Comprehensive comparison of large-scale tissue expression datasets

Alberto Santos[1]*[§], Kalliopi Tsafou[1]*, Christian Stolte[2], Sune Pletscher-Frankild[1#], Seán I. O'Donoghue[2,3] and Lars Juhl Jensen[1§]

[1] Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[2] CSIRO, Sydney, Australia

[3] Garvan Institute of Medical Research, Sydney, Australia

* These authors contributed equally

[#] Current affiliation: Ferring Pharmaceuticals, Copenhagen, Denmark

[§] Corresponding author

Email addresses:

AS: alberto.santos@cpr.ku.dk

KT: kalliopi.tsafou@cpr.ku.dk

CS: christian.stolte@csiro.au

SPF: sune.frankild@gmail.com

SIO: sean@odonoghuelab.org

LJJ: lars.juhl.jensen@cpr.ku.dk

Abstract

For tissues to carry out their functions, they rely on the right proteins to be present. Several high-throughput technologies have been used to map out which proteins are expressed in which tissues; however, the data have not previously been systematically compared and integrated. We present a comprehensive evaluation of tissue expression data from a variety of experimental techniques and show that these agree surprisingly well with each other and with results from literature curation and text mining. We further found that most datasets support the assumed but not demonstrated distinction between tissue-specific and ubiquitous expression. By developing comparable confidence scores for all types of evidence, we show that it is possible to improve both quality and coverage by combining the datasets. To facilitate use and visualization of our work, we have developed the TISSUES resource (http://tissues.jensenlab.org), which makes all the scored and integrated data available through a single user-friendly web interface.

Keywords

Tissue expression
Tissue-specificity
Immunohistochemistry
Mass spectrometry
RNA sequencing
Microarrays
Text-mining
Database

Introduction

Mapping out which proteins are present in each tissue is of major importance for understanding the functional differences between tissues as well as their development and differentiation [1,2]. Several high-throughput experimental technologies have been used for this, the most widely used of which are expressed sequence tags (ESTs) [3,4], high-density oligonucleotide microarrays (also called DNA chips) [5,6], and RNA sequencing (RNA-seq) [7,8].

ESTs are short sequence reads — typically around 400bp — derived from 5' or 3' ends of complementary DNA (cDNA) libraries from tissues or cell lines [9–11]. Consequently, for a highly expressed gene, one would expect to see a correspondingly high abundance of ESTs derived from its transcripts. A more recent sequencing-based approach to quantifying transcript levels is RNA-seq. The major difference to EST sequencing is that random cDNA fragments are sequenced instead of only the 5' and 3' ends. The resulting reads are aligned to a reference genome, producing a quantitative expression profile for each gene [12,13]. Because reads are generated from all parts of a transcript instead of only the ends, the number of reads observed for a gene depends on both its length and its level of expression. A major advantage of RNA-seq is the ability to resolve individual splice variants if enough reads are obtained for a gene. Microarrays are another extensively used technology for transcriptome analysis. Gene expression is quantified by measuring the fluorescence intensity of labeled cDNA that hybridizes to oligonucleotide probes [14–16]. Because a microarray can contain millions of different probes, the transcript levels of all genes can be measured simultaneously.

The above mentioned techniques are all based on measuring mRNA levels. Fewer techniques exist for high-throughput measurement of protein levels. One of them is multiplexed immunohistochemical staining of tissue samples embedded in paraffin blocks (sometimes referred to as tissue microarrays). Histological analysis of the resulting images of tissues stained with an antibody can semiquantitatively tell where the target protein is present [17]. The main challenge to using this approach at the proteome scale is the need for specific antibodies against all proteins [18]. Mass spectrometry has also been used for measuring protein abundances in tissue samples, mainly in bodily fluids [19–21], muscle biopsies [22], and tumor samples [23–25]. Two recent publications collected many of these experiments into a single repository [26] and for the first time used this technology for in-depth proteomic profiling of a broad selection of normal human tissues [27], respectively.

Large-scale tissue expression datasets have formed the basis for many analyses and discoveries related to roles of housekeeping and tissue-specific genes in protein complexes [2,28], biological processes [29–31], and diseases [32–36]. However, the majority of

these studies [28,30–32,34–36] are based solely on microarray data from the GNF Expression Atlas [5], which could bias the results. It is thus relevant to test to which extent the different technologies and datasets give congruent results.

We here present the first comparative evaluation of the quality of tissue associations from a variety of different datasets and experimental methods as well as from manual curation [37] and automatic text mining of the biomedical literature (Figure 1). We show that these datasets — despite the technological differences — agree surprisingly well with each other and can be combined to improve quality and coverage. Finally, as a result of the integration process, we have developed the TISSUES resource (http://tissues.jensenlab.org), which makes the above mentioned heterogeneous data more easily accessible to researchers by collecting them in a single place and assigning confidence scores.

Results

To systematically compare the different datasets, we standardized the varying names used for the same tissues to their respective terms in the Brenda Tissue Ontology [38] (Supplementary data 1). Because this ontology is structured as a directed acyclic graph, this also helps deal with the challenge of different datasets having different tissue resolution; for example, some datasets study the brain as a whole whereas others study different parts separately. We decided to base our analyses on the 21 major tissues shown in Figure 1.

Tissue-specific and ubiquitous transcripts

Many studies have made the distinction between housekeeping and tissue-specific genes, which are expressed in most or only a few tissues, respectively [29,39–44]. However, there are no strict definitions of these two classes of genes, and it is not clear to what extent this represents a natural classification. To answer the latter, we analyzed the expression breadth of five transcriptome datasets, i.e. how many genes are expressed in how many tissues. As this depends strongly on the threshold used to decide whether a gene is expressed in a given tissue, we performed the analysis with three different cutoffs, in the following referred to as low, medium, and high confidence (see Methods).

Figure 2 shows the expression breadths for five transcriptome datasets, each at the three different confidence levels. Most show a clear bimodal distribution with peaks at the extreme ends, i.e. the vast majority of genes are expressed either in only a few tissues or in most tissues measured. We thus show that data from several sources and technologies robustly support a natural distinction between tissue-specific and ubiquitously expressed genes.

Zhu and colleagues [42] also showed a bimodal trend when comparing the GNF

expression atlas and EST sequencing data; however, for the latter data type the bimodality was weak. We similarly find very few tissues-specific genes when analyzing UniGene at the low-confidence cutoff, but show that this trend is reversed when using more stringent cutoffs. We observe that the GNF dataset is atypical in that it identifies fewer ubiquitously expressed genes at all cutoffs than the rest of the datasets, including the other microarray-based study (Exon array).

Consistency of transcriptomic methods

The previous analysis showed that the global trends in terms of tissue specificity are similar across the transcriptome datasets. That, however, does not imply that the datasets necessarily agree on which genes are expressed where. To quantify the agreement, we focused on the five tissues and 3,254 genes covered by all the transcriptome datasets. Comparing the five transcriptome datasets, we saw that genes are assigned to tissues with high consistency between datasets at all three confidence levels (Figure 3). At medium confidence 39.2% (5679/14504) of gene–tissue associations are common to all datasets and 65.8% (9537/14504) are common to at least four of the five datasets (Supplementary data 3).

The largest discrepancy in the comparison is the large number of gene–tissue associations found by all datasets except GNF at all three confidence levels (Figure 3). This is likely because the GNF Expression Atlas was made using microarrays designed prior to the completion of the Human Genome Project, which consequently have suboptimal probe sets for many genes.

Conversely, the largest agreement is seen among the three most recent datasets, which were generated using RNA-seq or exon arrays. At medium confidence, their overlap makes up 72.65% (10538/14504) of all gene–tissue associations, 13.66% (1439/10538) of which are not found by any other dataset.

Correlation between expression values and confidence levels

The high consistency between the mRNA datasets demonstrates their quality; however, it does not guarantee that the selected cutoffs are comparable and represent the same level of confidence across datasets. To assess the assumed correlation between expression values and confidence, we compared all datasets to a gold standard of gene–tissue associations extracted from scientific literature by UniProtKB [37]. While reliable, UniProtKB annotations are very incomplete as they are restricted to what has been published. It is thus not possible to estimate the precision of a dataset; instead, we quantified the quality of the datasets in terms of its fold enrichment of correct gene–tissue associations compared to random chance.

The comparison showed that fold enrichment for gold-standard associations increased

steadily with expression value from all datasets (Figure 4A). This was expected because, in general, the more abundant a transcript, the more reliably it can be identified. Moreover, we find that the low-, medium-, and high-confidence cutoffs used in the preceding analyses correspond to the same quality in all datasets. However, a dataset of lower quality will give fewer associations at any given confidence cutoff.

The expression breadth distribution of UniProtKB is strongly skewed towards tissue-specific proteins; only 0.72% of proteins (106/14722) are annotated as expressed in more than five tissues. This likely reflects that many annotations describe proteins as widely or ubiquitously expressed but list only a few tissues. Also, UniProtKB annotations are incomplete, because many proteins have only been described in the literature as present in some of the tissues where they are expressed.

In light of this and the high quality of the mRNA datasets, we built a complementary set of gene–tissue associations, hereafter called the mRNA reference set, with high-confidence support from at least three datasets. This set exhibits the expected bimodal distribution of expression breadth (Figure 4C) and provides 7,384 gene–tissue associations not present in UniProtKB (Figure 4D, Supplementary data 3).

Quality of proteomics data

To complement the mRNA datasets with protein-level data, we investigated the Human Protein Atlas immunohistochemistry data (HPA IHC) [8] and the mass spectrometry data from the Human Proteome Map (HPM) [27].

To compare these with other datasets, we developed a quality scoring scheme for each. For HPM, we define the quality score as the number of unique tryptic peptides identified for a protein in a given tissue. The HPA IHC protein–tissue associations come with quality scores based on a combination of the staining level observed in the experiment, RNA-seq data and literature evidence. To make the IHC data independent of the other datasets, we instead developed a scoring scheme based purely on the staining levels and agreement between results from different antibodies with the same target (see Methods).

With the scoring schemes defined, we analyzed the two proteomics datasets with respect to enrichment for associations from both the UniProtKB and mRNA reference sets (Figure 5A). Higher scores were correlated with higher enrichment, giving some validation for the proposed scoring schemes. Despite looking at proteins instead of transcripts, the proteomics datasets show worse fold enrichment than the transcriptome datasets, when compared to the UniProtKB gold standard. This is consistent with the criticism raised over the quality of the HPM data based on analysis of olfactory receptors [45]. In case of HPA IHC, this is especially true for data derived based only on a single antibody.

HPM exhibits bimodal distributions of expression breadth at all confidence level consistent with the majority of the transcriptome datasets (Figure 5B). This consistency across confidence levels is in part due to a substantial fraction (23,440/107,935) of the associations from HPM being high confidence. Conversely, the HPA IHC dataset is dominated by low-confidence associations for proteins studied with only a single antibody or with multiple antibodies that gave different results. At low confidence, proteins tend to be associated with many tissues, which is likely due to unspecific antibodies. By contrast, most proteins have higher-confidence links to only a few tissues.

Complementary annotations from text mining

Automatic text mining of the biomedical literature has the potential to extract information that has been either overlooked by curators, not yet curated, or not annotated due to curation standards [46,47]. We used a previously published text-mining pipeline [48,49], expanded with a dictionary of tissues and cell lines, to extract associations between genes/proteins and tissues and scored them according to their co-occurrence in sentences and abstracts.

We evaluated the quality of these associations by comparing them to both the UniProtKB and mRNA reference sets (Supplementary figure 1a). This analysis shows that co-occurrence-based text mining performs well for this task. The high agreement with UniProtKB is not surprising considering that text mining and curation are both based on the available literature. The comparison to the mRNA reference set, however, shows that many of the associations found by text mining, but not by curators, are also supported by direct experimental evidence.

The distribution of expression breadths is, like for UniProtKB, skewed towards the tissue-specific end (Supplementary figure 1b), due to the same literature limitations. However, text mining associates each gene/protein with more tissues, even at high confidence. For example, 421 are linked to more than five tissues, which is four times more than what UniProtKB annotates. These results demonstrate the value of complementing manual annotation with automatic text mining.

Improved tissue profiles through data integration

So far we have shown that the quality of the different datasets is comparable at each of the chosen confidence levels. To assess the consistency and complementarity of different data sources, we compared the medium-confidence associations from UniProtKB and text mining to two pooled sets of high-confidence associations from transcriptomics and proteomics experiments, respectively.

Despite the inherent differences between data types and technologies compared, when

looking at the common proteins and tissues, 44.5% (18,904/42,435) of all associations are supported by at least two of the four sets (Figure 6A). The transcriptomics and proteomics sets show the largest pairwise agreement, which accounts for 33.5% (12,902/38,471) of the associations from the two sets and 30% (12,902/42,435) of all associations (Supplementary data 4). This agreement highlights the strong connection between transcription and final protein abundance; indeed, transcription was recently demonstrated to explain about 80% of the differences seen in protein expression [50].

Although all the sets are consistent on the proteins and tissues they have in common, they are also highly complementary because they cover different proteins and tissues. When not restricting the comparison to common proteins and tissues, 71% (102,574/144,525) of all the reported associations are unique to a single set (Figure 6B, Supplementary data 4). The analysis also reveals that only 6.2% (9,029/144,525) of the associations are unique to UniProtKB. Text mining alone captures 20% (5,551/27,596) of the curated literature results and complements them with 20,263 additional protein–tissue associations, 41.5% (8,423/20,263) of which are supported by the transcriptomics or proteomics sets.

Another way to illustrate the complementarity of the datasets is to compare the quality and coverage obtained when integrating many datasets compared to using a single dataset. To this end, we looked at the union of the transcriptomics and proteomics sets and compared it to the same number of top-scoring associations from the GNF atlas. Focusing on the 7,445 proteins and 17 tissues that GNF and UniProtKB have in common, 77% (11,489/14,978) of the associations from the integrated list were annotated in UniProtKB, whereas this was only the case for 60% (8,912/14,978) of the associations from GNF. Moreover, the integrated list includes 12,562 associations not covered by GNF (Supplementary data 5 and Supplementary figure 2).

The TISSUES web resource

In light of the clear advantages of combining multiple datasets, we believe the scientific community can benefit from having a resource that integrates and provides easy access to the available information on tissue expression. We thus developed the TISSUES web resource that is available at http://tissues.jensenlab.org. Several other resources provide gene–tissue associations, including TiGER [51], BioGPS [52], TissueDistributionDB [53], VeryGene [54], and EBI Gene Expression Atlas [55]. What makes TISSUES unique is that it integrates data from many different technologies and sources, quantifies the reliability of each gene–tissue association, and thereby makes results from different sources comparable.

The web interface allows the user to search for a human gene and get a complete overview of where it may be expressed. To provide an at-a-glance overview, we show a

body map with each the 21 major tissues colored according to the confidence that the gene of interest is expressed there (Supplementary figure 3). The figure also allows the user to see which sources of evidence support expression in a given tissue. Three interactive tables below the body map provide the user with more detailed information for the evidence from UniProtKB, high-throughput experiments, and text mining. This includes information on additional tissues, linkout to the source of the evidence whenever possible, and a unified confidence score ranging from 1 to 5 stars (see Methods).

TISSUES holds information for 21,294 genes and 5,305 different tissues and provides more than 2.2 million gene–tissue associations at varying confidence levels. These are all available for download under the Creative Commons Attribution License at http://tissues.jensenlab.org to facilitate large-scale studies.

Discussion

We have compared gene–tissue associations derived from different transcriptomic and proteomic methods, automatic text mining and manual curation of the scientific literature. To ensure robustness, the analysis was performed using different confidence levels. The comparison enabled us to assess the consistency across the various datasets, highlighting the differences between them in terms of quality and coverage.

Overall, the associations derived from the different high-throughput experimental methods show reasonably good overlap. Good agreement was also observed for the highest scoring associations when comparing the datasets with manually curated annotations, which are recognized as being of high reliability. This shows that high-confidence information can be extracted from any of these high-throughput methods when using appropriate scoring schemes. Conversely, many discrepancies were observed for low-scoring associations, and some datasets had poor coverage once filtered for quality.

The various experimental methods and other data sources have different strengths and weaknesses. Integration of the data from different sources is thus a prerequisite for getting the best possible overview of the tissue expression of proteins. Combining diverse high-throughput methods overcomes the limitations of single experiments where quality is often questioned; for example, the recently published drafts of the human proteome [26, 27] have been suggested to have high false positive rates [46]. We have also shown that integration of text-mining results extends the space of manually curated associations with unreported literature information, often supported by experimental methods.

We thus believe that the comparison presented in this paper demonstrates that high-throughput datasets should be neither trusted at face value nor entirely discarded.

Rather, they should be carefully analyzed and integrated to determine which parts of which datasets should be trusted, and how much. The publicly available TISSUES resource provides exactly that and thereby makes it easy for researchers to make the most of the many tissue-profiling efforts.

Methods

GNF Gene Expression Atlas

The experimental data from the Human U133A/GNF1H Gene Atlas) [5] was downloaded from the BioGPS portal (http://biogps.org/). The dataset contains information for 44,775 probe sets, which we filtered to remove probe sets associated with multiple targets (names ending with "_[r,i,f,x]_at" and control probe sets (names starting with "AFFX"). We mapped the remaining probe sets to gene identifiers using the probeset-to-gene annotation file (gnf1h.annot2007.tsv) and finally mapped these to 16,598 Ensembl protein identifiers using the alias file from the STRING database [56]. The GNF Gene Expression Atlas provides information for 79 tissues, 60 of which we could map to Brenda Tissue Ontology terms. We scored each gene–tissue association based on the normalized expression units obtained from the microarray analysis, under the assumption that transcripts identified with higher intensity are less likely to be false positives. When multiple probe sets mapped to the same gene, we used the mean expression value.

Affymetrix Exon tiling array

These high-density microarrays [6] contain probe sets for more than one million annotated and predicted exons. We downloaded the data from the Gene Expression Omnibus [57] (GSE5791 series matrix) and used the 565,690 probe sets mapped to a gene identifier according to the GPL4253 platform. We mapped the latter to 15,559 Ensembl protein identifiers. The Exon Array experiment examined 16 tissues mainly from the nervous system studying six sub-regions of the brain. All tissues could be mapped to BTO terms. As in the other microarray experiment, we used the mean normalized expression units as the score for each gene–tissue association.

UniGene

The UniGene database [3,4] clusters together Expressed Sequence Tags (EST) that belong to a single gene and includes information about the tissue where each EST was observed. We used the *Homo sapiens* UniGene Build #236, which contains 24,289 clusters that could be mapped to 18,493 Ensembl protein identifiers via the provided gene symbols or UniGene cluster identifiers. UniGene Human library (Hs.lib.info)

provides information for 80 tissues from which we discarded several with ambiguous names, e.g. “retina and testis” or ”uncharacterized tissue” (see Supplementary data 1), and finally obtained 60 BTO terms. The scoring scheme for UniGene is based on the number of ESTs clustered into a single gene that belong to the same tissue. When multiple clusters mapped to the same gene, we used the total number of ESTs from the clusters.

RNA-seq atlas

The RNA-seq Atlas [7] is a web-based resource that provides expression data for 21,399 genes in 11 tissues. We mapped the genes to 18,063 Ensembl protein identifiers using the STRING alias file; all the specified tissues mapped directly to BTO terms. We used the normalized Reads Per Kilobase per Million mapped reads (RPKM) as the confidence score for each gene–tissue association.

HPA RNA-seq data

The Human Protein Atlas version 12 [8] provides short-read high-throughput sequencing data (RNA-seq) in 27 non-disease tissues. We mapped 20,315 Ensembl gene identifiers for which the database contained expression levels to 18,491 Ensembl protein identifiers and all the tissue names to BTO terms. Similarly to the scoring scheme applied to the RNA-seq Atlas dataset, we assigned the normalized expression levels in Fragments Per Kilobase of exon per Million fragments mapped (FPKM) as the confidence score for each gene–tissue association.

HPA Immunohistochemistry

HPA also provides an atlas of protein expression derived from immunohistochemistry experiments over many tissues [8]. We obtained information on the expression of 16,384 genes in 45 tissues (data downloaded on 21st January 2014), which we mapped to 15,552 Ensembl Protein identifiers and 45 BTO terms. For each antibody and tissue, HPA provides a semiquantitative strength of staining ($staining_{a,t}$), which we translated into numeric values (not detected: 0, low: 1, medium: 3, high: 6). When only a single antibody was used to measure a protein, we simply used the staining values from that antibody as the confidence scores for the tissues.

When multiple antibodies for the same protein were used we used a more complex scoring scheme to combine the staining values from the individual antibodies:

$$score_{p,t} = \alpha \cdot quality_p \cdot level_{p,t}$$

where α is a scaling factor for making the multi-antibody scores comparable to the single-antibody scores, $quality_p$ captures the internal agreement among the

antibodies for the protein, and $level_{p,t}$ is a weighted average of staining values of the antibodies for the protein in a given tissue.

The correction factor for the quality of the antibodies is defined as:

$$quality_p = e^{-\beta \frac{R_p^2}{N_p}}$$

where β is a parameter optimized as described below, N_p is the number of antibodies for the protein, and R^2 measures the disagreement between the antibodies across all tissues:

$$R_p^2 = \sum_{a \in p} \sum_{t \in T} R_{a,t}^2$$

where T is the set of tissues studied and $R_{a,t}^2$ is the disagreement in a given tissue between one antibody and the average of the antibodies:

$$R_{a,t}^2 = \left(staining_{a,t} - \frac{1}{N_p} \sum_{a \in p} staining_{a,t} \right)^2$$

We defined the level of a protein in a given tissue ($level_{p,t}$) as a weighted average of the antibodies:

$$level_{p,t} = \frac{\sum_{a \in p} weight_{a,t} \cdot staining_{a,t}}{\sum_{a \in p} weight_{a,t}}$$

where the weights are defined based on the disagreements between the antibodies:

$$weight_{a,t} = 1 - \frac{R_{a,t}^2}{R_p^2}$$

We validated the scoring scheme and determined the values of the free parameters by calculating the fold enrichment (see Quality of proteomics data) against UniProtKB. The optimal values of the parameters were $\alpha = 3.0$ and $\beta = 0.7$.

Human Proteome Map

HPM is a large mass spectrometry-based catalogue of protein profiles in 30 normal human tissues [27], which contains more than 290,000 tryptic peptides. We mapped these to Ensembl by comparing the sequences to all theoretical tryptic peptides derived from Ensembl v75 protein sequences, allowing for up to two missed cleavages. We assigned each tryptic peptide to the corresponding Ensembl gene identifier and mapped these to

a total of 17,038 Ensembl protein identifiers using the STRING alias file. The 30 normal human tissues were comprised of 17 adult tissues, 7 fetal tissues, and 6 primary hematopoietic cell types. Because the corresponding adult and fetal tissues map to the same term in BTO, the 30 tissues mapped to only 26 different BTO terms. As confidence score for a protein being expressed in a given tissue, we used the number of different tryptic peptides observed.

UniProtKB tissue annotations

UniProtKB [37] provides manually curated protein annotations. This includes annotations of tissue expression for 17,075 human proteins. Whereas each protein is typically only annotated with one or a few tissues, the number of different tissue terms used is very high; we were able to manually map UniProtKB tissues for 401 different BTO terms in total. Because the annotations are manually curated, we considered all protein–tissue associations from UniProtKB to be of the highest confidence.

Text mining

The text mining pipeline used in this work has been described in detail elsewhere. It relies on an efficient dictionary-based named entity recognition algorithm [48] and a co-occurrence scoring scheme [49] to extract associations from Medline abstracts. To use the pipeline to extract of protein–tissue associations, we complemented the existing dictionary of human gene and protein names from STRING with a dictionary of tissue and cell types constructed from BTO. The pipeline extracted more than one million protein–tissue associations based on co-occurrences of 16,748 proteins and 5,300 BTO terms.

Evaluation and calibration of scores

To evaluate the quality of the gene–tissue associations from each dataset, we compared them to the UniProtKB gold standard. We quantified the agreement in terms of the fold enrichment, which we define as the fraction of pairs in a dataset that are also in the gold standard divided by the fraction expected by random chance. The latter is defined as the fraction of possible gene–tissue pairs that are found in the gold standard. For these fold-enrichment calculations we considered only the genes and tissues that are shared between the dataset and the gold standard.

We calculated the fold enrichment for score windows of 100 gene–tissue associations to capture the relationship between fold enrichment and the quality scores defined in the previous sections. To be able to convert the quality scores from individual datasets into confidence scores that are comparable between datasets, we first fit the relationships between quality scores and fold enrichments with mathematical functions with only a few parameters. We used these to define the low-, medium-, and high-confidence

cutoffs for the comparisons of the datasets (Supplementary Table 1). Next, we performed a global transformation of the fold enrichments into the "star" confidence scores used in the COMPARTMENTS resource [49] based on the text-mining scores, which the two resources have in common. The combined, calibrated functions for translating quality scores into the final confidence scores are listed in Supplementary Table 1 (Supplementary figure 4).

Web resource

To make the protein–tissue associations available for query by a web resource, we store all data in a PostGreSQL database. The web interface is implemented through the same Python web framework used for the COMPARTMENTS database [49]. The body map onto which the data is visualized was manually created in Adobe Illustrator and saved as a Scalable Vector Graphics (SVG). In the user's browser, JavaScript is then used to provide interactive coloring and labelling of tissues.

Authors' contributions

AS analyzed and compared the experimental datasets. KT parsed and mapped the manually curated information from UniProtKB. SPF and LJJ adapted the text-mining pipeline to protein–tissue associations. CS and SIOD developed the body map visualization. LJJ, SPF, and CS developed the web resource. AS, KT, and LJJ conceived the ideas and wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

The authors thank Janos X. Binder for help with the web resource. This work was in part supported by the Novo Nordisk Foundation Center for Protein Research and by CSIRO's OCE Science Leader program.

References

1. Pontén, F. *et al.* A global view of protein expression in human cells, tissues, and organs. *Mol. Syst. Biol.* **5,** 337 (2009).

2. Emig, D. & Albrecht, M. Tissue-specific proteins and functional implications. *J. Proteome Res.* **10,** 1893–1903 (2011).

3. UniGene: A Unified View of the Transcriptome. (2003). at <http://www.ncbi.nlm.nih.gov/books/NBK21083/>

4. Wheeler, D. L. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **31,** 28–33 (2003).

5. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 6062–7 (2004).

6. Clark, T. A. *et al.* Discovery of tissue-specific exons using comprehensive human exon microarrays. *Genome Biol.* **8,** R64 (2007).

7. Krupp, M. *et al.* RNA-Seq Atlas--a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics* **28,** 1184–5 (2012).

8. Fagerberg, L. *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).

9. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–6 (1991).

10. Bailey, L., Searls, D. & Overton, G. Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* **8,** 362–76 (1998).

11. Nagaraj, S. H., Gasser, R. B. & Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief. Bioinform.* **8,** 6–21 (2007).

12. Nagalakshmi, U., Waern, K. & Snyder, M. RNA-Seq: a method for comprehensive transcriptome analysis. *Curr. Protoc. Mol. Biol.* **Chapter 4,** Unit 4.11.1–13 (2010).

13. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

14. Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. & Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nat. Genet.* (1999).

15. Harrington, C. A., Rosenow, C. & Retief, J. Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol.* **3,** 285–291 (2000).

16. Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* (2002).

17. Kampf, C., Olsson, I., Ryberg, U., Sjöstedt, E. & Pontén, F. Production of tissue microarrays, immunohistochemistry staining and digitalization within the human protein atlas. *J. Vis. Exp.* (2012). doi:10.3791/3620

18. Buchwalow, I., Samoilova, V., Boecker, W. & Tiemann, M. Non-specific binding of antibodies in immunohistochemistry: fallacies and facts. *Sci. Rep.* **1,** 28 (2011).

19. Adkins, J. N. Toward a Human Blood Serum Proteome: Analysis By Multidimensional Separation Coupled With Mass Spectrometry. *Mol. Cell. Proteomics* **1,** 947–955 (2002).

20. Schmidt, A. & Aebersold, R. High-accuracy proteome maps of human body fluids. *Genome Biol.* **7,** 242 (2006).

21. Aretz, S. *et al.* In-depth mass spectrometric mapping of the human vitreous proteome. *Proteome Sci.* **11,** 22 (2013).

22. Lundby, A. *et al.* Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nat. Commun.* **3,** 876 (2012).

23. Schwartz, S. A. Protein Profiling in Brain Tumors Using Mass Spectrometry: Feasibility of a New Technique for the Analysis of Protein Expression. *Clin. Cancer Res.* **10,** 981–987 (2004).

24. Seeley, E. H. & Caprioli, R. M. Molecular imaging of proteins in tissues by mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 18126–31 (2008).

25. Paul, D., Kumar, A., Gajbhiye, A., Santra, M. K. & Srikanth, R. Mass spectrometry-based proteomics in molecular diagnostics: discovery of cancer biomarkers using tissue culture. *Biomed Res. Int.* **2013,** 783131 (2013).

26. Wilhelm, M., Schlegl, J. & Hahne, H. Mass-spectrometry-based draft of the human proteome. *Nature* **509,** 582–7 (2014).

27. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).

28. Bossi, A. & Lehner, B. Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* **5,** 260 (2009).

29. Zhu, J., He, F., Hu, S. & Yu, J. On the nature of human housekeeping genes. *Trends Genet.* **24,** 481–4 (2008).

30. Chang, C.-W. *et al.* Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One* **6,** e22859 (2011).

31. Schaefer, M. H. *et al.* Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. *PLoS Comput. Biol.* **9,** e1002860 (2013).

32. Shyamsundar, R. *et al.* A DNA microarray survey of gene expression in normal human tissues. *Genome Biol.* **6,** R22 (2005).

33. Vasmatzis, G., Klee, E. W., Kube, D. M., Therneau, T. M. & Kosari, F. Quantitating tissue specificity of human genes to facilitate biomarker discovery. *Bioinformatics* **23,** 1348–55 (2007).

34. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 20870–5 (2008).

35. Magger, O., Waldman, Y. Y., Ruppin, E. & Sharan, R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.* **8,** e1002690 (2012).

36. Börnigen, D. *et al.* Concordance of gene expression in human protein complexes reveals tissue specificity and pathology. *Nucleic Acids Res.* **41,** e171 (2013).

37. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42,** D191–8 (2014).

38. Schomburg, I. *et al.* BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* **41,** D764–72 (2013).

39. Hsiao, L. L. *et al.* A compendium of gene expression in normal human tissues. *Physiol. Genomics* **7,** 97–104 (2001).

40. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31,** 180–3 (2002).

41. Liang, S., Li, Y., Be, X., Howes, S. & Liu, W. Detecting and profiling tissue-selective genes. *Physiol. Genomics* **26,** 158–62 (2006).

42. Zhu, J., He, F., Song, S., Wang, J. & Yu, J. How many human genes can be defined as housekeeping with current expression data? *BMC Genomics* **9,** 172 (2008).

43. Dezso, Z. *et al.* A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol.* **6,** 49 (2008).

44. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends Genet.* **29,** 569–74 (2013).

45. Ezkurdia, I. & Vázquez, J. Analyzing the First Drafts of the Human Proteome. *J. proteome …* (2014). doi:10.1021/pr500572z

46. Aerts, S. *et al.* Text-mining assisted regulatory annotation. *Genome Biol.* **9,** R31 (2008).

47. Van Auken, K. *et al.* Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase and TAIR. *Database (Oxford).* **2012,** bas040 (2012).

48. Pafilis, E. *et al.* The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS One* **8,** e65390 (2013).

49. Binder, J. X. *et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* **2014,** bau012–bau012 (2014).

50. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2,** e270 (2014).

51. Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9,** 271 (2008).

52. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10,** R130 (2009).

53. Kogenaru, S., Val, C., Hotz-Wagenblatt, A. & Glatting, K.-H. TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles. *Theor. Chem. Acc.* **125,** 651–658 (2009).

54. Yang, X. *et al.* VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery. *Physiol. Genomics* **43,** 457–60 (2011).

55. Kapushesky, M. *et al.* Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.* **38,** D690–D698 (2010).

56. Franceschini, A. & Szklarczyk, D. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids …* **41,** D808–15 (2013).

57. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res.* **39,** D1005–10 (2011).

Figures

Figure 1. Summary of the tissues and number of proteins present in each dataset. For our analyses, we mapped 9 datasets to 21 major tissues of interest. This figure shows which datasets cover which of these major tissues and how many proteins each dataset identified.

Figure 2. Distribution of expression breadth of the transcriptome datasets. For each of the five mRNA datasets, the histograms show the number of protein-coding genes expressed at low, medium, and high confidence as function of number of tissues. With the exception of UniGene, the distributions are bimodal, with most proteins occurring in either few tissues or in most tissues measured, supporting the notion of distinguishing between tissue-specific and ubiquitous expression.

Figure 3. Consistency of the transcriptome datasets. We assessed the consistency of the five transcriptome datasets by calculating the overlap of gene–tissue associations for the shared genes and tissues. At all levels of confidence, we observe surprisingly good agreement, with the largest count in each Venn diagram representing associations found by all five datasets.

Figure 4. Quality of the transcriptome datasets. *a.* To assess the correlation between expression level and confidence, we compared the transcriptome datasets to a gold standard, namely UniProtKB. We quantify the quality of the datasets in terms of its fold enrichment for correct gene–tissue associations compared to random chance. The comparison shows that higher expression values imply higher quality and that the three confidence cutoffs (vertical dotted lines) used correspond to equivalent quality in all datasets. *b.* The distribution of expression breadth for UniProtKB is strongly skewed towards tissue-specific proteins, contrary to what was seen for transcriptome datasets. *c.* We thus constructed a consensus mRNA reference set; its expression breadth distribution is in line with that of the individual mRNA datasets. *d.* The mRNA reference set is highly complementary to the UniProtKB gold standard, providing 7,384 gene–tissue association that are not in the latter.

Figure 5. Analysis of the proteomic datasets. *a.* To make the data from HPA IHC and HPM comparable with other datasets, we developed a quality scoring scheme for each. The quality scores show good correlation with the fold enrichment for associations from the UniProtKB and the mRNA reference sets. *b.* The distribution of expression breadth is consistent with the results of the transcriptome datasets in case of HPM, whereas the results for HPA IHC vary qualitatively between confidence levels.

Figure 6. Consistency and complementarity of evidence types. To assess the

consistency and complementarity of the associations supported by different types of evidence, we compared the medium-confidence associations from UniProtKB and text mining to two pooled sets of high-confidence associations from transcriptomics and proteomics experiments, respectively. The white numbers show the overlap of protein–tissue associations when considering only at the common proteins and tissues among all sets. The black numbers show the overlap when not restricting the comparison to common proteins and tissues. Together these analyses show that the different sources of evidence have high consistency across the common proteins and tissues, but that they are at the same time complementary because they cover different proteins and tissues.

Supplementary information

All the data and R code necessary to reproduce the analyses performed in this systematic comparison can be downloaded at:

https://github.com/albsantosdel/TISSUES-database-reproducible-analyses

Supplementary Table 1 - Definition of cutoffs. This table shows the different confidence cutoffs used in the analyses for each dataset, the quality score and how each quality score is converted to the unified confidence score used in the TISSUES web resource.

Supplementary data 1 – Mapping of tissue names to Brenda Tissue Ontology terms

This excel file contains the mapping from the tissue names from the original sources to the standardized BTO terms.

Supplementary data 2 - Common associations transcriptomic methods

This file contains the following information:

- The list of genes studied at the different cutoffs
- The list of common associations to all datasets at the different cutoffs
- The list of common association for at least 4 datasets at the different cutoffs

Supplementary data 3 - mRNA reference set associations

This excel file contains the gene–tissue associations that form the mRNA reference set used in the fold-enrichment analysis.

Supplementary data 4 - Common and unique gene–tissue associations to all the sets

This file contains:

- Overlap between all the sets (transcriptomic set, UniProtKB, Text-mining and proteomics set)
- Overlap between the transcriptomic and the proteomic set
- The list of gene–tissue associations unique to each set

Supplementary data 5 - Gene–Tissue associations coverage and quality analysis

This file contains:

- Gene–tissue associations from the integration of the transcriptomics and

proteomics datasets
- GNF atlas gene–tissue associations used in the analysis
- Overlap between the integrated set and UniProtKB
- Overlap between the GNF atlas studied set and UniProtKB

Supplementary figure 1. Complementary annotations from text-mining *a.* We used text-mining to extract associations between genes/proteins and tissues and score them based on their co-occurrence in sentences and abstracts. Comparing these associations to the UniProtKB and mRNA reference sets showed both the expected high agreement with UniProtKB and that many of the text-mined associations not annotated by curators are nonetheless supported by experimental evidence. *b.* The distribution of expression breadth for text mining is subject to the same literature limitations as UniProtKB. However, text mining associates each gene/protein with more tissues than the latter, even at high confidence, which demonstrates the value of complementing manual annotations with automatic text mining.

Supplementary figure 2. Quality and coverage. This figure shows how the overlap between UniProtKB and the sets derived from the GNF atlas alone (panel a) and the combined transcriptomics and proteomics data (panel b), respectively.

Supplementary figure 3. TISSUES: all data accessible in a single resource. The TISSUES web resource integrates all the data compared in this study, quantifies the reliability of each gene–tissue association, and thereby makes associations from different sources comparable. When searching for a human protein, the user is presented with a body map that provides a complete overview of where the protein is likely expressed by coloring the 21 major tissues according to the confidence of the protein–tissue association. The body map is interactive and allows the user to see which sources of evidence support expression in a given tissue. The TISSUES web resource is available at http://tissues.jensenlab.org.

Supplementary figure 4. Score calibration. The figure shows that after score calibration, the same confidence score corresponds to the same quality irrespective of the source of the evidence.

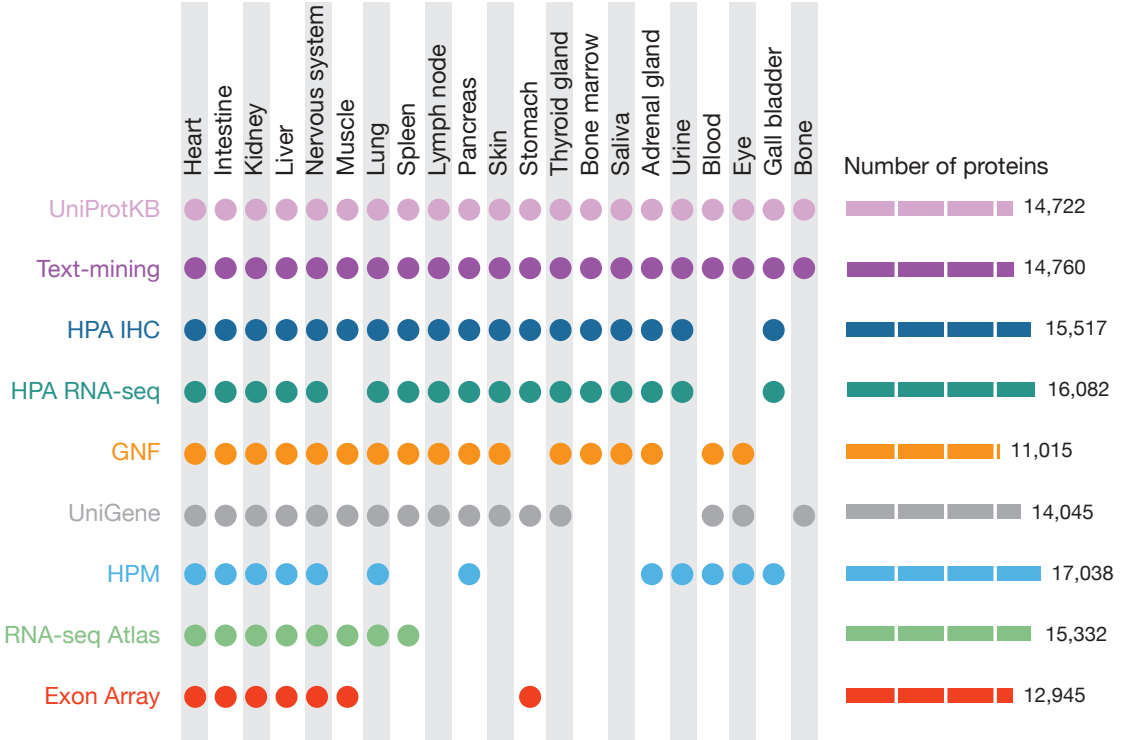

Heart
Intestine
Kidney
Liver
Nervous system
Muscle
Lung
Spleen
Lymph node
Pancreas
Skin
Stomach
Thyroid gland
Bone marrow
Saliva
Adrenal gland
Urine
Blood
Eye
Gall bladder
Bone
Number of proteins
UniProtKB
14,722
Text-mining
14,760
HPA IHC
15,517
HPA RNA-seq
16,082
GNF
11,015
UniGene
14,045
HPM
17,038
RNA-seq Atlas
15,332
Exon Array
12,945

Figure 1

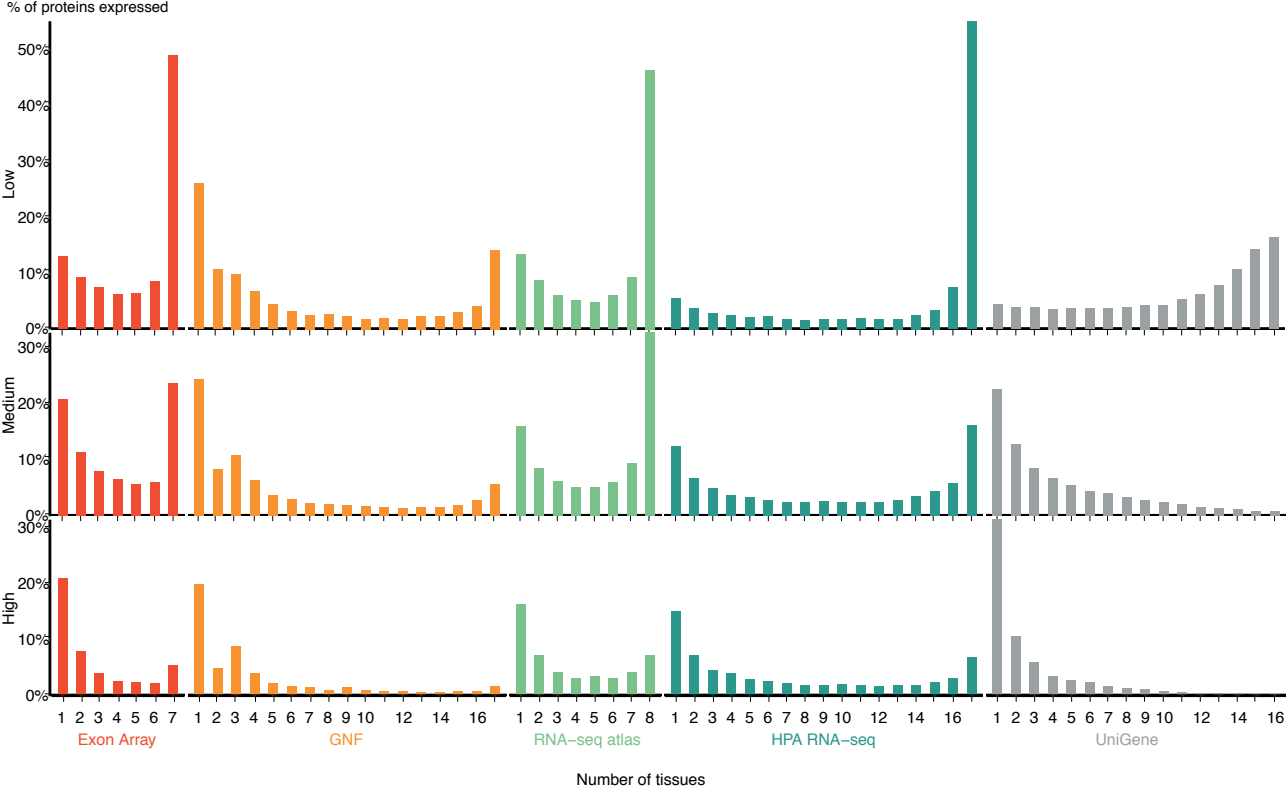
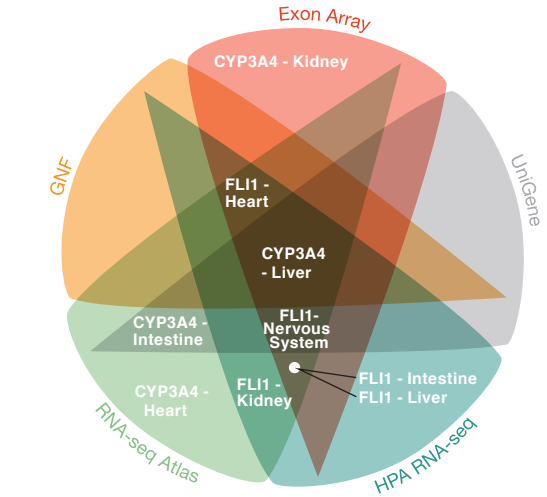
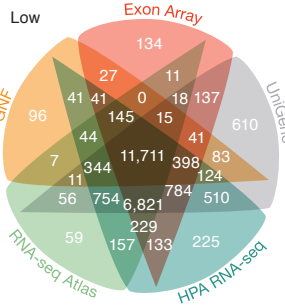
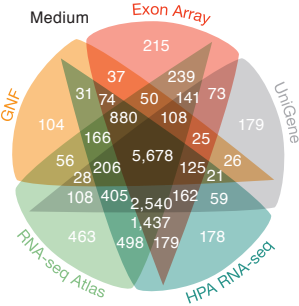

% of proteins expressed
Low
Medium
High
0%
10%
20%
30%
40%
50%
1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10 12 14 16
Exon Array
GNF
RNA−seq atlas
HPA RNA−seq
UniGene
Number of tissues

Figure 2

Example: CYP3A4 and FLI1







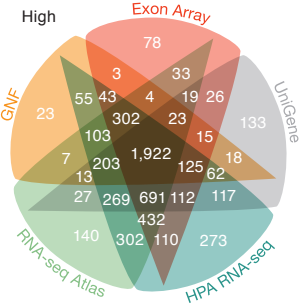
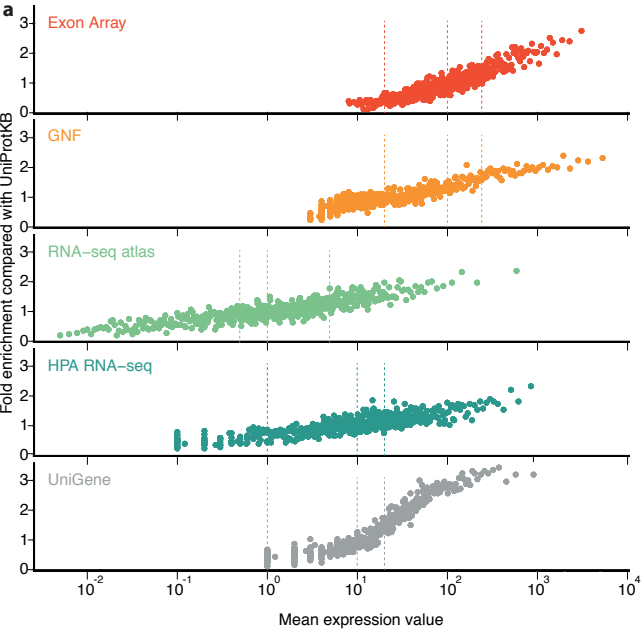
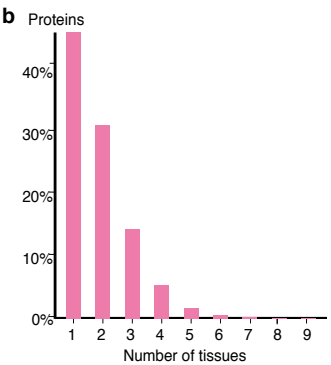
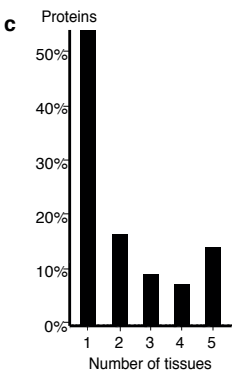


Figure 3


a
Exon Array
GNF
RNA−seq atlas
HPA RNA−seq
UniGene
Fold enrichment compared with UniProtKB
Mean expression value


b
Proteins
Number of tissues


c
Proteins
Number of tissues

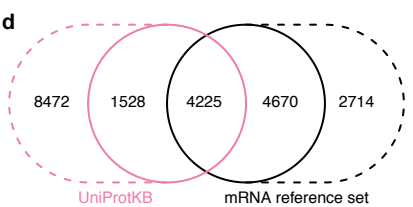

d
8472
1528
4225
4670
2714
UniProtKB
mRNA reference set

Figure 4

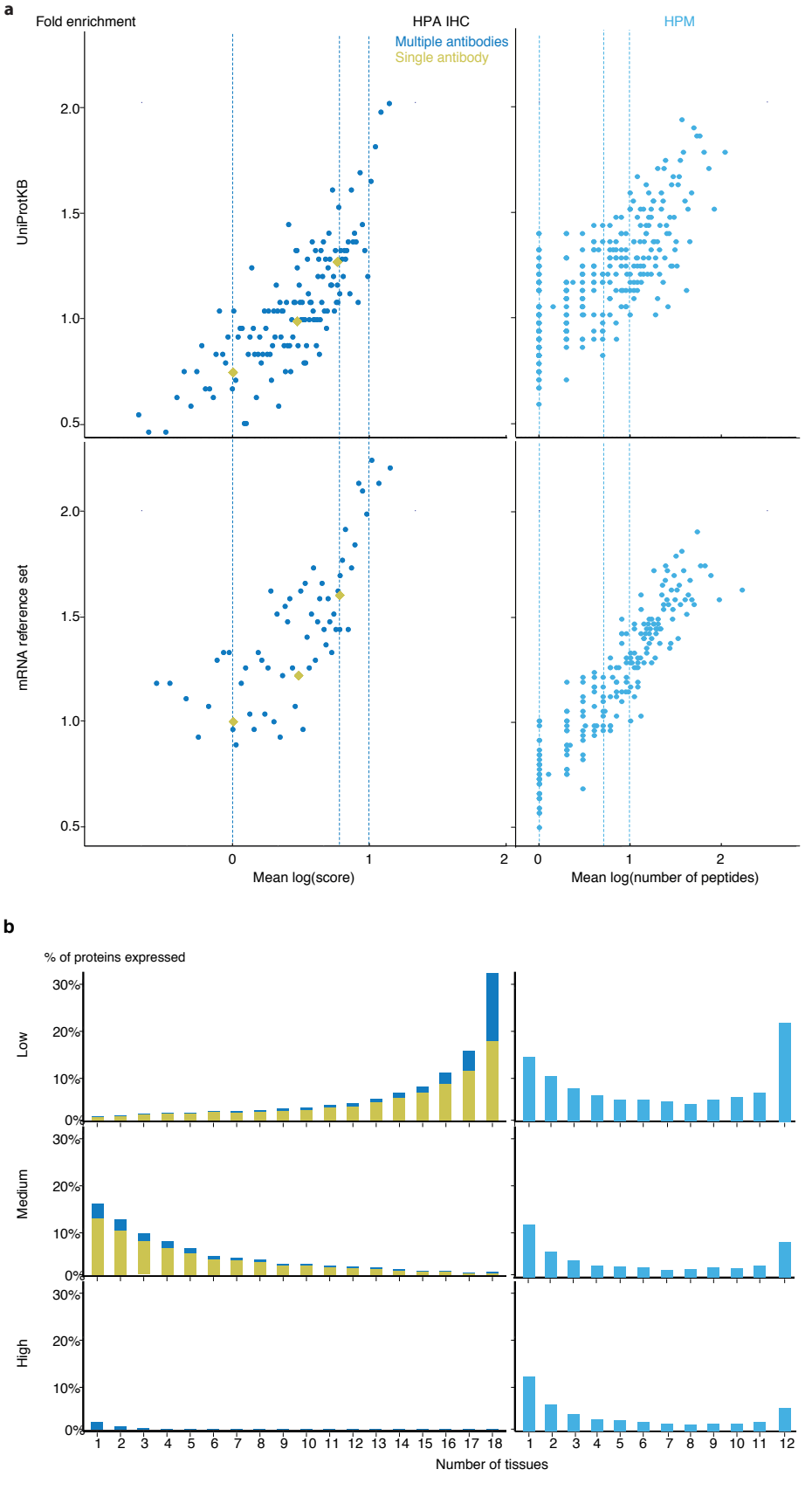

a
Fold enrichment
HPA IHC
HPM
Multiple antibodies
Single antibody
UniProtKB
mRNA reference set
Mean log(score)
Mean log(number of peptides)
b
% of proteins expressed
Low
Medium
High
Number of tissues

Figure 5

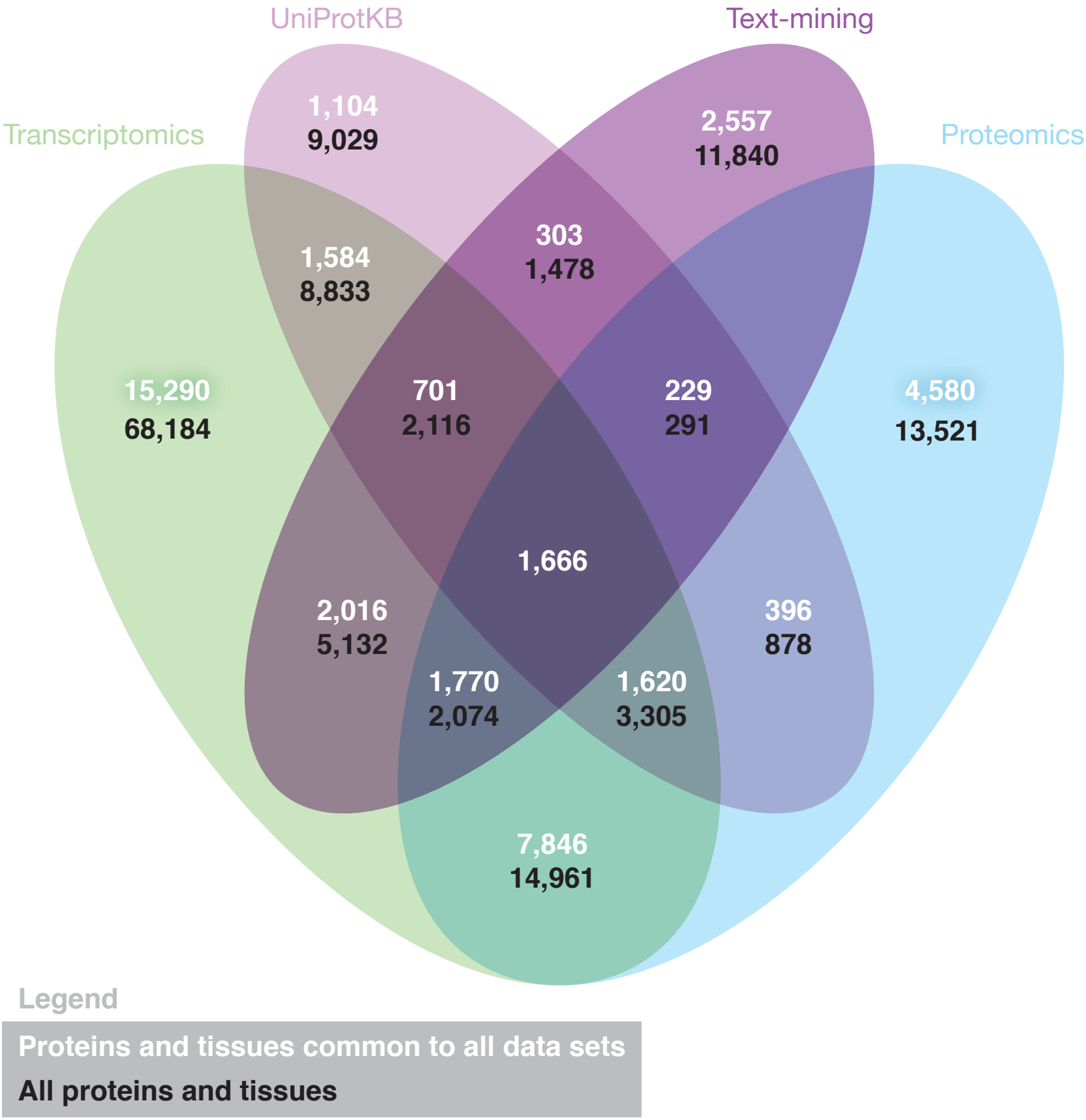

UniProtKB
Text-mining
Transcriptomics
Proteomics
1,104
9,029
2,557
11,840
1,584
8,833
303
1,478
15,290
68,184
701
2,116
229
291
4,580
13,521
1,666
2,016
5,132
396
878
1,770
2,074
1,620
3,305
7,846
14,961
Legend
Proteins and tissues common to all data sets
All proteins and tissues

Figure 6