Published in final edited form as:

Wiley Interdiscip Rev Syst Biol Med. 2016 November ; 8(6): 472–484. doi:10.1002/wsbm.1354.

Data sources for *in vivo* molecular profiling of human phenotypes

Timothy Cardozo^{1,*}, Priyanka Gupta^{1,2}, Eric Ni^{1,2}, Lauren M. Young³, Doreen Tivon², and Klara Felsovalyi²

¹Department of Biochemistry and Molecular Pharmacology, NYU School of Medicine, New York, NY, USA

²GeneCentrix Inc., New York, NY, USA

³Department of Pathology, NYU School of Medicine, New York, NY, USA

Abstract

Molecular profiling of human diseases has been approached at the genetic (DNA), expression (RNA), and proteomic (protein) levels. An important goal of these efforts is to map observed molecular patterns to specific, mechanistic organic entities, such as loci in the genome, individual RNA molecules or defined proteins or protein assemblies. Importantly, such maps have been historically approached in the more intuitive context of a theoretical individual cell, but diseases are better described in reality using an *in vivo* framework, namely a library of several tissue-specific maps. In this article, we review the existing data atlases that can be used for this purpose and identify critical gaps that could move the field forward from cellular to *in vivo* dimensions.

INTRODUCTION

The landmark achievement of the first sequencing of the human genome 15 years ago suggested vast and imminent change in our understanding of biology. The promise of rapid application of this knowledge to improvement in medicine was not quickly realized, and we now know that path is significantly more protracted and complex than previously appreciated. The major driver of progress since 2001 has been the generation of improved technologies to rapidly and inexpensively generate Omics data (genomic sequences, transcriptomes, epigenetic and proteomic maps); however, the attainment of new knowledge from this data has lagged behind. Specifically, cellular and human physiology is known to be a mind-boggling interplay of a vast (and as yet incompletely documented) array of DNA, RNA, protein, lipid, carbohydrate, and other molecules operating in various interlinked compartments and environmental scenarios. The causes or mechanisms of biological phenotypes (knowledge) are most easily identified for downstream of effectors of this network. For example, changes in the modification of the central cell regulator Ras by a lipidation enzyme are unambiguous, ‘on-off’ switches for certain cellular phenotypes, while the expression level of Ras in the same cell may not be correlated with any phenotype.

*Correspondence to: Timothy.Cardozo@nyumc.org.

Conflict of interest: The authors have declared no conflicts of interest for this article.

Conversely, molecular profiling data of cells and tissues at the genetic level dwarfs the data profiling of molecules at the downstream level, creating an inverse relationship between data availability and biological causality for Omics data (Figure 1(a)). Systems and computational biology in medicine have historically been widely concerned with the challenge of overcoming this inverse relationship to generate knowledge.

This challenge is even more daunting when one considers the variation among cells in the organism. One can imagine a scenario in which the specific mechanism of a disease is molecular variation resulting in barely abnormal functioning of one machine in one type of cell synergizing with barely abnormal functioning in several other cell types located in distant tissues in the body and converging into a dramatic disease phenotype in the whole organism. The genome sequences of the cell types in this example are identical, but one would expect wide and divergent variations in their transcriptomes, epigenomes, and proteomes. Considering genomic or proteomic variation in only one of these cell types, or some kind of average of the cell types, would make it impossible to understand such a disease, as would considering only one level such as the transcriptome or epigenome. Clearly, vertical layers of Omics data in this example would need to be integrated across tissues to generate knowledge about the disease: an *in vivo* framework would have to be engaged. Conversely, the horizontal variations in molecular profiles *in vivo* between different tissues or vertically across levels might provide valuable informatics reference points to deconvolute the true molecular drivers of the disease.

As one of many well-known examples, consider the vertically (genome → epigenome → transcriptome → proteome → signal transduction) and horizontally (tissues) variable molecular signature of the action of selective estrogen response modifiers (SERMs). As a drug class with tissue-specific dependence on the differential co-expression and co-recruitment of gene expression modulators, SERMs are an elegant example of conflicting phenotypes in distant and disparate tissues for the same target gene/gene product. SERMs have been a mainstay in targeting the estrogen receptor as an antagonist in breast tissue, opposing its effects by recruitment of corepressors to target genes, thereby reducing the proliferation of mammary tissue in both prevention and treatment of breast cancer and in male gynecomastia. However, tamoxifen, the founding member of the class and resident on the WHO's List of Essential Medicines, is instead estrogen-mimicking in uterine tissue, where it partners with coactivators in the canonical agonist manner of estradiol to upregulate gene targets such as *c-myc*, driving endometrial proliferation and potentiating malignancy. Tamoxifen's effects also imitate those of endogenous estrogen in preventing or reversing bone demineralization in post-menopausal women, in lowering serum low-density lipoprotein (LDL) levels and reducing coronary artery disease risk, and in creating a prothrombotic state. Conversely, tamoxifen use decreases estrogen-enabled protection from cataracts and can potentiate menopausal symptoms, inducing hot flashes and insomnia—via thermoregulatory and sleep centers in the brain—as a result of estrogen antagonism. To further muddy the picture, a closely related drug in the same class, raloxifene, acts identically to tamoxifen in some contexts, such as treatment of breast cancer and osteoporosis, but often differs remarkably in others, most notably acting without the adverse consequence of endometrial hyperplasia and transformation. These complex effects and their emerging mechanisms would be impossible to decipher from genomics data alone, or even

from vertically complete data for a single tissue: an *in vivo* framework would have to be engaged.

Even Mendelian genetic diseases engage a tissue complexity. Genetic mutations can lead to tissue-specific pathologies,^{1–3} but how, for many complex and prevalent diseases, the tissue-specific expression pattern of affected genes mechanistically leads to the disease pathology, which may have a dramatically different tissue pattern, may be obscure. For example, the LDL receptor is expressed most highly in adrenal and lung tissue in normal humans (<http://www.proteinatlas.org/ENSG00000130164-LDLR/tissue>), but in the tissue pattern of the disease phenotype of an LDL receptor deficiency, such as familial hypercholesterolemia, these are not the dominant tissues exhibiting pathology.⁴ This phenomenon can be clearly confounding for Omics molecular profiling efforts, which might focus on the Omics of the wrong cells from the disease. Thus, complex, prevalent multi-tissue diseases, such as heart disease, psychiatric diseases, metabolic diseases and neurological diseases, whether Mendelian or multigenetic, remain resistant to knowledge attainment via Omics approaches. These types of diseases are the motivation for this review.

Diseases that are intrinsically defined by a tissue, such as cancer, are a notable exception. There are many clear simple and direct connections between genetic mutations that result in enhanced cell growth or metastasis (b-RAF, BRCA, etc) and cancer disease phenotypes, which are the direct result of the growth and spread of that single tissue. Not surprisingly, Omics analyses of these diseases have been most productive to date. Accordingly, diseases that are intrinsically defined by tissues, such as cancer, are not the main focus of this review, although the relevant data sources are included for context.

Drugs used to treat human diseases and chemicals that cause diseases or adverse effects also act on specific cellular components that may vary between body tissues. Thus, systems pharmacologic approaches may also benefit from taking tissue specificity into account. Recently, we introduced the concept of ‘historeceptomics’ signatures of the actions of drugs.⁵ Most drugs exert their beneficial and adverse effects through their combined action on several different molecular targets (polypharmacology). The true molecular fingerprint of the direct action of a drug in the human body has two components: the ensemble of all the receptors upon which a drug acts, and the differential level of expression of these receptors across organs/tissues. We integrated drug–receptor interaction scores, or affinities, with whole-genome gene expression data for human receptors across a panel of human tissues to produce drug-specific tissue-receptor (historeceptomics) scores. A statistical model was designed to identify significant scores, which define an improved fingerprint representing the unique physiological activity of any drug. These multi-dimensional historeceptomic fingerprints describe, in a novel, intuitive style, the holistic, *in vivo* picture of the mechanism of any drug’s action. Importantly, high-dimensional pharmacology and genomic data are integrated and mapped to fewer dimensions, using the tissue atlas of the human as the integrative data structure.

This novel approach to profiling the *in vivo* signature of diseases and *in vivo* action of drugs could be used for better knowledge generation. The critical entity distinguishing the perspective in this review article from others is the *gene–tissue* pair, or its counter-parts the

protein-tissue or target-tissue pair. We review large biomedical data resources that promise insight into the molecular basis of diseases, evaluating them simply in terms of whether they can display gene (or protein) level variation across tissues, or be queried with a gene–tissue pair (or protein-tissue pair) or set of pairs for the purposes of knowledge discovery. We define the term ‘tissue-agnostic’ to mean that the molecular profiling data to which we apply the term does not take into account the expression or distribution of the gene or gene product across different tissues. Further-more, we consider three primary levels of resolution or definition for ‘tissues’: anatomically derived tissues (e.g., cardiac tissue), primary cells, and cell lines.

GENOMIC BIG DATA RESOURCES, CLASSIFIED BY TISSUE PROFILING POTENTIAL

Genome-wide association studies (GWAS) identify nucleotide variations (SNPs) at specific locations in the genome (loci) that correlate with diseases.^{6,7} Theoretically, a variable nucleotide that truly contributes to disease initiation, development, or progression could do so through differential tissue expression. However, *a priori*, GWAS data are *tissue-agnostic*, in the sense that the base genome is considered to be constant across tissues. Thus, these data cannot be queried with a gene–tissue pair.

Expression quantitative trait loci (eQTL) are the first step in potentially translating GWAS variations into functions. In this case, the expression of a specific gene is linked to SNPs by correlating the presence of the SNP with a change in the expression level of the gene as compared to wild-type. These types of data do indeed have a tissue profile, since the SNP may correlate with expression changes in cells from some tissues, but not from other tissues.⁸ However, eQTL need to be collected in a standardized fashion for all genes and for a range of human tissues in order for this data to be queried with gene–tissue pairs.

The Genotype-Tissue Expression project (GTEx) aims to generate eQTL for all genes for a range of human tissues. Its main purpose is to work back from a GWAS signal to identify specific genes associated with the disease or phenotype for which the GWAS was performed⁹ (Figure 1(b)). GTEx is a *tissue profile resource* since gene expression variations across tissues can be generated for any gene. In addition, a meta-expression (multi-gene) profile can be generated for each eQTL, and conversely, an eQTL cluster can theoretically be generated for any gene–tissue pair.¹⁰ Thus, in theory, GTEx can be queried with gene–tissue pairs associated with a disease and yield specific loci associated with or regulating those gene–tissue pairs.

The ENCODE database is a dictionary of functional nucleotide sites in the genome, especially in noncoding regions such as promoters, enhancers, epigenetic marks, etc. Its main purpose is to provide information that may then be used to decipher the exact molecular chain of events between the nucleotide entity (e.g., SNP) and the expression of specific genes.¹¹ Accordingly, databases of protein–protein interactions, small RNAs, signaling pathways and transcription factor (TF) activities may all secondarily contribute to this goal by integrated analysis with ENCODE. Currently, due to the extreme diversity and lack of standardization of these data sources, the ENCODE database, protein-interaction

databases, small RNA, signaling and TF databases are in practice *tissue agnostic*, although they contain dispersed tissue-specific information. Thus, these data cannot be queried with a gene–tissue pair.

dbGaP is a collection of genome profiles for specific phenotypes. Since there is no tissue requirement in this database, different phenotypes have different tissue sources, with some having only a single cell or tissue sample representing the phenotype. Thus, these data are tissue-agnostic, and cannot be directly queried with a gene–tissue pair. While dbGaP appears able to be curated into a tissue-organized resource, its utility is currently limited in this respect.¹²

BioGPS is a gene annotation portal that collaborates with other available gene-centric databases to provide a user interface to query genes or gene sets. While BioGPS itself provides an unstructured report on differential gene-expression patterns across various tissues and cell types, it provides a library of plugins for other databases that facilitates retrieval of the desired gene annotation data in a customized layout (Figure 2).¹³ Though BioGPS does not allow querying with a gene–tissue pair, it is a tissue profile resource because it readily displays normalized gene expression across tissues.

The Cancer Cell Line Encyclopedia (CCLE) describes a large set of human cancer cell lines. Although cancer is not a focus of this review, as explained above, the CCLE is a valuable tissue-profile resource, but only for variation across cancer cell lines. It contains gene expression profiling, chromosomal copy number data, and mutational data for a panel of 947 human cancer cell lines, representing 36 tumor types. For a given cell line, users can access these data; alternatively, one can search a gene and compare expression levels and copy number variation across the different tumor types. Several additional visualization tools are provided through the web interface, such as for differential expression or gene co-expression.

The Gene Expression Omnibus (GEO) is a repository for high-throughput gene expression data,¹⁴ and also contains other types of data: profiles of noncoding RNA, chromatin-IP, genome methylation, SNPs, genome variation and SAGE. In terms of gene expression data, GEO includes profiles across cells and tissues, and can be mined for studies that compare control and experimental conditions. These can include a disease, small molecule or genetic perturbation, and a gene expression signature can be created for any given perturbation.¹⁵ However, as with dbGaP, these data are not curated into a form that is practical for a general tissue-gene lookup since it represents many heterogeneous experimental sets, therefore, GEO itself is currently tissue-agnostic.

Although the GEO database and interface itself is tissue-agnostic, several tools have been build on top of this data repository to use it for tissue-specific queries. The Gene Expression Barcode is particularly intuitive in providing binary genomic code classifying every gene as either expressed or not expressed for a large number of human and mouse tissues and cells. This allows the generation of a profile for a gene across a large number of cells and tissues in which it is easy to see cross-tissue variation.¹⁶ Another tool is Concordia, where user-provided gene expression profiles are compared against a curated database of over 3000

samples from GEO that have annotation for their tissue of origin and disease to identify single-tissue or cell line-specific expression signatures.¹⁷ Gene Expression Commons also yields an estimate of absolute expression levels and dynamic ranges in a number of cell types.¹⁸

Expression Atlas is a collection of human gene expression patterns, curated from published literature and other genomic databases. This dataset provides information about gene expression profiles under different biological conditions, and different biological conditions affecting a specific gene.¹⁹ This information is organized into different biological groups (representing cell or tissue type, normal or diseased state, etc.) and provides an easy search function using both gene and tissue query system. Hence, Expression Atlas is a tissue profile resource.

TiGER is a comprehensive database dedicated to tissue-specific gene expression and regulation by TFs and *Cis*-regulatory elements.²⁰ It is a tissue profile resource as it not only lists genes that are preferentially expressed in a specific tissue but also provides data on TF interactions and their occurrence across tissues. Also, based on the TF interaction data, this resource provides a distribution of *cis*-regulatory modules in the promoter regions of curated tissue-specific genes. Overall this resource could prove to be very useful in understanding tissue-specific gene-expression and the elements involved in gene regulation. Though it does not allow queries with a gene–tissue pair, it has an integrated query system that lists all tissues for a gene entry and similarly all genes, TFs, etc. for a tissue entry.

Roadmap Epigenomics is a project that aims to create an epigenomic map for stem cells and primary tissues. The data involves DNA methylation sites, histone modification sites, chromatin accessibility, and RNA expression profiles.²¹ These epigenomic indicators vary among individuals, so this resource provides a series of normal epigenomes for reference and comparison.²² This can provide insight into how changes in the epigenome are related to disease, and can create an epigenomic signature for each. Although tissue-profiles can theoretically be derived, the resource does not support gene–tissue queries for every tissue and gene therefore, it is considered tissue agnostic at this time.

DiseaseMeth is a DNA methylation repository that focuses on integration and in-depth analysis of DNA methylation data on human diseases.²³ Genes that are differentially methylated in diseases can be identified by querying this database, and reference methylation data for normal cells and tissues are available. This resource provides a wide range of search options, including gene–tissue pair for a selected disease, and thus can be referred as an epigenomic tissue profile resource.

HIstome is a database of human histones, their post-translational modifications (PTMs) and the enzymes resulting in these modifications.²⁴ Some entities are annotated by their association with human diseases. However, this database does not allow querying with a tissue entry, nor does it provide tissue-specific expression profiles, so it is a tissue-agnostic resource.

MethylomeDb is a database of human and mouse brain methylation profiles.²⁵ However, this database neither supports searching data for a specific tissue, nor provides tissue-specific DNA methylation profiles as output. Hence MethylomeDb is tissue-agnostic.

PROTEOMIC BIG DATA RESOURCES, CLASSIFIED BY TISSUE PROFILING POTENTIAL

Proteomics may have greater potential to link tissue variation to function. The proteomics analogue of GTEx is the Human Protein Atlas, where a direct study of proteomic variation across tissues was recently published.²⁶ The evaluable proteome is smaller than the genome, and the expressed product of many genes cannot be currently detected using antibodies. Nonetheless, this tissue profile resource is particularly valuable in integrating and visualizing, in an intuitive side-by-side manner, both gene and protein expression levels across tissues and cells.

The Human Proteome Map (HPM) uses a large-scale mass spectrometry survey to detect translated peptides that are then mapped back to genes using informatics. By this method, protein-coding genes, translated pseudogenes, noncoding RNAs and upstream open reading frames (ORFs) that have not yet been included in annotated protein databases are detected across various tissues.²⁷ Since this map exhibits a cell/tissue-specific expression of proteins, the data can be queried with gene–tissue pairs or a set of gene–tissue pairs. Thus, HPM is a tissue profile resource.

ProteomicsDb houses proteomics data that is curated from different repositories, peer-reviewed publications and in-house experimental studies. There is a component concerned with identification of a common core and tissue-specific proteome, followed by functional analyses with the help of protein-expression profiles in various human tissues, cell-lines, and body fluids.²⁸ It also focuses on identification of PTMs, translated lincRNAs and protein markers of resistance and sensitivity. As the resource provides tissue-expression profiles for a queried protein of interest, it is also a tissue profile resource.

The Global Proteome Machine Database (GPMDB) is a proteomics database that features a search engine, which can be queried for genes, proteins, cells, tissues, or any keyword. Users can submit data to the resource, which takes raw MS/MS data and reprocesses it using X! Tandem for protein identification analysis.²⁹ Predicted peptides are assigned a confidence score for their identification, and a visualization of their coverage can be viewed in a web interface. However, GPMDB does not calculate the abundance of a protein after identifying it, and will only output the number of observations of a protein for a tissue lookup. Still, it is considered a tissue profile resource.

PeptideAtlas is another peptide/protein identification resource and protein expression database. Using shotgun proteomics, it reprocesses MS/MS data to identify peptides and assign them to proteins with *P*-values.³⁰ Each protein and peptide is annotated with the number of its observations in all tissues. PeptideAtlas SRM Experiment Library (PASSEL) integrates selection reaction monitoring (SRM) methods into its workflow.³¹ This technique is particularly useful for quantification of pre-determined peptides in a sample. This allows

for measurements of protein abundance in samples, making PeptideAtlas a tissue profile resource.

The PRoteomics IDentifications (PRIDE) database is a public repository where users can upload and share their proteomics data, which can include peptide/protein identification, PTMs, and peptide/protein expression values.³² These data have analysis as originally provided by the researchers, making PRIDE a tissue-agnostic resource. Though tissue–protein expression values are present, they cannot be queried in a reliable way since all the data did not go through a common pipeline.

In the same respect, the Mass spectrometry Interactive Virtual Environment (MassIVE) (<https://massive.ucsd.edu/>) and Chorus Project (<https://chorusproject.org/>) are also tissue-agnostic resources. Both databases serve the purpose of storing and sharing of MS data, but do not allow for querying of tissue–protein pairs.

MaxQB is a repository for MS-based proteomic experiments and serves as a web interface for public research and analysis. MaxQB contains only data on experiments performed by the Mann laboratory, making it a very homogenous dataset. This allows for higher level analysis and more accurate quantification of the proteome. The database makes use of the MaxQuant engine to estimate expression of a protein in a sample.³³ Since the web interface allows for querying of protein expression profiles across tissues, this is a tissue profile resource.

Human Proteinpedia is a repository for proteomic data, where users can upload and share files. The database includes data not only for MS-based experiments, but also includes immunohistochemistry, protein/peptide arrays, yeast two-hybrid, and fluorescence microscopy.³⁴ All data annotation and analysis are provided by the researchers and does not get reprocessed. Human Proteinpedia allows for querying of proteins and tissues, but due to the heterogeneity of the data, this remains a tissue-agnostic resource.

The ProteomeXchange consortium was established to provide a single point of submission for MS-based experimental data and promote the public dissemination of proteomic data.³⁵ As of now, the consortium accepts data in the form of MS/MS proteomics through PRIDE and MassIVE, and SRM data through PASSEL. The web interface only allows users to browse raw or processed data provided by researchers, and there is currently no practical way to query for protein expression. Thus, this database is tissue-agnostic.

COMPOSITE/SPECIALIZED BIG DATA RESOURCES, CLASSIFIED BY TISSUE PROFILING POTENTIAL

The Allen Brain Atlas is a gene expression and anatomic database for brain tissues. The data include microarray and RNA-seq expression profiles for adult human and mouse brains. There is also an atlas for the developing human and mouse brains detailing gene expression from early embryonic stage through adulthood. Each data point is accompanied by structural imaging showing the 3D spatial origin of the sampled tissue.³⁶ While limited only to brain tissues, the Allen Brain Atlas is an important tissue profile resource. Notably, due to the

profiling technologies used, the focus on circuitry and the imaging aspect, this atlas is the most advanced of all the existing resources toward the futuristic goal of a ‘cytoreceptomics’ profile, namely maps of gene expression across the higher resolution of every cell in the human brain or body.

TISSUES is a visualization tool (plugin to BioGPS) wherein mRNA and protein expression profiles across all human tissues are normalized into a single profile.³⁷ The data are curated from various other resources, including GEO and the Human Protein Atlas, as well as text-mining of biochemical literature. A gene–tissue query will return a useful result, and this database is a tissue profile resource. Similar intuitive visualizations are available in other tools such as the Human Protein Atlas.

Project Achilles aims to enhance the understanding of genetic vulnerability of human cancer cells. The project houses data on the effect of gene alterations on cell viability and proliferation in cancer cells as well as lineage-specific genetic dependencies.³⁸ Though the project provides a suite of tools for data mining/analysis and a dataset that has phenotypic associations, the resource does not have a tissue entry for querying data. Thus, Project Achilles is tissue-agnostic.

MotifMap primarily focuses on the identification of TF binding sites in noncoding and nonrepeat regions to derive a comprehensive map of candidate regulatory motifs encoded in the human genome.³⁹ In addition, the resource also contains data on SNPs overlapping or neighboring motif sites. Since this motif site map is based on computational analysis and additional mapping studies are required to build a more comprehensive map, there is a long road ahead before such data can be used for specific gene expression studies. Hence, MotifMap is tissue-agnostic. The dataset cannot be queried with a gene–tissue pair.

TRANSFAC is a relational database that comprises a variety of data on TFs, their target genes, and regulatory binding sites. It is accompanied by CYTOMER, a database of anatomic structures and developmental stages, as well as a tool called Match, that together not only provide tissue-specific expression profiles, but also facilitate the creation and modification of custom profiles.⁴⁰ TRANSFAC does not have a direct provision to query gene–tissue pairs, but its integration with the CYTOMER database provides expression profiles for a queried TF or gene. Thus, it is a tissue profile resource, albeit one that suffers from the validation difficulties inherent to sequence-based TF binding site prediction.

The ChIP-X database is a collection of data obtained by various ChIP-X/ChIP-Seq experiments studying the binding profiles of TFs.⁴¹ This database is a part of ChEA, a web-based application that allows retrieval of gene targets for specific TFs and lists all TFs upstream of a specific gene. ChIP-X provides data on target genes shared by two or more TFs, suggesting the mechanism of regulation of gene expression in cells, and allowing the possibility of building networks that connect TFs; however, it is still a tissue-agnostic dataset. It does not allow queries with gene–tissue pairs.

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets including genes that have a common biological function, pathway, or regulation.⁴² The database is tissue agnostic, excluding any information on tissues. However, its value is in

analyzing the results of gene expression profile studies. Where single genes may not show up as statistically significant, multiple related genes will display a signal. The differential of these gene sets also carry a higher biological significance, due to their shared pathway. While it is not a tissue profile resource, MSigDB could be an important secondary tool for achieving specificity at the target-tissue profile level.

PhosphoMouse is a murine phosphorylation atlas that has been developed based on studies focusing on tissue-specific effects of post-translational regulation, involving differential phosphorylation of proteins.⁴³ This resource provides detailed information on abundance of phosphoproteins and phosphorylation sites. PhosphoMouse is a tissue profile resource that comprises data on differential phosphorylation patterns observed across tissues specific for a protein and their co-relation with cell-signaling networks. By separating tissue-specific phosphorylation from tissue-specific gene expression, unique connections between complex and variable intracellular signaling pathways and physiological differences among tissues may become apparent. This can further help in explaining tissue-specific disease and drug resistance. This resource supports gene–tissue queries, and the only reservation is the usual one: to what degree does the mouse model human physiology?

CircNet is a database of circular RNAs, their tissue-specific expression profiles and integrated microRNA regulatory networks. It also provides data on potential miRNAs sponged by a selected circRNA. Preliminary studies^{44,45} of circular noncoding RNAs have suggested their abundance, evolutionary conservation and potential role in posttranscriptional gene regulation. Some works⁴⁶ have also shown correlation of non-coding circular RNAs with certain diseases. Hence it seems that CircNet can prove to be a useful tool in further exploring the role of circRNAs in pathogenesis. CircNet is indeed a tissue profile resource, however, the dataset does not allow querying through a gene–tissue pair.

ILLUSTRATIVE EXAMPLE

A composite cross-tissue profile of human Glut1 (gene SLC2A1) reveals some concordance across resources, but also significant variation in the profiles resulting in different knowledge implications. For example, the Human Protein Atlas shows concordant, highest expression and protein levels of Glut1 in brain, skin, the mouth/throat and placenta (Figure 2(a)). Glut1 deficiency syndrome(OMIM link), a human genetic disease of Glut1, is characterized by neurological/movement phenotypes, and it is not hard to connect this phenotype with the expression in the brain, the most functionally, phenotypically sensitive of these organs. However, inspection of the histology data reveals a pattern of blood vessel expression (Figure 2(a)), and red blood cells are not a listed cell type from this resource. BioGPS, on the other hand shows significant expression in early erythroid cells (Figure 2(b)). The true *in vivo* profile may be that Glut1 is expressed in blood and *secondarily* is detected in the most oxygen needy organs, such as the brain. The clinical phenotype of this gene product maps to those organs and not to the cells in which the molecular defect is resident. This *in vivo* molecular profile of a human phenotype appears to be only available by integration of several tissue-specific resources and certainly not by genomic, epigenomic, transcriptomic or proteomic data alone, no matter how powerful the informatics method.

METHODS AND CHALLENGES

A variety of informatics approaches have historically been used to interpret Omics data, most commonly variations of machine learning and network analysis/graphical approaches. These approaches have traditionally been relatively agnostic to tissue specificity and have relied upon the mathematical approaches as a blunt tool to clarify data. There have been a number of recent attempts to incorporate tissue specificity into the assessment of molecular profiling signals.^{47–51} The main challenge is, not surprisingly, distinguishing signal from noise, that is, is the difference in expression levels of a gene between two different tissues functionally significant? This is also a challenge for tissue-agnostic analytic methods, so pursuing differential analysis between tissues may actually offer opportunities to improve the signal-to-noise ratio and better generate knowledge from such data.⁵

SUMMARY

Tissue atlases have the potential to integrate disparate big biomedical data into simple, intuitive and physiologically accurate profiles of complex *in vivo* phenotypes.⁵ We have reviewed the spectrum of tissue atlas-compliant big biomedical data resources (selected tissue-profile resources are shown in Table 1). Many of the existing resources are sufficiently tissue-target-profile-compliant to already enable diverse informatics and systems biology approaches to translate Omics molecular profiling data into medical knowledge, with suitable informatics approaches. Most common data science approaches, notably approach to decipher regulatory networks from Omics data and other such common computational biology approaches, tend to be tissue-agnostic and may benefit from adaptation to a cross-tissue approach in seeking *in vivo* applicability. One challenge is that tissue-target/gene-target profiles may vary significantly across different data sources or that true knowledge may require inference across the sources (Figure 2). New informatics methods to achieve this may significantly increase knowledge generation and *in vivo* molecular profiling of human phenotypes from the ever-growing volume of Omics data.

ACKNOWLEDGMENTS

This work was supported by NIH grant R43GM113546 (to KF). All authors except LMY are employees or consultants to GeneCentrix, Inc., which develops products based on some of the data sources cited in this work.

REFERENCES

1. Chao EC, Lipkin SM. Molecular models for the tissue specificity of DNA mismatch repair-deficient carcinogenesis. *Nucleic Acids Res* 2006, 34:840–852. doi:10.1093/nar/gkj489. [PubMed: 16464822]
2. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci U S A* 2007, 104:8685–8690. doi:10.1073/pnas.0701361104. [PubMed: 17502601]
3. Winter EE, Goodstadt L, Ponting CP. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res* 2004, 14:54–61. doi:10.1101/gr.1924004. [PubMed: 14707169]
4. Marais AD. Familial hypercholesterolaemia. *Clin Biochem Rev* 2004, 25:49–68. [PubMed: 18516203]

5. Shmelkov E, Grigoryan A, Swetnam J, Xin J, Tivon D, Shmelkov SV, Cardozo T. Historeceptomic fingerprints for drug-like compounds. *Front Physiol* 2015, 6:371. doi:10.3389/fphys.2015.00371. [PubMed: 26733872]
6. Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, Wong MP, Sham PC, Chanock SJ, Wang J. GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* 2012, 40:D1047–D1054. doi:10.1093/nar/gkr1182. [PubMed: 22139925]
7. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorf L, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014, 42(Database issue):D1001–D1006. doi:10.1093/nar/gkt1229. [PubMed: 24316577]
8. Hernandez DG, Nalls MA, Moore M, Chong S, Dillman A, Trabzuni D, Gibbs JR, Ryten M, Arepalli S, Weale ME, et al. Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol Dis* 2012, 47:20–28. [PubMed: 22433082]
9. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, Li H. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am J Hum Genet* 2013, 92:667–680. doi:10.1016/j.ajhg.2013.03.022. [PubMed: 23643380]
10. Consortium GTEx. The genotype-tissue expression (GTEx) project. *Nat Genet* 2013, 45:580–585. doi:10.1038/ng.2653. [PubMed: 23715323]
11. Consortium ENCODE Project. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011, 9:e1001046. doi:10.1371/journal.pbio.1001046. [PubMed: 21526222]
12. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, et al. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res* 2014, 42 (Database issue):D975–D979. doi:10.1093/nar/gkt1211. [PubMed: 24297256]
13. Wu C, Macleod I, Su AI. BioGPS and [MyGene.info](http://mygene.info): organizing online, gene-centric information. *Nucleic Acids Res* 2013, 41(Database issue):D561–D565. doi:10.1093/nar/gks1114. [PubMed: 23175613]
14. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* 2013, 41(Database issue):D991–D995. doi:10.1093/nar/gks1193. [PubMed: 23193258]
15. Dudley JT, Tibshirani R, Deshpande T, Butte AJ. Disease signatures are robust across tissues and experiments. *Mol Syst Biol* 2009, 5:307. doi:10.1038/msb.2009.66. [PubMed: 19756046]
16. McCall MN, Jaffee HA, Zelisko SJ, Sinha N, Hooiveld G, Irizarry RA, Zilliox MJ. The Gene expression barcode 3.0: improved data processing and mining tools. *Nucleic Acids Res* 2014, 42(Database issue): D938–D943. doi:10.1093/nar/gkt1204. [PubMed: 24271388]
17. Schmid PR, Palmer NP, Kohane IS, Berger B. Making sense out of massive data by going beyond differential expression. *Proc Natl Acad Sci U S A* 2012, 109:5594–5599. doi:10.1073/pnas.1118792109. [PubMed: 22447773]
18. Seita J, Sahoo D, Rossi DJ, Bhattacharya D, Serwold T, Inlay MA, Ehrlich LI, Fathman JW, Dill DL, Weissman IL. Gene expression commons: an open platform for absolute gene expression profiling. *PLoS One* 2012, 7:e40321. doi:10.1371/journal.pone.0040321. [PubMed: 22815738]
19. Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, et al. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* 2014, 42(Database issue):D926–D932. doi:10.1093/nar/gkt1270. [PubMed: 24304889]
20. Liu X, Yu X, Zack DJ, Zhu H, Qian J. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics* 2008, 9:271. doi:10.1186/1471-2105-9-271. [PubMed: 18541026]
21. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol* 2010, 28:1045–1048. doi:10.1038/nbt1010-1045. [PubMed: 20944595]
22. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human epigenomes. *Nature* 2015, 518:317–330. doi:10.1038/nature14248. [PubMed: 25693563]

23. Lv J, Liu H, Su J, Wu X, Liu H, Li B, Xiao X, Wang F, Wu Q, Zhang Y. DiseaseMeth: a human disease methylation database. *Nucleic Acids Res* 2012, 40(Database issue):D1030–D1035. doi: 10.1093/nar/gkr1169. [PubMed: 22135302]
24. Khare SP, Habib F, Sharma R, Gadewal N, Gupta S, Galande S. H1stome – a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res* 2012, 40(Database issue): D337–D342. doi:10.1093/nar/gkr1125. [PubMed: 22140112]
25. Xin Y, Chanrion B, O'Donnell AH, Milekic M, Costa R, Ge Y, Haghighi FG. MethylomeDB: a database of DNA methylation profiles of the brain. *Nucleic Acids Res* 2012, 40(Database issue):D1245–D1249. doi:10.1093/nar/gkr1193. [PubMed: 22140101]
26. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. Proteomics. Tissue-based map of the human proteome. *Science* 2015, 347:1260419. doi:10.1126/science.1260419. [PubMed: 25613900]
27. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. A draft map of the human proteome. *Nature* 2014, 509:575–581. doi: 10.1038/nature13302. [PubMed: 24870542]
28. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. Mass-spectrometry-based draft of the human proteome. *Nature* 2014, 509:582–587. doi:10.1038/nature13319. [PubMed: 24870543]
29. Beavis RC. Using the global proteome machine for protein identification. *Methods Mol Biol* 2006, 328:217–228. doi:10.1385/1-59745-026-X:217. [PubMed: 16785652]
30. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. *Nucleic Acids Res* 2006, 34(Database issue):D655–D658. doi:10.1093/nar/gkj040. [PubMed: 16381952]
31. Farrah T, Deutsch EW, Kreisberg R, Sun Z, Campbell DS, Mendoza L, Kusebauch U, Brusniak MY, Huttenhain R, Schiess R, et al. PASSEL: the PeptideAtlas SRM experiment library. *Proteomics* 2012, 12:1170–1175. doi:10.1002/pmic.201100515. [PubMed: 22318887]
32. Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, et al. The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* 2013, 41(Database issue): D1063–D1069. doi:10.1093/nar/gks1262. [PubMed: 23203882]
33. Schaab C, Geiger T, Stoeckl G, Cox J, Mann M. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol Cell Proteomics* 2012, 11: M111.014068. doi:10.1074/mcp.M111.014068.
34. Kandasamy K, Keerthikumar S, Goel R, Mathivanan S, Patankar N, Shafreen B, Renuse S, Pawar H, Ramachandra YL, Acharya PK, et al. Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Res* 2009, 37(Database issue): D773–D781. doi:10.1093/nar/gkn701. [PubMed: 18948298]
35. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* 2014, 32:223–226. doi:10.1038/nbt.2839. [PubMed: 24727771]
36. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de Lagemaat LN, Smith KA, Ebbert A, Riley ZL, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 2012, 489:391–399. doi:10.1038/nature11405. [PubMed: 22996553]
37. Santos A, Tsaou K, Stolte C, Pletscher-Frankild S, O'Donoghue SI, Jensen LJ. Comprehensive comparison of large-scale tissue expression datasets. *PeerJ* 2015, 3:e1054. doi:10.7717/peerj.1054. [PubMed: 26157623]
38. Cowley GS, Weir BA, Vazquez F, Tamayo P, Scott JA, Rusin S, East-Seletsky A, Ali LD, Gerath WF, Pantel SE, et al. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci Data* 2014, 1:140035. doi:10.1038/sdata.2014.35. [PubMed: 25984343]
39. Xie X, Rigor P, Baldi P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics* 2009, 25:167–174. doi:10.1093/bioinformatics/btn605. [PubMed: 19017655]

40. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003, 31:374–378. [PubMed: 12520026]
41. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 2010, 26:2438–2444. doi:10.1093/bioinformatics/btq466. [PubMed: 20709693]
42. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005, 102:15545–15550. doi:10.1073/pnas.0506580102. [PubMed: 16199517]
43. Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* 2010, 143: 1174–1189. doi:10.1016/j.cell.2010.12.001. [PubMed: 21183079]
44. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 2013, 19:141–157. doi:10.1261/rna.035667.112. [PubMed: 23249747]
45. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013, 495:333–338. doi:10.1038/nature11928. [PubMed: 23446348]
46. Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet* 2010, 6:e1001233. doi:10.1371/journal.pgen.1001233. [PubMed: 21151960]
47. Antanaviciute A, Daly C, Crinnion LA, Markham AF, Watson CM, Bonthon DT, Carr IM. GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles. *Bioinformatics* 2015, 31:2728–2735. doi:10.1093/bioinformatics/btv196. [PubMed: 25861967]
48. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc Natl Acad Sci U S A* 2008, 105:20870–20875. doi:10.1073/pnas.0810772105. [PubMed: 19104045]
49. Chen Y, Wang W, Zhou Y, Shields R, Chanda SK, Elston RC, Li J. In silico gene prioritization by integrating multiple data sources. *PLoS One* 2011, 6: e21137. doi:10.1371/journal.pone.0021137. [PubMed: 21731658]
50. Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 2007, 23:215–221. doi:10.1093/bioinformatics/btl569. [PubMed: 17098772]
51. Nitsch D, Tranchevent LC, Goncalves JP, Vogt JK, Madeira SC, Moreau Y. PINTA: a web server for network-based gene prioritization from expression data. *Nucleic Acids Res* 2011, 39(Web Server Issue): W334–W338. doi:10.1093/nar/gkr289. [PubMed: 21602267]

FIGURE 1.

(a) Inverse relationship of Omics data volume and causality (PTMome = post-translational modification-ome). (b) Relationships of GWAS and true biological entities.

FIGURE 2. Cross-tissue expression profile of human Glut1 (gene SLC2A1) from the Human Protein Atlas (a) and BioGPS (b), arrow indicates expression in early erythroid cells.

TABLE 1 |

Selected Tissue-Profile Resources

Dataset Name	Type of Data	Source of Data	Nature of Data	Type of Study Sample		
				Human/ Animal	Healthy/ Diseased	Tissue Type
Human Protein Atlas	Protein and mRNA expression profiles	Proteomics Microarray, RNA-seq, immunohistochemistry, immunofluorescence	Processed data	Human and mouse	Healthy and diseased	Tissues, primary cells, cell lines
The Human Proteome Map (HPM)	Protein/peptide expression profiles	Mass spectrometry	Processed data	Human	Healthy	Tissues, primary cells
ProteomicsDb	Protein/peptide expression profiles	Mass spectrometry	Processed data	Human	Healthy and diseased	Tissues, primary cells, cell lines
Global Proteome Machine Database (GPMDB)	Protein/peptide identification	Mass spectrometry	Processed data	Human and animal	Healthy and diseased	Tissues, primary cells, cell
PeptideAtlas and PeptideAtlas SRM Experiment Library (PASSEL)	Protein/peptide identification	Mass spectrometry	Processed data	Human and animal	Healthy and diseased	Tissues, primary cells, cell lines
PRoteomics IDentifications (PRIDE)	Protein/peptide identification, post-translational modifications	Mass spectrometry	Raw and processed data	Human and animal	Healthy and diseased	Tissues, primary cells, cell lines
MaxQB	Protein identification and expression profiles	Mass spectrometry	Processed data	Human and mouse	Healthy and diseased	Tissues (mouse), cell lines (human)
Roadmap Epigenomics	DNA accessibility profiles, DNA methylation profiles, mRNA expression profiles, histone modification profiles	Genomics DNase-seq, bisulfite sequencing, MeDIP-seq, MRE-seq, RNA-seq, ChIP-seq	Processed data	Human	Healthy and diseased	Tissues, primary cells, cell lines, hybrid cell lines
BioGPS	mRNA expression profiles	Whole-genome microarrays	Processed data	Human and animal	Healthy and diseased	Tissues, primary cells, cell lines
Gene Expression Omnibus (GEO)	Profiles of gene expression, methylation, genome variation, genome binding/occupancy, non-coding RNA	Microarray, RNA-seq, DNase-seq, ChIP-seq, Methylseq, Next-gen sequencing	Processed data	Human and animal	Healthy and diseased	Tissues, primary cells, cell lines, hybrid cell lines
Genotype-Tissue Expression (GTEx)	eQTL data, gene expression across tissues and genotypes	Microarray, RNA-seq, whole-genome sequencing	Processed data	Human	Healthy	Tissues, cell lines
Expression Atlas	Gene expression profiles	Other datasets: Gene Expression Omnibus (GEO), ArrayExpress	Processed data	Human and animal	Healthy	Tissues, cell lines
Encyclopedia of DNA Elements (ENCODE)	Transcription factor binding sites, histone modifications, DNA methylation	Microarray, RNA-seq, DNase-seq, ChIP-seq, methylseq	Processed data	Human and animal	Healthy and diseased	Tissues, primary cells, cell lines
Tissue-specific Gene Expression and Regulation (TIGER)	Gene expression profiles, transcription factor interactions,	Computationally predicted	Raw and processed data	Human	Healthy	Tissues

Dataset Name	Type of Data	Source of Data	Nature of Data	Type of Study Sample		
				Human/ Animal	Healthy/ Diseased	Tissue Type
Cancer Cell Line Encyclopedia (CCLE)	<i>cis</i> -regulatory modules, expressed sequence tags	Expression microarray, SNP array, genome sequencing	Raw and Processed data	Human	Diseased	Cell Lines
	Profiles of gene expression, chromosomal copy number, sequencing data	Composite/specialized datasets				
Allen Brain Atlas	Anatomic and gene expression profiles for brain tissues, neural connections	Microarray, RNA-seq, and <i>in situ</i> hybridization data	Raw and Processed data	Human and animal	Healthy and diseased	Tissues, cell lines
TISSUES	Gene and protein expression profiles	Text mining, microarray, RNA-seq, immunohistochemistry, mass spectrometry	Processed data	Human	Healthy	Tissues
PhosphoMouse	Phosphoprotein expression profiles, tissue-specific phosphorylation	Mass spectrometry	Raw data	Mouse	Healthy	Tissues
CircNet	circRNA expression profiles, circRNA-miRNA-gene regulatory networks	Reported back-spliced junction sites, RNA-seq samples from the NCBI Sequence Read Archive	Processed data	Human	Healthy and diseased	Tissues, primary cells, cell lines