches can vary upon the nature of peptides, such as their size, the charge state of precursor peptide ions or missed enzymatic cleavage (Extended Data Fig. 1d–f and Supplementary Information).

We compared our data set with two of the largest human peptide-based resources, PeptideAtlas and GPMDB. These two databases contain curated peptide information that has been collected from the proteomics community over the past decade. Notably, almost half of the peptides we identified were not deposited in either one of these resources. Also, the novel peptides in our data set constitute 37% of the peptides in PeptideAtlas and 54% of peptides in the case of GPMDB (Extended Data Fig. 1g, h). This marked increase in the coverage of human proteomic data was made possible by the breadth and depth of our analysis as most of the cells and tissues that we analysed have not previously been studied using similar methods. The depth of our analysis enabled us to identify protein products derived from two-thirds (2,535 out of 3,844) of proteins designated as 'missing proteins'[19] for lack of protein-based evidence. Several hypothetical proteins that we identified have a broad tissue distribution, indicating the inadequate sampling of the human proteome thus far (Extended Data Fig. 2a).

Landscape of protein expression in cells and tissues

On the basis of gene expression studies, it is clear that there are several genes that are involved in basic cellular functions that are constitutively expressed in almost all the cells/tissues. Although the concept of 'housekeeping genes' as genes that are expressed in all tissues and cell types is widespread among biologists, there is no readily available catalogue of such genes. Moreover, the extent to which these transcripts are translated into proteins remains unknown. We detected proteins encoded by 2,350 genes across all human cells/tissues with these highly abundant 'housekeeping proteins' constituting approximately 75% of total protein mass based on spectral counts (Extended Data Fig. 2b). The large majority of these highly expressed housekeeping proteins include histones, ribosomal proteins, metabolic enzymes and cytoskeletal proteins. One of the caveats of tissue proteomics is the contribution of vasculature, blood and haematopoietic cells. Thus, proteins designated as housekeeping proteins based on analysis of tissue proteomes could be broadly grouped into two categories, those that are truly expressed in every single cell type and those that are found in every tissue (for example, endothelial cells). Another caveat to be noted here is that some proteins that are indeed expressed in all tissues might not be detected in some of the tissues because of inadequate sampling by mass spectrometry. Thus, this list of housekeeping proteins will continue to be refined as additional in-depth analyses are carried out.

We used a label-free method based on spectral counting to quantify protein expression across cells/tissues. Although more variable as compared

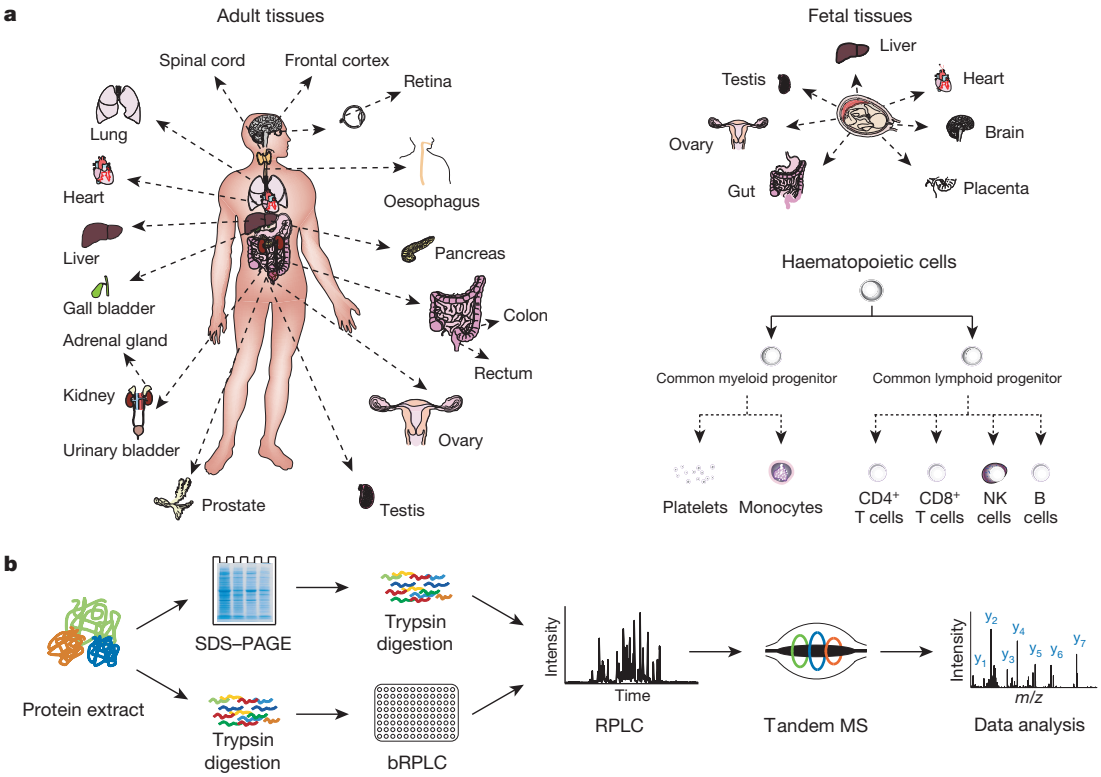


Figure 1 | Overview of the workflow and comparison of data with public repositories. a, The adult/fetal tissues and haematopoietic cell types that were analysed to generate a draft map of the normal human proteome are shown. **b**, The samples were fractionated, digested and analysed on the high-resolution and high-accuracy Orbitrap mass analyser as shown. Tandem mass spectrometry data were searched against a known protein database using SEQUEST and MASCOT database search algorithms.

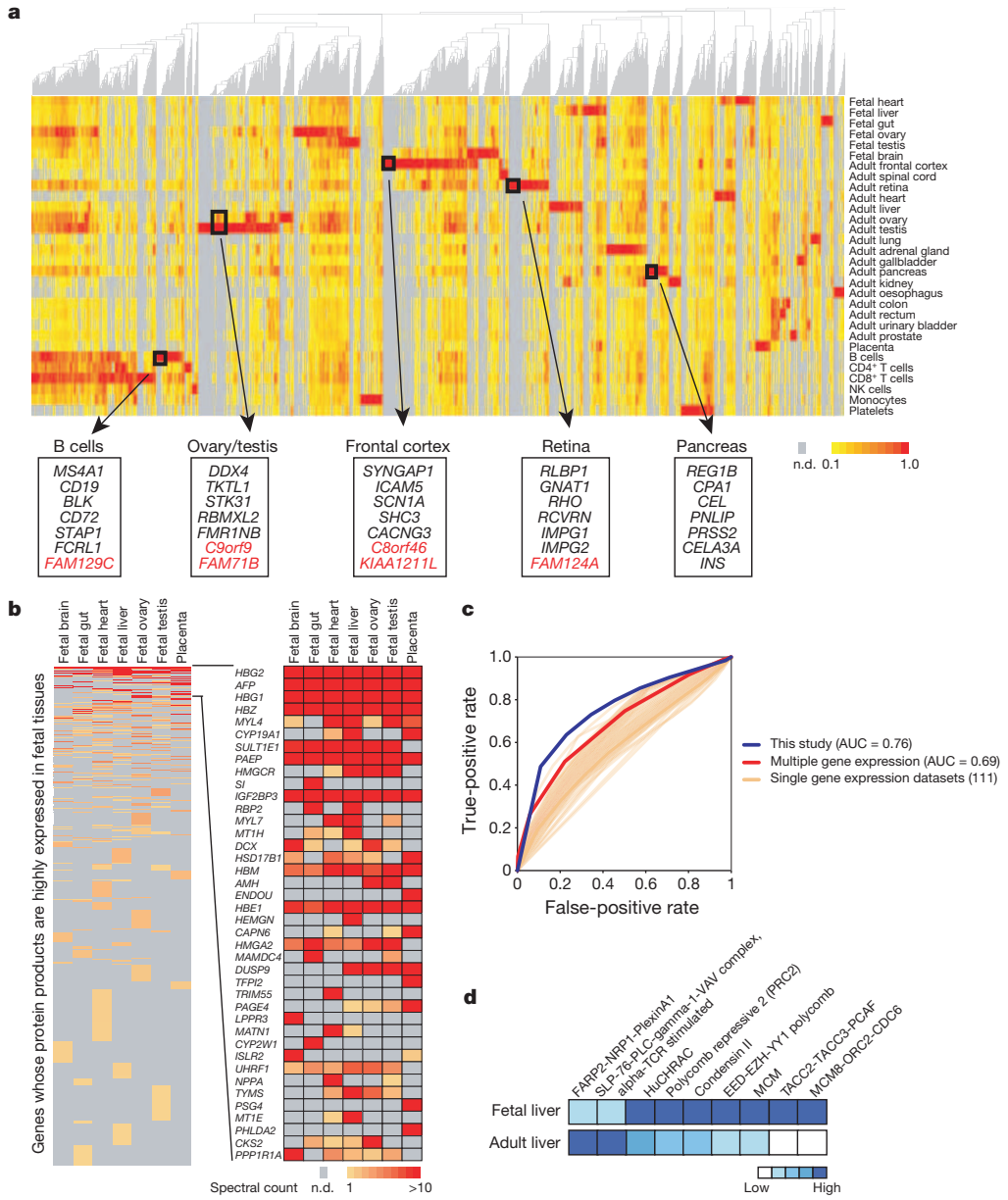


Figure 2 | Landscape of the normal human proteome. **a**, Tissue-supervised hierarchical clustering reveals the landscape of gene expression across the analysed cells and tissues. Selected tissue-restricted genes are highlighted in boxes to show some well-studied genes (black) as well as hypothetical proteins of unknown function (red). The colour key indicates the normalized spectral counts per gene detected across the tissues. n.d., not determined. **b**, A heat map showing tissue expression of fetal tissue-restricted genes ordered by average expression across fetal tissues (left) and a zoom-in of the top 40 most abundant genes (right). The colour key indicates the spectral counts per gene. **c**, An ROC curve showing a comparison of the performance of the current data set (blue, area under the curve (AUC) = 0.762) with 111 individual gene expression data sets (orange) and a composite of the 111 individual data sets (red, AUC = 0.692). **d**, Developmental stage-specific differential expression of protein complexes in fetal and adult liver tissues. Heat map shows protein complexes with less than or equal to half of their subunits expressed in one of the tissue types. The darker the colour, the greater the number of expressed subunits.

to label-based methods, this method is readily applicable to analysis of a large number of samples[8] and has been shown to be reproducible[20]. Supervised hierarchical clustering showed proteins encoded by some genes to be expressed in only a few cells/tissues, whereas others were more broadly expressed (Fig. 2a). Some proteins detected in only one sample were encoded by well-known genes like *CD19* in B cells, *SCN1A* in the frontal cortex and *GNAT1* in the retina, whereas others were encoded by ill-characterized genes. For example, *C8orf46* was expressed in the adult frontal cortex whereas *C9orf9* was expressed in adult ovary and testis. Overall, we detected proteins encoded by 1,537 genes only in one of 30 human samples examined in this study (Extended Data Fig. 2c). These may or may not be tissue-specific genes because of the limit of detection of mass spectrometry and because this analysis did not sample every human cell or tissue type. Because methods based on antibody-based detection can be more sensitive, we performed western blotting experiments to confirm the tissue-restricted expression pattern of some of the proteins against which appropriate antibodies were available. Of 32 proteins tested, eight proteins exhibited a tissue-specific expression in agreement with mass spectrometry-derived data (Extended Data Fig. 3a). Four proteins exhibited a more widespread expression, although in each of these cases extra bands were detected (Extended Data Fig. 3b). In eighteen cases, the antibody did not recognize a protein in the expected size range at all, and no band was detectable in the remaining two cases.

A number of proteins are expressed during development in fetal tissues but not in normal adult tissues. Although earlier studies have focused on a few fetal tissues like fetal brain[21] or liver[22], our study provides the first general survey of the fetal proteome. We detected proteins encoded by 735 genes that are expressed more than tenfold in fetal samples compared to adult tissues/cells. A heat map highlights the expression level of putative fetal-tissue-restricted genes across various fetal tissues (Fig. 2b). The list includes the well-known oncofetal antigens, alpha fetoprotein (*AFP*) and insulin-like growth factor-2 binding protein-3/IGF-2 mRNA binding protein-3 (*IGF2BP3*). High levels of AFP in serum and cerebrospinal fluid are clinically used as biomarkers for neural tube defects, teratoma and yolk sac tumours. Some of the proteins expressed during development in ovary and testis can serve as potential biological markers for identifying cancers of different lineages in the future.

In the past, gene expression profiles across various experimental conditions or tissues have been used to investigate the likelihood of co-expressed genes to physically interact at the protein level[23,24]. We proposed

that protein expression patterns should be a better predictor of protein–protein interactions than gene expression measured at the messenger RNA level. We correlated the expression level for each available protein pair across all 30 cells/tissues in our data using Pearson correlation and compared this to known protein–protein interactions. We then repeated this analysis using correlations obtained from 111 published gene expression data sets. Receiver operating characteristic (ROC) curve analysis clearly shows that data from the human proteome map outperforms that from gene expression profiles for predicting protein–protein interactions, even if all the gene expression data sets are combined and used as a single predictor (Fig. 2c). It should be pointed out, however, that although the use of protein expression data are useful for predicting protein–protein interactions, it is unlikely by itself to be sufficient for such predictions.

Many proteins interact with different partners in different tissues or at different stages of development, although they are not traditionally studied in this fashion. To investigate tissue-restricted expression of protein complex subunits, we evaluated 938 complexes with three or more subunits from the CORUM database[25] and found 679 protein complex subunits showing differential expression in at least one of the 30 tissues. In contrast, there were 34 complexes where all the subunits were expressed concurrently in all tissues. Interestingly, there were 201 instances where differential expression of subunits of complexes was observed across the adult and fetal tissues, indicating that these complexes are dynamic and probably have distinct composition during ontogeny. This dynamic composition is probably related to developmental stage-specific processes in which these complexes are involved. For example, mini-chromosome maintenance (MCM) complex components are highly expressed in fetal liver in contrast to FARP2–NRP1–PLXNA1 complex members, which are highly expressed in the adult liver (Fig. 2d).

Detecting protein isoforms

Alternative splicing gives rise to a large number of splice variants at the RNA level, some of which can encode distinct protein isoforms. Multiple protein isoforms are contributed by only one-third of annotated genes, whereas the remaining two-thirds generate only a single protein product according to the RefSeq database[15] (Extended Data Fig. 2d). Although our primary goal was not to obtain complete coverage of splice isoforms, we identified isoform-specific peptides for 2,861 protein isoforms derived from 2,450 genes. For example, we detected isoform 1 of FYN protein tyrosine kinase in brain and isoform 2 in haematopoietic cells (Fig. 3a). This is significant because although we did not detect the third known isoform of *FYN*, the two isoforms that we identified are known to possess distinct functional properties.

An interactive portal for exploring the human proteome

We have developed a portal for the Human Proteome Map (http://www.humanproteomemap.org) that makes it possible to test and generate hypotheses regarding gene families, protein complexes, signalling pathways, biomarkers, therapeutic targets, immune function and human development. As an illustration, one can explore the protein components of the 20S constitutive proteasome and immunoproteasome complexes (Fig. 3b and Extended Data Fig. 2e). Three of the subunits in 20S constitutive proteasome (PSMB5, PSMB6 and PSMB7; coloured red in Fig. 3b left panel) are known to be replaced by three other subunits (PSMB8, PSMB9 and PSMB10; coloured green) in the 20S immunoproteasome[26]. As shown in the heat map from Human Proteome Map, PSMB8, PSMB9 and PSMB10 are highly expressed in immune cell types.

Novel protein-coding regions in the human genome

The evidence for protein-coding potential is still largely driven by gene prediction algorithms or complementary DNAs and does not routinely include direct detection/measurement of proteins/peptides. Following protein database search of our large-scale proteome analysis, we extracted approximately 16 million MS/MS spectra that did not match currently annotated proteins. These unmatched spectra were searched using a

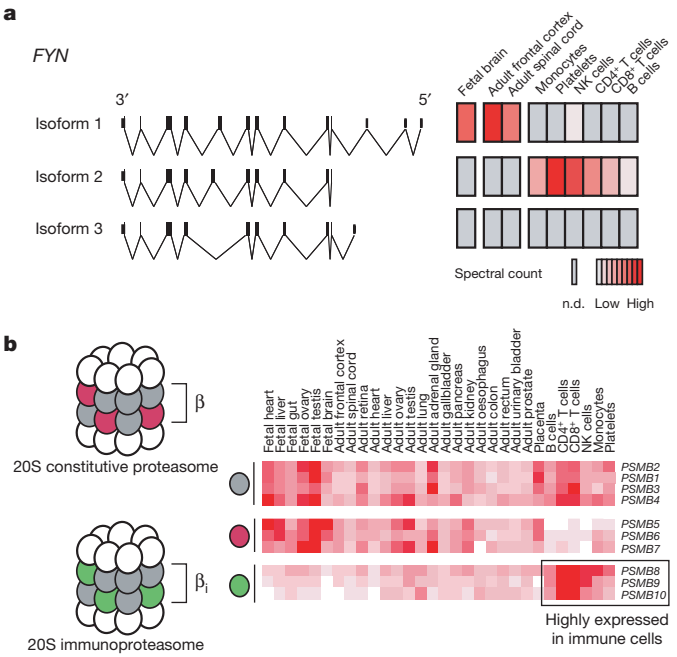


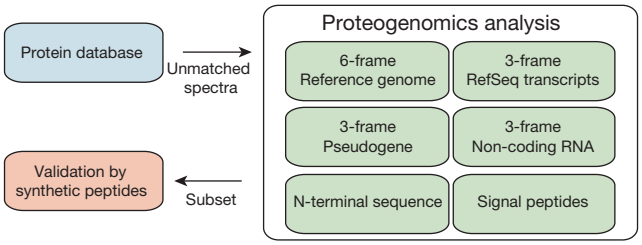
Figure 3 | Isoform-specific expression. **a**, Exon structure of three known isoforms of *FYN* (left) along with abundance of isoform-specific peptides detected in the indicated cells/tissues (right). The colour key indicates a relative expression based on the spectral counts of isoform-specific peptides detected. **b**, 20S constitutive proteasome and 20S immunoproteasome core complexes. Expression of their corresponding components are depicted by a heat map (red indicates higher expression) in the Human Proteome Map portal.

unique proteogenomic analysis strategy developed by our group against conceptually translated human reference genome, RefSeq transcript sequences, non-coding RNAs and pseudogenes. In addition, the data were searched against theoretical protein amino termini and predicted signal sequences (Fig. 4a). As a result, we confirmed translation of 17,294 annotated protein-coding genes that include 4,105 protein N termini, 223,385 exons and 66,947 splice junctions (Fig. 4b). More importantly, we discovered 808 novel annotations of the human genome including translation of 140 pseudogenes, 44 novel ORFs, 106 novel coding regions/exons within annotated gene structures and 110 gene/protein/exon extension events in addition to 198 novel protein N termini and 201 novel signal peptide cleavage sites (Fig. 4b and Supplementary Table 1). Although the peptides were identified at less than 1% false discovery rate (FDR) with parts per billion median mass measurement error (Extended Data Fig. 1i), we carried out an additional level of manual inspection of the MS/MS spectra to reduce false positives in this type of analysis[27]. In addition, because these novel findings are based on matches of tandem mass spectra to translated genomic/transcriptomic sequences, we experimentally confirmed a subset of the identifications from various proteogenomic categories through matching of MS/MS spectra from 98 synthetic peptides with those obtained from analysis of human cells/tissues (Supplementary Data).

Upstream and other ORFs

We identified 28 genes with upstream ORFs (uORFs) where peptides mapped to 5′ untranslated regions (UTRs) of known human RefSeq transcripts. For example, we identified peptides that mapped to the 5′ UTR of solute carrier family 35, member A4 transcript (*SLC35A4*). Although we did not identify any unique peptides from the protein encoded by *SLC35A4*, which is probably related to its low abundance or cell/tissue-specific expression, we identified four peptides derived from the protein encoded by the uORF in 25 out of 30 samples (Extended Data Fig. 5a). This uORF encodes a 103-amino-acid protein that contains a transmembrane region within a putative BNIP3 domain (BCL2/adenovirus

a



b

	Number of genes involved or annotation confirmed /added/altered
Genes whose products were detected	17,294
Confirmed exons	223,385
Confirmed N termini	4,105
Confirmed exon–exon junctions	66,947
Signal peptide cleavage site	329
Confirmation of annotated sites	128
Unannotated cleavage sites	201
Novel protein-coding regions	193
Pseudogenes	140
Non-coding RNAs	9
Upstream ORFs	29
Other ORFs	15
Novel coding regions/exons	106
Novel N termini	198
Gene/protein extensions	70
N-terminal extension	58
C-terminal extension	12
Exon extension	40

Figure 4 | Proteogenomic analysis. **a**, An overview of the multiple databases used in the proteogenomic analysis. A subset of peptides corresponding to genome search-specific peptides were synthesized and analysed by mass spectrometry. **b**, Overall summary of the results from the current study.

E1B 19 kDa protein-interacting protein 3) that is required for mitochondrial localization, suggesting that this protein may play a role in mitochondrial function.

We identified eight cases where we observed peptides that mapped to an ORF located in an alternate reading frame within coding regions of annotated genes. For example, we identified peptides that mapped to a novel ORF of 159 amino acids within the *C11orf48* gene. The protein encoded by the *C11orf48* gene was identified only in the adult retina, although we identified three peptides encoded by the novel ORF from 17 various cells/tissues. We also identified peptide matches to seven ORFs located within 3′UTRs. As an example, a novel ORF comprising of 524 amino acids in the 3′UTR of the *CHTF8* gene was identified on the basis of multiple peptides. The translation initiation site of this novel ORF overlaps the stop codon of the *CHTF8* gene (Extended Data Fig. 4a). Remarkably, the protein encoded by this novel ORF was observed in haematopoietic cells where we did not detect the CHTF8 protein. In addition, this novel gene product was expressed at higher levels in fetal ovary and adult testis than the protein encoded by *CHTF8*. These observations suggest that the translational control for these two proteins encoded by the same gene structure is likely to be different. We also identified a peptide encoded by an ORF within a human endogenous retrovirus (Extended Data Fig. 5b). Domain analysis revealed the presence of a signal peptide at the N terminus along with other domains including Furin-like repeats. In fact, during preparation of this manuscript, a report was published in which this protein was designated as suppressyn and shown to inhibit cell–cell fusion in trophoblast cells[28].

Non-coding RNAs and pseudogenes

We detected several peptides that were not mapped to RefSeq proteins but corresponded to nine transcripts annotated as non-coding RNAs. As shown in Extended Data Fig. 4b, five peptides (including one across an exon-exon junction) were mapped to a region on chromosome 1 that was annotated as a non-coding RNA. Although long non-coding RNAs are, by definition, not supposed to encode proteins[29], our data indicate that some of these encode proteins, which can be discovered by deep proteomic profiling of cells/tissues.

A few studies have recently examined the transcription of pseudogenes on a global scale and found approximately 2,000 pseudogenes to be transcribed either ubiquitously or in a tissue/cancer-specific manner[30,31]. In our data set, we identified more than 200 peptides that are encoded by 140 pseudogenes. These 140 pseudogenes originate from 110 parental genes. To derive this unambiguous set, we filtered out several peptides for which sequence change could be explained by SNPs reported for corresponding genes in dbSNP. When we looked at the tissue distribution of the translated pseudogenes, we observed that roughly half of the pseudogenes were translated in a cell/tissue-restricted manner whereas a small minority was expressed globally (Fig. 5a), a pattern similar to that described for pseudogene transcripts[30,31]. For instance, VDAC1P7 was found to be translated globally (22 of 30 cells/tissues analysed) whereas MAGEB6P1 was detected only in adult testes although its parental gene, *EIF4B*, is widely expressed (Fig. 5a). Extended Data Fig. 5c shows two peptides that map to *MAGEB6P1*, a pseudogene, along with an alignment of the identified peptides with the corresponding peptides in the parental gene. It should be noted that there is still a small chance that some SNPs are not represented in our analysis because we did not analyse the genomes of these individuals. However, this is unlikely to affect our results greatly because it has been estimated that approximately 98% SNPs with 1% frequency have been detected from the 1000 Genomes Project[32].

Protein N termini and signal peptides

We confirmed 4,105 annotated N termini in RefSeq (from 4,132 protein-coding genes) to generate the largest catalogue of experimentally validated translational start sites in humans (~20% of annotated genes). Annotation of start sites in proteins can be imprecise because initiation AUG codons are generally predicted by computation and are prone to errors. For instance, the optimal Kozak consensus sequence (CCACC [AUG]G), which is widely believed to surround the true initiation codon, occurs only in half of the human genes[33]. By searching unmatched MS/MS spectra against customized databases, we identified 3 cases that map upstream and 195 cases that map downstream (within 100 amino acids) of the currently annotated translational initiation sites (Fig. 5b). The nucleotide sequences that surround the AUG codon of these novel translational start sites show the same pattern as bona fide translational start sites (Extended Data Fig. 6), strongly indicating that we have identified true novel or alternative start sites. We also confirmed 201 annotated signal peptide cleavage sites and revised the predictions in 128 other cases (Supplementary Information).

Discussion

Here we present a high-coverage map (covering >84% of human protein-coding genes) of the human proteome using high-resolution mass spectrometry. This demonstrates the feasibility of comprehensively exploring the human proteome on an even larger scale in the near future such as the initiative recently announced by the Human Proteome Organization; the Chromosome-Centric Human Proteome Project[34–36]. Our findings also highlight the need for using direct protein sequencing technologies like mass spectrometry to complement genome annotation efforts. This process can be facilitated by workflows suitable for proteogenomic analyses. Even greater depth can lead to an increase in sequence coverage by employing methodologies such as multiple proteases, direct capture of N termini, enrichment of post-translationally modified peptides, exhaustive fractionation and incorporating alternative sequencing technologies such as electron transfer dissociation (ETD) and top-down mass spectrometry. Finally, these strategies can be further complemented by sampling of individual cell types of human tissues/organs to develop a 'human cell map'. With the availability of both genomic and proteomic landscapes, integrating the information from both resources is likely to accelerate

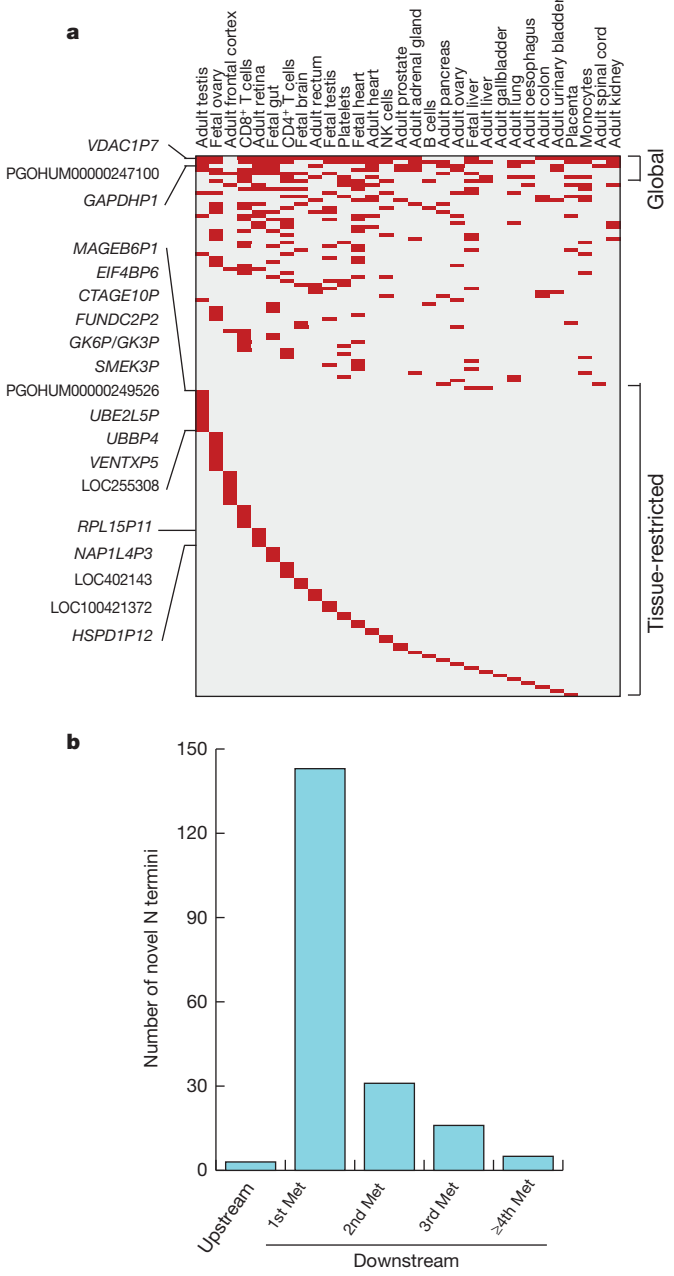


Figure 5 | Translation of pseudogenes and identification of novel N termini. **a**, A heat map shows the expression of pseudogenes (listed by name or accession code number) across the analysed cells/tissues. Some pseudogenes such as *VDAC1P7* and *GAPDHP1* were found to be globally expressed, whereas others were more restricted in their expression or were detected only in a single cell type/tissue as indicated. **b**, The distribution of novel N termini detected with N-terminal acetylation is shown with respect to the location of the annotated translational start site. All sites in the 5′ UTR are labelled upstream whereas those located downstream of the annotated AUG start sites are labelled as 1st Met, 2nd Met and so on.

basic as well as translational research in the years to come through a better understanding of gene-protein-pathway networks in health and disease.

METHODS SUMMARY

Protein was extracted from sorted cells or tissue/organs from three individuals and pooled before processing. The complex mixture of peptides obtained from tryptic digestion of gel bands from SDS–PAGE or basic RPLC fractions was separated on a micro-capillary RPLC column, ionized by nanoflow electrospray ionization and analysed on LTQ-Orbitrap Elite or LTQ-Orbitrap Velos mass spectrometers using HCD fragmentation. Raw mass spectrometry files were processed in Proteome Discoverer and peak lists were first searched against a human RefSeq protein sequence database. The unmatched spectra were subsequently searched against a series of custom polypeptide sequence databases. Only those genome-search-specific peptides that had unique hits in the human genome were investigated further for proteogenomic analysis. A subset of genome-search-specific peptides was generated as synthetic peptides and their fragmentation pattern was compared with corresponding tandem mass spectra identified from proteogenomic analysis. Hierarchical clustering was carried out using normalized spectral counting-based quantification values. Full methods and associated references are available in the online version of this paper.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 August 2013; accepted 31 March 2014.

1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
2. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
3. Bensimon, A., Heck, A. J. & Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annu. Rev. Biochem.* **81,** 379–405 (2012).
4. Cravatt, B. F., Simon, G. M. & Yates, J. R. III. The biological impact of mass-spectrometry-based proteomics. *Nature* **450,** 991–1000 (2007).
5. Nagaraj, N. *et al.* System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. *Mol. Cell. Proteomics* **11,** M111.013722 (2012).
6. Picotti, P. *et al.* A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. *Nature* **494,** 266–270 (2013).
7. Kelkar, D. S. *et al.* Proteogenomic analysis of *Mycobacterium tuberculosis* by high resolution mass spectrometry. *Mol. Cell. Proteomics* **10,** M111.011627 (2011).
8. Huttlin, E. L. *et al.* A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143,** 1174–1189 (2010).
9. Gholami, A. M. *et al.* Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.* **4,** 609–620 (2013).
10. Branca, R. M. *et al.* HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature Methods* **11,** 59–62 (2014).
11. Farrah, T. *et al.* The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.* **12,** 162–171 (2013).
12. Craig, R., Cortens, J. P. & Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3,** 1234–1242 (2004).
13. Gaudet, P. *et al.* neXtProt: organizing protein knowledge in the context of human proteome projects. *J. Proteome Res.* **12,** 293–298 (2013).
14. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature Biotechnol.* **28,** 1248–1250 (2010).
15. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–D763 (2014).
16. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
17. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
18. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **4,** 923–925 (2007).
19. Lane, L. *et al.* Metrics for the human proteome project 2013–2014 and strategies for finding missing proteins. *J. Proteome Res.* **13,** 15–20 (2014).
20. Mosley, A. L. *et al.* Highly reproducible label free quantitative proteomic analysis of RNA polymerase complexes. *Mol. Cell. Proteomics* **10,** M110.000687 (2011).
21. Fountoulakis, M., Juranville, J. F., Dierssen, M. & Lubec, G. Proteomic analysis of the fetal brain. *Proteomics* **2,** 1547–1576 (2002).
22. Ying, W. *et al.* A dataset of human fetal liver proteome identified by subcellular fractionation and multiple protein separation and identification technology. *Mol. Cell. Proteomics* **5,** 1703–1707 (2006).
23. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12,** 37–46 (2002).
24. Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29,** 482–486 (2001).
25. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.* **38,** D497–D501 (2010).
26. Ferrington, D. A. & Gregerson, D. S. Immunoproteasomes: structure, function, and antigen presentation. *Prog. Mol. Biol. Transl. Sci.* **109,** 75–112 (2012).
27. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nature Rev. Mol. Cell Biol.* **5,** 699–711 (2004).
28. Sugimoto, J., Sugimoto, M., Bernstein, H., Jinno, Y. & Schust, D. A novel human endogenous retroviral protein inhibits cell-cell fusion. *Sci. Rep.* **3,** 1462 (2013).
29. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–251 (2013).
30. Kalyana-Sundaram, S. *et al.* Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* **149,** 1622–1634 (2012).
31. Pei, B. *et al.* The GENCODE pseudogene resource. *Genome Biol.* **13,** R51 (2012).

32. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
33. Peri, S. & Pandey, A. A reassessment of the translation initiation codon in vertebrates. *Trends Genet.* **17,** 685–687 (2001).
34. Legrain, P. *et al.* The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **10,** M111.009993 (2011).
35. Paik, Y. K. *et al.* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature Biotechnol.* **30,** 221–223 (2012).
36. Marko-Varga, G., Omenn, G. S., Paik, Y. K. & Hancock, W. S. A first step toward completion of a genome-wide characterization of the human proteome. *J. Proteome Res.* **12,** 1–5 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to acknowledge the National Development and Research Institutes for some of the tissues. We acknowledge the assistance of V. Sandhya, V. Puttamallesh, U. Guha and B. Cole for help with analysis of some of the samples. We thank L. Lane and B. Amos for their assistance with the list of missing genes. This work was supported by an NIH roadmap grant for Technology Centers of Networks and Pathways (U54GM103520), NCI's Clinical Proteomic Tumor Analysis Consortium initiative (U24CA160036), a contract (HHSN268201000032C) from the National Heart, Lung and Blood Institute and the Sol Goldman Pancreatic Cancer Research Center. The authors acknowledge the joint participation by the Adrienne Helis Malvin Medical Research Foundation and the Diana Helis Henry Medical Research Foundation through its direct engagement in the continuous active conduct of medical research in conjunction with The Johns Hopkins Hospital and the Johns Hopkins University School of Medicine and the Foundation's Parkinson's Disease Programs. The analysis work was partially supported by the National Resource for Network Biology (P41GM103504). A.Mah., S.K.Sh., P.S. and T.S.K.P. are supported by DBT Program Support on Neuroproteomics (BT/01/COE/08/05) to IOB and NIMHANS. H.G. is a Wellcome Trust-DBT India Alliance Early Career Fellow. We thank Council of Scientific and Industrial Research, University Grants Commission and Department of Science and Technology, Government of India for research fellowships for S.M.P., R.S.N., A.R., M.K., G.J.S., S.C., P.R., J.S., S.S.M., D.S.K., S.R., S.K.Sr., K.K.D., Y.S., A.S., S.D.Y., N.S., S.A. and G.D.

Author Contributions A.P., H.G., R.C., M.-S.K. designed the study; A.P., H.G., M.-S.K. managed the study; D.G., C.L.K., C.A.I.-D., K.R.M. collected human cells/tissues; M.-S.K., R.C., D.G. developed the pipeline of experiment and analysis; D.G., M.-S.K., S.M.P., K.M., R.C., S.R., J.Z., X.W., P.G.S., M.S.Z., T.-C.H. prepared peptide samples for LC-MS/MS; M.-S.K., R.S.N., S.M.P., R.C., D.S.K., S.R., G.J.S. performed LC-MS/MS; M.-S.K., S.M.P., S.P., S.S.M., C.J.M., J.A. and A.K.M. processed MS data and managed data; A.K.M., S.S.M., B.G., A.H.P., Y.S., M.-S.K. performed comparison analysis with PeptideAtlas, neXtProt and GPMDB; R.I., S.Jai., G.D.B. performed interaction and complex analysis; M.-S.K., S.M.P., S.S.M., P.K., A.K.M., N.A.S., R.S.N., L.B., L.D.N.S., D.S.K., V.N., A.R., T.S., M.K., S.K.Sr., G.D., A.Mar., R.R., S.C., K.K.D., A.S., S.D.Y., S.Jay., P.R., A.H.P., B.G., J.S., N.S., R.G., G.J.S., A.A.K., S.A., D.F., T.S.K.P., H.G., A.P. performed proteogenomic analysis; A.C., H.L., R.S., J.T.S., K.K.M., S.S., A.Mah., S.K.Sh., P.S., S.D.L., C.G.D., A.Mai., M.K.H., R.H.H., C.L.K., C.A.I.-D. assisted with analysis of the data; M.-S.K., S.M.P., T.-C.H., P.L.-R. performed western blot experiments; M.-S.K., J.K.T., A.K.M., B.M., S.P., S.M.P. designed the Human Proteome Map web portal; M.-S.K., A.K.M., J.K.T. generated selected reaction monitoring (SRM) database; M.-S.K., K.M., G.D., S.M.P., S.S.M. illustrated figures with help of other authors; A.P., M.-S.K., H.G. wrote the manuscript with inputs from other authors.

Author Information The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD000561. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.G. (harsha@ibioinformatics.org) or A.P. (pandey@jhmi.edu).

METHODS

Cell and tissue samples. Human tissues were collected post mortem as part of a rapid autopsy program from three adult donors for each tissue by author C.A.I.-D. or obtained from National Disease Research Interchange (NDRI, PA, USA). All tissues were histologically confirmed to be normal before analysis. Fetal tissues were obtained by author C.L.K. Haematopoietic cells were isolated from leukopaks obtained from healthy volunteers participating in routine platelet pheresis. This study was approved by the Johns Hopkins University's Institutional Review Board for use of human tissues and informed consent was obtained from all subjects from whom blood samples were obtained for isolation of haematopoietic cells. Haematopoietic cell populations were isolated sequentially using Miltenyi Biotec magnetic beads from leukopaks in the following order; $CD14^+$ Monocytes, pan $CD56^+$ NK Cells, $CD20^+$ B cells, $CD8^+$ T cells, pan $CD4^+$ T cells as per manufacturer's instructions. In brief, peripheral blood mononuclear cells (PBMCs) were enriched by centrifugation over Ficoll gradient at 700*g* for 15 min. Cells were incubated with magnetic beads in modified MACS buffer (0.05% BSA and 2 mM EDTA) and separated using autoMACS (Miltenyi Biotec). Purity was assessed by flow cytometry. Isolated cells were washed with PBS and stored at −80 °C until use. Platelets were obtained from platelet rich plasma from healthy donors. Platelets were pelleted by centrifugation, resuspended and washed with modified Tyrode's buffer before lysis.

Peptide preparation. The samples were lysed in lysis buffer containing 4% SDS, 100 mM DTT and 100 mM Tris pH 7.5. Tissues were homogenized and sonicated in lysis buffer. The protein concentration of the cleared lysates was estimated using BCA assay and equal amounts of protein from each donor was pooled for further fractionation. Proteins from SDS lysates were separated on SDS–PAGE and in-gel digestion was carried out as described earlier[37]. Briefly, the protein bands were destained, reduced and alkylated and subjected to in-gel digestion using trypsin. The peptides were extracted, vacuum dried and stored at −80 °C until further analysis. The samples (~450 µg proteins) were reduced, alkylated and digested using trypsin overnight at 37 °C. The peptide digests were desalted using Sep-Pak C_{18} columns (Waters Corporation, Milford, MA, USA) and lyophilized. The lyophilized samples were reconstituted in solvent A (10 mM triethylammonium bicarbonate, pH 8.5) and loaded onto XBridge C_{18}, 5 µm 250 × 4.6 mm column (Waters, Milford, MA, USA). The digests were resolved by basic RPLC[38] method using a gradient of 0 to 100% solvent B (10 mM triethylammonium bicarbonate in acetonitrile, pH 8.5) in 50 min. The total fractionation time was 60 min. A total of 96 fractions were collected which were then concatenated to 24 fractions, vacuum dried and stored at −80 °C until further LC-MS analysis.

LC-MS/MS. Peptide samples from in-gel digestion and basic RPLC fractionation were analysed on LTQ-Orbitrap Velos or LTQ-Orbitrap Elite mass spectrometers (Thermo Electron, Bremen, Germany) interfaced with Easy-nLC II nanoflow liquid chromatography systems (Thermo Scientific, Odense, Southern Denmark). The peptide digests from each fraction were reconstituted in Solvent A (0.1% formic acid) and loaded onto a trap column (75 µm × 2 cm) packed in-house with Magic C_{18} AQ (Michrom Bioresources, Inc., Auburn, CA, USA) (5 µm particle size, pore size 100 Å) at a flow rate of 5 µl min^{-1} with solvent A (0.1% formic acid in water). Peptides were resolved on an analytical column (75 µm × 10 cm- on Velos and 20 cm on Elite) at a flow rate of 350 nl min^{-1} using a linear gradient of 7–30% solvent B (0.1% formic acid in 95% acetonitrile) over 60 min. Mass spectrometry analysis was carried out in a data dependent manner with full scans (350–1,800 *m/z*) acquired using an Orbitrap mass analyser at a mass resolution of 60,000 at 400 *m/z* in Velos and 120,000 in Elite at 400 *m/z*. Twenty most intense precursor ions from a survey scan were selected for MS/MS from each duty cycle and detected at a mass resolution of 15,000 at *m/z* of 400 in Velos and 30,000 at *m/z* of 400 in Orbitrap analyser. All the tandem mass spectra were produced by higher-energy collision dissociation (HCD) method. Dynamic exclusion was set for 30 s with a 10 p.p.m. mass window. The automatic gain control for full FT MS was set to 1 million ions and for FT MS/MS was set to 0.05 million ions with a maximum ion injection times of 100 ms and 200 ms, respectively. Internal calibration was carried out using lock-mass from ambient air (*m/z* 445.1200025)[39].

Database searching. MS/MS data obtained from all LC-MS analysis were searched against Human RefSeq database (version 50 containing 33,833 entries along with common contaminants) using SEQUEST and MASCOT (version 2.2) search algorithms through Proteome Discoverer 1.3 platform (Thermo Scientific, Bremen, Germany). The parameters used for data analysis included trypsin as the protease with a maximum of two missed cleavages allowed. Carbamidomethylation of cysteine was specified as a fixed modification and oxidation of methionine, acetylation of protein N termini and cyclization of N-terminal glutamine and alkylated cysteine were included as variable modifications. The minimum peptide length was specified to be 6 amino acids. The mass error was set to 10 p.p.m. for precursor ions and 0.05 Da for fragment ions. The data was also searched against a decoy database and the results used to estimate *q* values using the Percolator algorithm within the Proteome Discoverer suite. Peptides were considered identified at a *q* value <0.01 with a median mass measurement error of approximately 260 parts per billion (Extended Data Fig. 1a). The mass spectrometry data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE[40] partner repository with the data set identifier PXD000561. For assessing the effect of parameters such as size, missed enzymatic cleavage or charge state of precursor peptide ions on FDRs, all PSMs identified at 1% FDR were collated from all fetal tissues and subjected to additional analyses presented in Extended Data Fig. 1f–h (discussed in Supplementary Information).

Mapping peptides onto the human genome. Non-redundant peptide sequences were mapped onto RefSeq protein sequences and the corresponding RefSeq genes. When a peptide was mapped uniquely to a gene but to multiple protein isoforms, the peptide was considered 'gene-specific.' When a peptide was mapped uniquely to a single protein product of a gene, the peptide was considered 'isoform-specific.' Otherwise, peptides were considered 'shared.' To explore gene expression, gene-specific peptides were used. To map the human proteome, all the identified peptides were used. For the isoform expression analysis, isoform-specific peptides were used when multiple isoforms were annotated in RefSeq protein database (version 50).

Spectral count-based quantification. Spectral counts per gene per experiment (for example, SDS–PAGE) were first summed from all peptides mapped to each gene. Total acquired tandem mass spectra were used to normalize between experiments and then spectral counts per gene were averaged across multiple experiments (for example, SDS–PAGE and basic RPLC fractionation) per sample (for example, pancreas). Final averaged spectral counts per gene per sample were used to plot the white-to-red gradient heat map for genes of interest, while summed spectral counts per peptide were used to depict the heat map for all peptides identified and mapped to a gene of interest.

Comparison of human proteome map data with resources. Our data was compared with the neXtProt genes (http://nextprot.org/; release date: 07-16-2013) with expression evidence at protein level. All the peptides from our study were also compared with the peptide databases, GPMDB (http://gpmdb.thegpm.org/; release date: 02-17-2012) and PeptideAtlas (http://www.proteinatlas.org; release: 201308). Processed Illumina Human Body Map 2.0 data for 10 tissues (adrenal glands, ovary, testis, prostate, lung, colon, liver, heart, brain and kidney) were downloaded as individual GTF files from Broad Institute (ftp://ftp.broadinstitute.org/). All the proteins with isoform-specific peptides corresponding to the 10 tissues were mapped to respective mRNA accessions from RefSeq database. The mRNA accessions were used to get the fragments per kilobase per million (FPKM) values of each isoform across tissues from the gene transfer format (GTF) files. These expression values were tabulated and used for comparison with normalized spectral counts of proteins. Normalized spectral count zero was taken as the protein being unidentified in the tissue. Similarly, FPKM of zero was considered for the transcript to be not expressed.

Western blot analysis. Equal amounts of adult tissue protein (25 µg) were separated by sodium dodecyl sulphate-polyacrylamide gel electrophoresis in 4–12% NuPAGE Bis-Tris Precast Gels (Novex, Life Technologies Corporation) and transferred onto nitrocellulose membranes (Whatman, GE Healthcare, Life Science, Pittsburgh, PA, USA) using a semi-dry transfer unit Hoefer TE 70 (Amersham Bioscience). The membranes were blocked with 5% low-fat milk dissolved in phosphate buffered saline containing 0.05% Tween (PBST) for 1 h at room temperature and incubated overnight at 4 °C with primary antibodies. After washing with PBST three times each for 10 min, the membranes were further incubated with the corresponding horseradish peroxidase-conjugated secondary antibodies for 1 h at room temperature. After washing with PBST three times each for 10 min, antibody-bound protein bands were detected with ECL Western Blotting Detection Reagents (RPN 2106V1 and RPN2106V2, GE Healthcare Life Science, Pittsburgh, PA, USA) and photographed with Amersham Hyperfilm ECL autoradiography film (GE Healthcare Life Science, Pittsburgh, PA, USA). Anti-GAPDH antibody was used for a loading control.

Proteogenomic analysis. To enable identification of novel peptides and correction of existing gene annotations in the human genome, we searched six different databases generated in-house using python scripts for proteogenomics analysis. The database used were: (1) six-frame-translated human genome database (2) three-frame-translated RefSeq mRNA sequences (3) three-frame-translated pseudogene database with sequences derived from NCBI and Gerstein's pseudogene database (4) three-frame-translated non-coding RNAs from NONCODE (5) N-terminal peptide database derived from RefSeq mRNA sequences from NCBI and (6) signal peptide database from SignalP and HPRD. The sequences from common contaminants including trypsin were appended to each database. A decoy database was created for each database by reversing the sequences from a target database. Peptide identification was carried out using X!Tandem[41]. The following parameters were common to all searches: precursor mass error set at 10 p.p.m., fragment mass error set at 0.05 Da, carbamidomethylation of cysteine defined as fixed modification, oxidation of methionine defined as variable modification. Only tryptic peptides with up to 2 missed cleavages were considered. Unmatched MS/MS spectra generated from

the protein database search were then searched against these database using X!Tandem search engine installed locally. Specific details about generation of each database, search parameters and post-processing of the search outcome are provided below.

Customization of sequence databases. The human reference genome assembly hg19 was downloaded from NCBI and translated in six reading frames using in-house python scripts. The six-frame translation included stop codon to stop codon translation of the template sequence. Translated peptide sequences smaller than 7 amino acids were not included in the database. Three frame translated mRNA sequence database was created from mRNA sequences downloaded from NCBI RefSeq (RefSeq version 56 containing 33,580 sequences) and translated in three reading frames. Pseudogene sequences downloaded from NCBI (11,160 sequences from NCBI) and Gerstein's pseudogene database (16,881 sequences from http://pseudogene.org/, version 68) were translated into three frames. Similarly, non-coding RNA sequences from NONCODE (91,687 sequences from http://www.noncode.org/NONCODERv3/, version 3) and mRNA sequences from RefSeq version 56 (33,580 sequences) were translated in three reading frames. A custom N-terminal tryptic peptides database was created by fetching all the peptide sequences which begin with methionine and end with K/R from RefSeq mRNA sequences (RefSeq 56). 5′ UTRs of known mRNAs were translated in three reading frames and all possible peptides starting with a methionine until the next tryptic cleavage site were generated. Peptides starting with a methionine until the next tryptic cleavage site from translated protein were also included in the database. Only peptide sequences with up to one missed cleavage and 6–25 amino acids in length were considered. 'Quick acetyl' search option was defined in X!Tandem searches as it allows up to 2 N-terminal amino acids clipping. Signal peptide annotations were obtained from HPRD (http://www.hprd.org) and from predictions by SignalP version 4.0. The signal peptide database was created by fetching N-terminal peptides starting from the cleavage site and extending up to the next tryptic site.

Manual validation of GSSPs. Peptide sequences identified from each of the alternate database searches were filtered and only peptides passing 1% FDR score threshold were considered. Spectra that matched to multiple sequences with equal score were not considered for further analysis. Unique list of genome search specific peptides (GSSPs), pseudogene specific peptides and transcriptome search specific peptides were generated by comparing the unique peptide data with the protein database. The genomic regions to which these peptides mapped were further analysed to identify novel protein coding regions or corrections/alterations to existing annotations. The peptides identified were categorized either as (1) mapping intergenic regions (2) overlapping annotated genes (3) mapping to the intronic regions of existing gene models (4) overlapping annotated genes but translated in alternate reading frame. In case of peptides identified uniquely from pseudogene database search, the amino acid change compared to parent sequence was checked against dbSNP to rule out SNP cases. Peptides identified with at least one mismatch and filtered after comparison with dbSNP were considered for further analysis. The filtered novel peptides were further manually verified to check the confidence of spectral assignment by the search engines. A list of novel peptides identified from this study is available (Supplementary Table 1).

UCSC genome browser[42] (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) was used as visualization tool for manual genome annotation using these novel peptides. Novel peptides, ESTs, ncRNA transcripts, Ensembl RNASeq models and gene prediction models from GENSCAN tracks were enabled on the genome browser. For manual genome annotation, genomic regions where novel peptides mapped were examined and novel genes or changes in gene structures were determined.

Validation of novel findings using synthetic peptides. Peptide sequences were randomly chosen from several categories of proteogenomic analysis and synthesized (JPT Peptide Technologies, Berlin, Germany). Dried peptides were diluted using 0.1% formic acid and approximately 1 pmol of each peptide was separately subjected to LC-MS/MS analysis on the LTQ-Orbitrap mass spectrometers. Fragmentation patterns of 98 peptides were validated manually by comparing them with that identified from proteogenomic analysis (Supplementary Information).

Development of web-based portal. The Human Proteome Map (HPM) portal was developed using the 3-tier web architecture with presentation, application and persistence layers. A MySQL database was used as a persistence layer. PHP and HTML were used for the application and presentation layers, respectively, and AJAX was used for dynamic functionality. Protein and peptide identifications obtained from SEQUEST and MASCOT search engines were converted into MySQL tables. NCBI RefSeq56 annotations were used as standard reference for fetching additional information about the genes from various public resources. Normalized spectral counts were used to represent expression of proteins and peptides. For each peptide, a high resolution MS/MS spectrum from the best scoring identification is shown on the spectrum viewer page using the Lorikeet JQuery plugin (https://code.google.com/p/lorikeet/). The download page allows the user to download the data in standard formats.

Protein–protein interaction (PPI) networks and complexes. PPI networks were generated for each tissue by filtering known interactions from the iRefWeb[43] database (http://wodaklab.org/iRefWeb/search/index) based on tissue spectral count data. If two proteins physically interact and both have spectral count data in a tissue, then the protein interaction is included in the corresponding tissue-specific network. First, proteins identified were mapped to UniProt protein accessions using data from the UniProt, HGNC, and GeneCards databases. 87,027 pairwise and experimental PPIs with confidence score >0 (iRefWeb MINT-like score) were downloaded from iRefWeb. These interactions were used to generate 30 tissue-specific PPI networks. We compared our tissue-specific interaction networks to the CORUM database. There are 938 complexes with three or more subunits in our protein expression data. We evaluated the differential expression of these complexes within corresponding tissue types of adult and fetal samples.

Using protein expression data for PPI prediction. Gene expression data sets were downloaded from GeneMANIA[44]. Only gene expression data sets without any missing values were used. Average Pearson correlation over multiple gene (111) expression data sets was calculated using Fisher's Z transformation.

$$z_i = \frac{1}{2}\ln\left(\frac{1+r_n}{1-r_n}\right)$$

where r_n is the Pearson correlation between two genes for the nth experiment. Then, an appropriate estimate of the true mean is calculated as,

$$z_n = N^{-1}\sum_{i=1}^{N} z_i$$

where N is the total number of experiments. Then, by inversion, average correlation is calculated as,

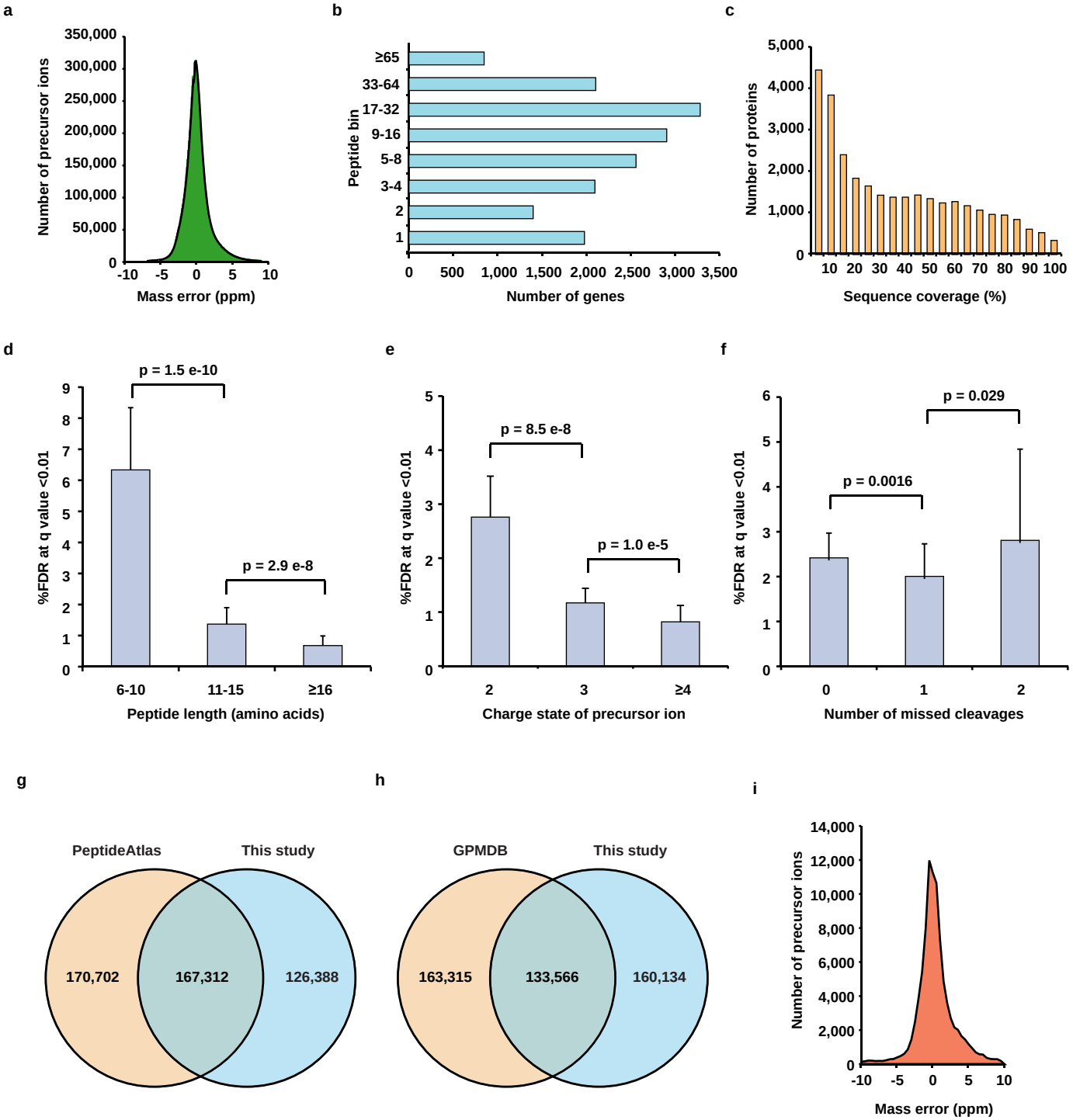
$$R = \frac{e^{2z_n}-1}{e^{2z_n}+1}$$

Human proteome data was filtered by removing all genes with zero spectral count in more than 26 tissue samples. Pearson correlation values lie within the range $[-1, +1]$. They were converted to fall between [0, 1] using a logistic function for both human proteome and gene expression data sets. A positive set of 5,523 protein–protein interactions was downloaded from iRefWeb with MINT-like confidence score >0.9 and where human proteome and gene expression information was available for both interaction participants. A negative set of an equal number of randomly selected protein pairs not part of the complete iRefWeb interaction set was created to compute ROC statistics.

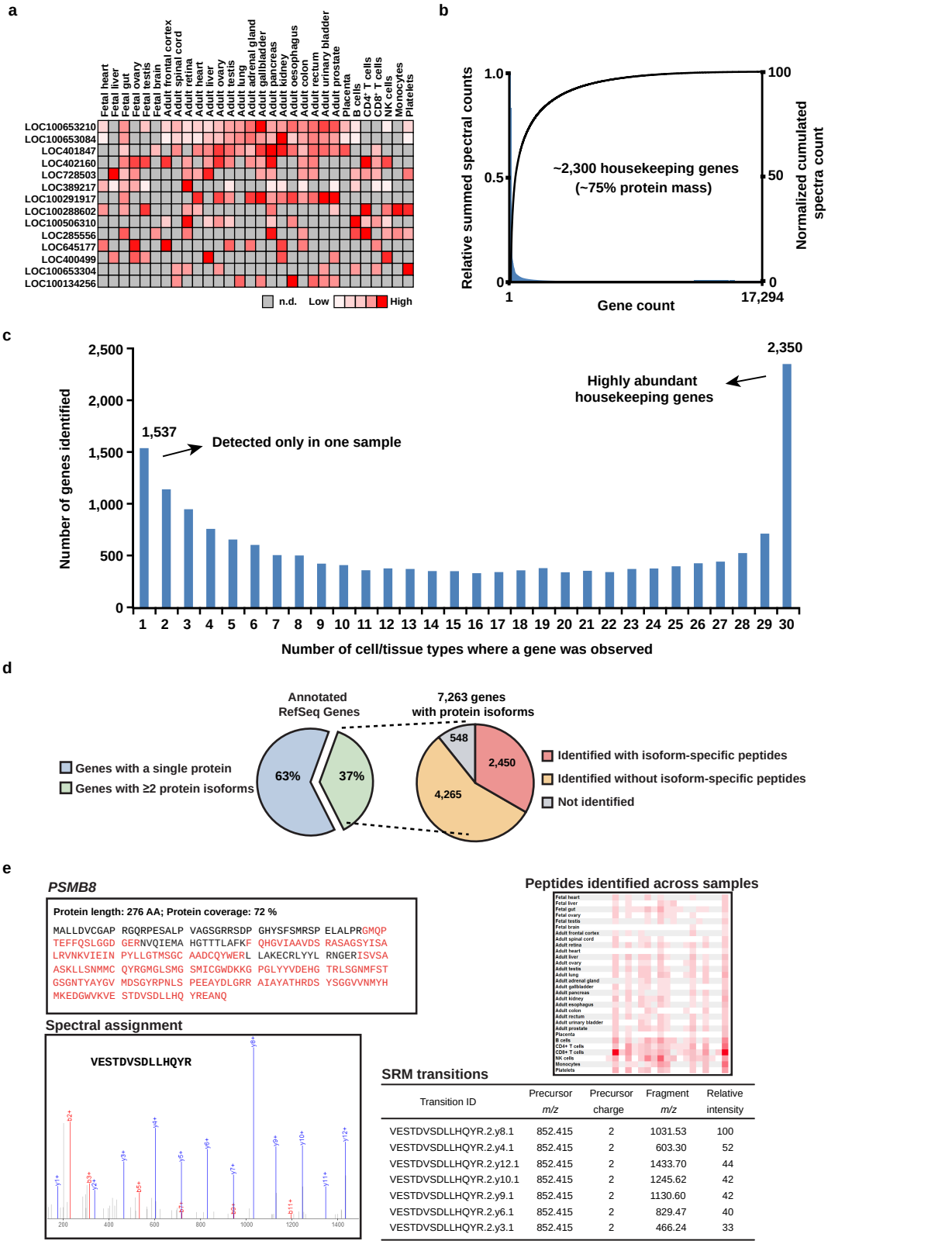
Tissue-supervised clustering of proteome expression across cells/tissues. Hierarchical clustering explorer (version 3.5, http://www.cs.umd.edu) was used to create tissue-supervised clustering using genes whose protein products were identified. Normalized/averaged spectral counts across cells/tissues were loaded onto the software and spectral counts per gene were further normalized by rescaling from 0 to 1 (0 being non-detected; 1 being the highest spectral counts). When clustering, average linkage and Pearson's correlation were set. Colour gradient was used; grey being non-detected, yellow being the lowest and red being the highest expression.

Visualization of proteogenomic results using the NGS module in GeneSpring. GeneSpring 12.6 was used to visualize the GSSPs against the human reference genome hg19. The GSSPs were loaded as a region list into GeneSpring. RefSeq annotation was downloaded from the Agilent server. All GTF files were loaded onto GeneSpring and the tracks were visualized in a genome browser view using a custom colouring scheme. Individual images were exported in JPEG format.

37. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols* **1,** 2856–2860 (2007).
38. Wang, Y. *et al.* Reversed-phase chromatography with multiple fraction concatenation strategy for proteome profiling of human MCF10A cells. *Proteomics* **11,** 2019–2026 (2011).
39. Olsen, J. V. *et al.* Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **4,** 2010–2021 (2005).
40. Vizcaíno, J. A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41,** D1063–D1069 (2013).
41. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467 (2004).
42. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* **41,** D64–D69 (2013).
43. Razick, S., Magklaras, G. & Donaldson, I. M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9,** 405 (2008).
44. Zuberi, K. *et al.* GeneMANIA prediction server 2013 update. *Nucleic Acids Res.* **41,** W115–W122 (2013).

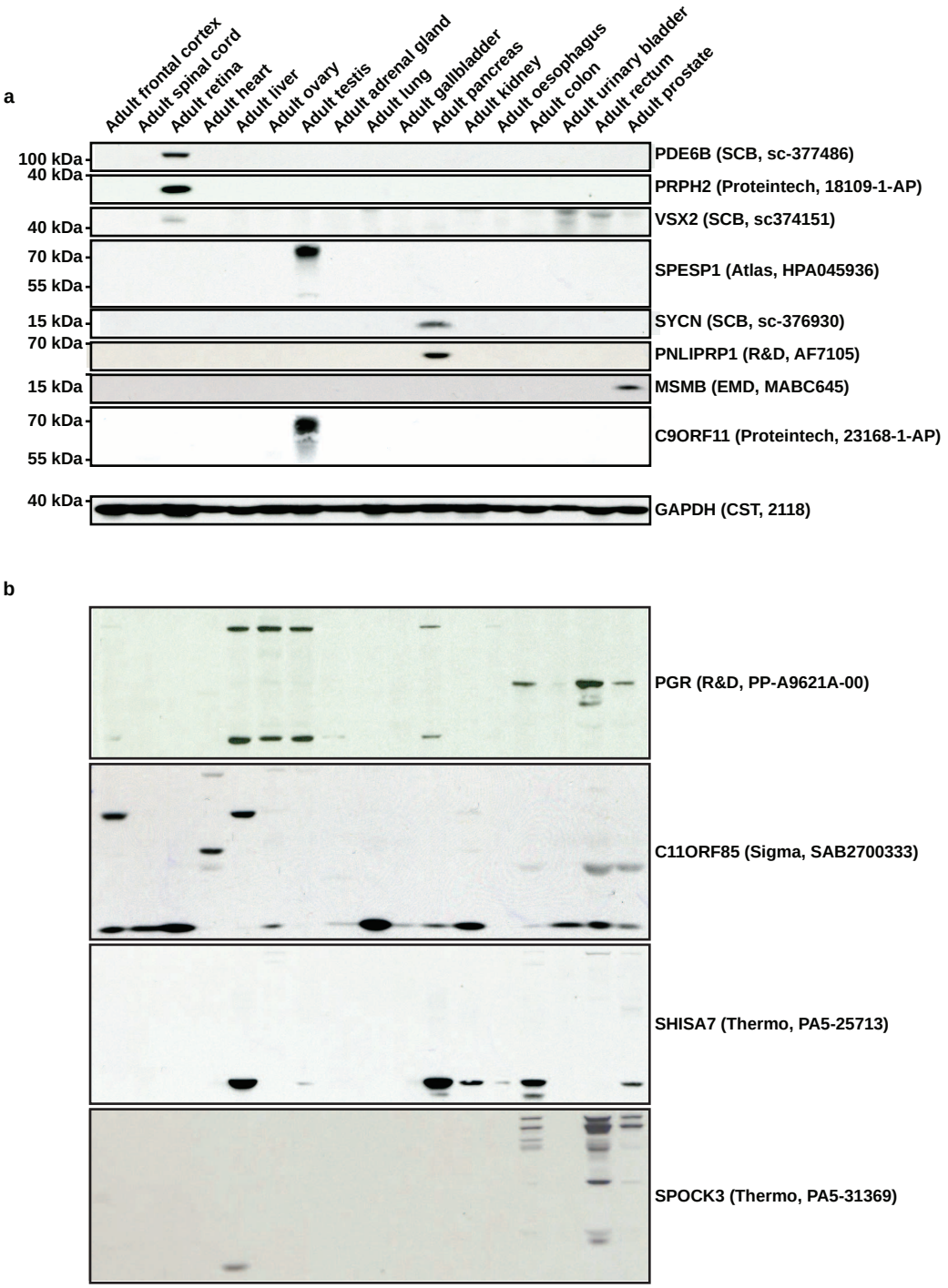


Extended Data Figure 1 | Summary of proteome analysis. **a**, Mass error in parts per million for precursor ions of all identified peptides. **b**, Number of peptides detected per gene binned as shown. **c**, Distribution of sequence coverage of identified proteins. **d**–**f**, %FDR with a q value of <0.01 plotted against peptide length in number of amino acids, charge state of peptide ion and number of cleavage sites missed by enzyme. P values computed from two-tailed t-test are shown. Error bars indicate s.d. calculated from FDRs of multiple fetal samples. **g**, **h**, A comparison of peptides identified in this study with PeptideAtlas and GPMDB. **i**, Mass error in parts per million for precursor ions identified from proteogenomics analysis.

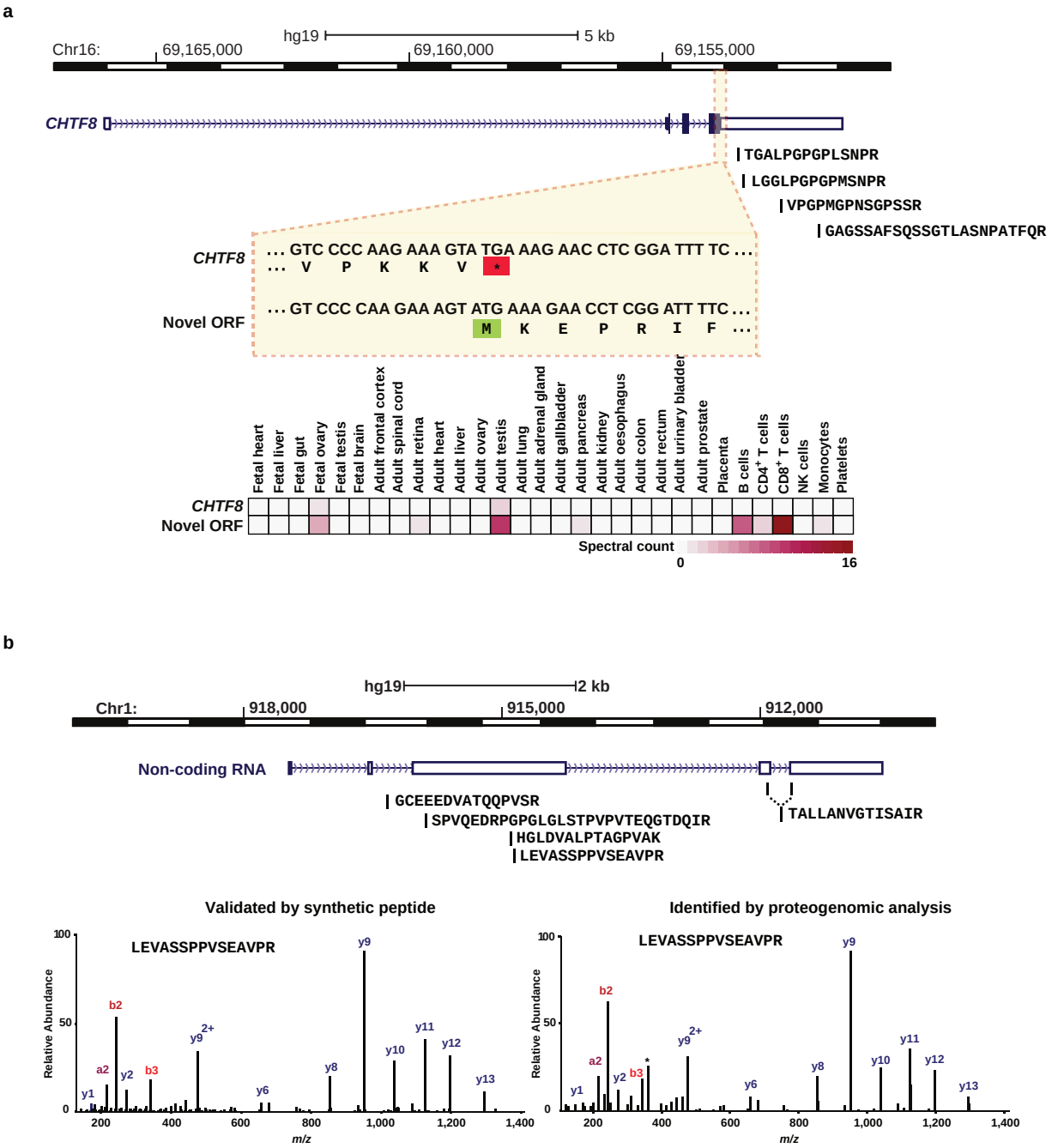


Transition ID	Precursor *m/z*	Precursor charge	Fragment *m/z*	Relative intensity
VESTDVSDLLHQYR.2.y8.1	852.415	2	1031.53	100
VESTDVSDLLHQYR.2.y4.1	852.415	2	603.30	52
VESTDVSDLLHQYR.2.y12.1	852.415	2	1433.70	44
VESTDVSDLLHQYR.2.y10.1	852.415	2	1245.62	42
VESTDVSDLLHQYR.2.y9.1	852.415	2	1130.60	42
VESTDVSDLLHQYR.2.y6.1	852.415	2	829.47	40
VESTDVSDLLHQYR.2.y3.1	852.415	2	466.24	33

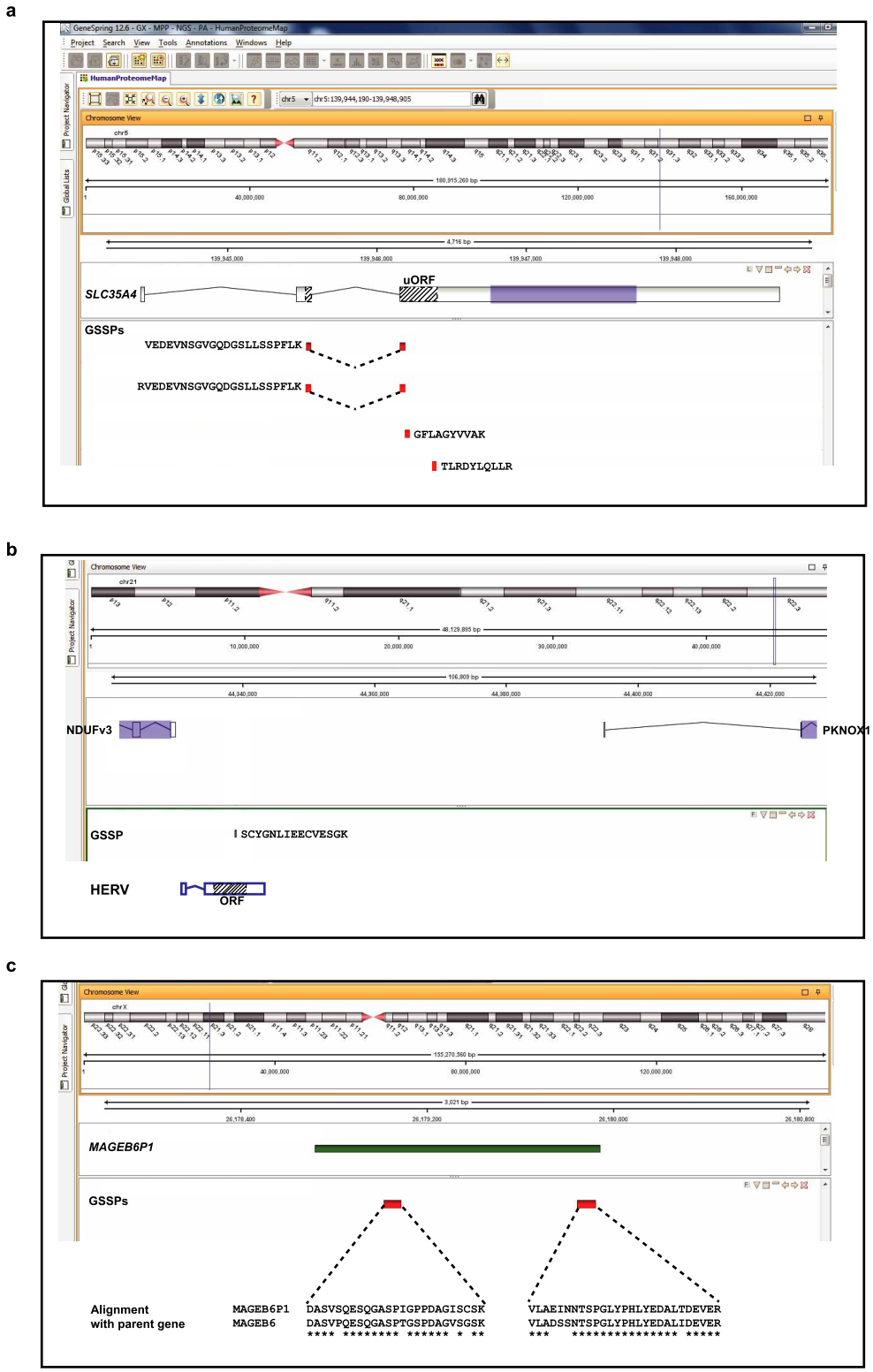
Extended Data Figure 2 | Tissue-wise gene expression and housekeeping proteins. **a**, A heat map shows a partial list of not well-characterized, hypothetical genes. **b**, The bulk of protein mass is contributed by only a small number of genes. Only 2,350 'housekeeping genes' account for ~75% of proteome mass. **c**, The number of cell/tissue types where a gene was observed was counted. Some genes were found to be specifically restricted in a few samples while others were observed in the majority of samples analysed. For example, 1,537 genes were detected only in one sample, and 2,350 genes were found in all samples. These latter genes can be defined as highly abundant 'housekeeping proteins'. **d**, Distribution of genes in the RefSeq database based on the number of protein isoforms resulting from their annotated transcripts (left). Distribution of the transcripts with two or more protein isoforms annotated based on the number of isoform-specific or shared peptides (right). **e**, A representative example of sequence coverage of PSMB8 protein along with tissue distribution of all of its identified peptides and the MS/MS spectrum of one of the peptides is shown along with seven selected reaction monitoring (SRM) transitions.



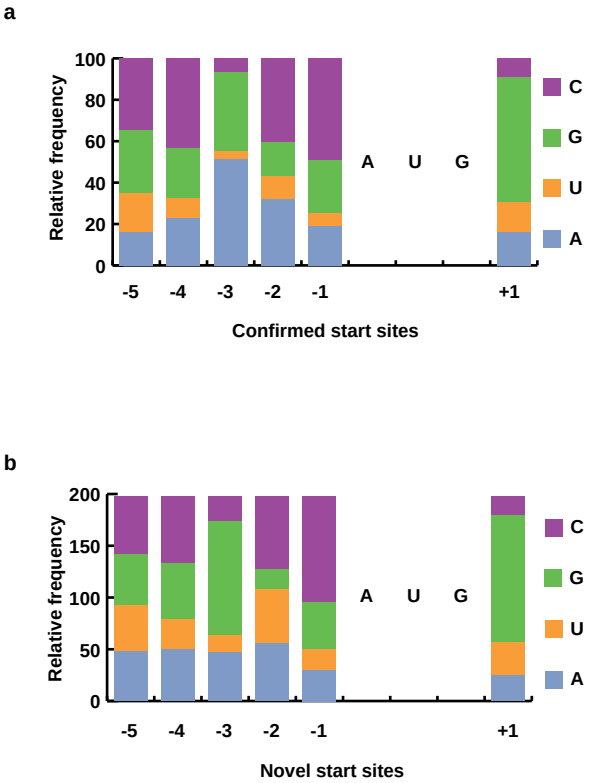
Extended Data Figure 3 | Western blot analysis of select tissue-restricted proteins. **a**, Eight proteins showing tissue-restricted expression were tested using western blot analysis in 17 adult tissues. GAPDH was used as a loading control. **b**, Four proteins found to be expressed in a broad range of tissues, although bands that do not correspond to the expected molecular weight are also observed. CST, Cell Signalling Technology; SCB, Santa Cruz Biotechnology.



Extended Data Figure 4 | Identification of novel genes/ORFs and translated non-coding RNAs. a, An example of a novel ORF in an alternate reading frame located in the 3′ UTR of *CHTF8* gene. The relative abundance of peptides from the CHTF8 protein and the protein encoded by the novel ORF is shown (bottom). **b**, An example of translated non-coding RNA (NR_027693.1) identified by searching 3-frame-translated transcript database. The MS/MS spectrum of one of the five identified peptides (LEVASSPPVSEAVPR) is shown along with a similar fragmentation pattern observed from the corresponding synthetic peptide.



Extended Data Figure 5 | Human genome annotation through proteogenomic analysis using GeneSpring. **a**, Four genome search specific peptides (GSSPs; red boxes) map to an upstream ORF (denoted as black hashes) located in 5′ UTR of the *SLC35A4* gene (ORF shown as blue rectangle). **b**, GSSP mapping in the intergenic region between two RefSeq annotated genes *NDUFv3* and *PKNOX1*. The ORF region is depicted in dotted lines of human endogenous retroviral element (HERV). **c**, GSSPs mapping to an annotated pseudogene *MAGEB6P1*, the alignments of parent gene and pseudogene are shown below the peptides.



Extended Data Figure 6 | Frequency of nucleotides surrounding translational start sites. **a**, Frequency of nucleotides at positions ranging from −5 to +1 surrounding the AUG codon for confirmed translational start sites. **b**, Frequency of nucleotides at positions ranging from −5 to +1 surrounding the AUG codon for novel translational start sites identified in this study.