ANALYSIS

nature genetics

Understanding multicellular function and disease with human tissue-specific networks

Casey S Greene[1–3,13], Arjun Krishnan[4,13], Aaron K Wong[5,13], Emanuela Ricciotti[6,7], Rene A Zelaya[1], Daniel S Himmelstein[8], Ran Zhang[9], Boris M Hartmann[10], Elena Zaslavsky[10], Stuart C Sealfon[10], Daniel I Chasman[11], Garret A FitzGerald[6,7], Kara Dolinski[4], Tilo Grosser[6,7] & Olga G Troyanskaya[4,5,12]

Tissue and cell-type identity lie at the core of human physiology and disease. Understanding the genetic underpinnings of complex tissues and individual cell lineages is crucial for developing improved diagnostics and therapeutics. We present genome-wide functional interaction networks for 144 human tissues and cell types developed using a data-driven Bayesian methodology that integrates thousands of diverse experiments spanning tissue and disease states. Tissue-specific networks predict lineage-specific responses to perturbation, identify the changing functional roles of genes across tissues and illuminate relationships among diseases. We introduce NetWAS, which combines genes with nominally significant genome-wide association study (GWAS) *P* values and tissue-specific networks to identify disease-gene associations more accurately than GWAS alone. Our webserver, GIANT, provides an interface to human tissue networks through multi-gene queries, network visualization, analysis tools including NetWAS and downloadable networks. GIANT enables systematic exploration of the landscape of interacting genes that shape specialized cellular functions across more than a hundred human tissues and cell types.

The precise actions of genes are frequently dependent on their tissue context, and human diseases result from the disordered interplay of tissue- and cell lineage–specific processes[1–4]. These factors combine to make the understanding of tissue-specific gene functions, disease pathophysiology and gene-disease associations particularly challenging. Projects such as the Encyclopedia of DNA Elements (ENCODE)[5] and The Cancer Genome Atlas (TCGA)[6] provide comprehensive genomic profiles for cell lines and cancers, but the challenge of understanding human tissues and cell lineages in the multicellular context of a whole organism remains[7]. Integrative methods that infer functional gene interaction networks can capture the interplay of pathways, but existing networks lack tissue specificity[8].

Although direct assay of tissue-specific features remains infeasible in many normal human tissues, computational methods can infer these features from large data compendia. We recently found that even samples measuring mixed cell lineages contain extractable information related to lineage-specific expression[9]. In addition to tissue specificity, we[10–13] and others[14–17] have shown that heterogeneous genomic data contain functional information, for example, of gene expression regulation by protein-DNA, protein-RNA, protein-protein and metabolite-protein interactions. Here we develop and evaluate methods that simultaneously extract functional and tissue or cell-type signals to construct accurate maps of both where and how proteins act.

We build genome-scale functional maps of human tissues by integrating a collection of data sets covering thousands of experiments contained in more than 14,000 distinct publications. To integrate these data, we automatically assess each data set for its relevance to each of 144 tissue- and cell lineage–specific functional contexts. The resulting functional maps provide a detailed portrait of protein function and interactions in specific human tissues and cell lineages ranging from B lymphocytes to the renal glomerulus and the whole brain. This approach allows us to profile the specialized function of genes in a high-throughput manner, even in tissues and cell lineages for which no or few tissue-specific data exist.

In contrast with tissue-naive networks, which assume that the function of genes remains constant across tissues[8], these maps can answer biological questions that are specific to a single gene in a single tissue. For example, we use these maps for the gene *IL1B* (encoding interleukin (IL)-1β) in the blood vessel network, where it has a key role in inflammation[18], to predict lineage-specific responses to IL-1β stimulation, which we experimentally confirm. Examination of parallel networks shows changes in gene and pathway functions and

[1]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA. [2]Dartmouth-Hitchcock Norris Cotton Cancer Center, Lebanon, New Hampshire, USA. [3]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, New Hampshire, USA. [4]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA. [5]Department of Computer Science, Princeton University, Princeton, New Jersey, USA. [6]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [7]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [8]Biology and Medical Informatics, University of California, San Francisco, San Francisco, California, USA. [9]Department of Molecular Biology, Princeton University, Princeton, New Jersey, USA. [10]Department of Neurology, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [11]Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [12]Simons Center for Data Analysis, Simons Foundation, New York, New York, USA. [13]These authors contributed equally to this work. Correspondence should be addressed to O.G.T. (ogt@genomics.princeton.edu).

Received 10 December 2014; accepted 6 March 2015; published online 27 April 2015; doi:10.1038/ng.3259

interactions across tissues, identifying tissue-specific rewiring. We demonstrate that several tissue-specific functions of *LEF1*, encoding the multifunctional lymphoid enhancer-binding factor 1, are evident from the way its connectivity changes in distinct tissues.

Tissue-specific networks provide a new means to generate hypotheses related to the molecular basis of human disease. We developed an approach, termed network-wide association study (NetWAS). In NetWAS, the statistical associations from a standard GWAS[19] guide the analysis of functional networks. This reprioritization method is driven by discovery and does not depend on prior disease knowledge. NetWAS, in conjunction with tissue-specific networks, effectively reprioritizes statistical associations from distinct GWAS to identify disease-associated genes, and tissue-specific NetWAS better identifies genes associated with hypertension than either GWAS or tissue-naive NetWAS.

Our tissue-specific maps are available through the Genome-Scale Integrated Analysis of Networks in Tissues (GIANT) interface, which provides interactive visualization and exploration of tissue-specific networks, including a comparative view that can highlight the tissue-specific rewiring of genes and pathways. GIANT also provides NetWAS analysis for biomedical researchers to reprioritize their gene-based GWAS results in the context of our human tissue-specific networks.

RESULTS

We integrated diverse genome-scale data in a tissue-specific manner to construct 144 human tissue- and cell lineage–specific networks and demonstrated their broad usefulness for generating specific, testable hypotheses, summarizing tissue-specific relationships between diseases and reprioritizing results from genetic association studies (**Fig. 1a**). Our findings underscore the importance of considering tissue specificity when integrating heterogeneous data to understand the pathophysiology of common human diseases.

Integrated tissue-specific functional interaction networks

We isolated tissue-relevant signals from data not resolved to a particular cell lineage or tissue using a Bayesian integration that incorporated the hierarchical relationships between tissues. We collected tissue-specific functional interactions for each tissue from known functional relationships and low-throughput tissue-specific gene expression data, and we mapped tissue-specific gene annotations from the Human Protein Reference Database[20] (HPRD) to the BRENDA Tissue Ontology[21] (BTO). We leveraged this hierarchy to increase gene and tissue coverage and to make the interactions consistent with tissue organization (Online Methods). Using these known tissue-specific interactions, we constructed a Bayesian model of tissue-specific functional information from diverse experiments for each of 144 human tissues. Each tissue network represents the tissue-specific posterior probability of a functional relationship between each pair of genes from an ensemble of data covering more than 14,000 publications (**Fig. 1a**).

Our approach accurately identified tissue-relevant signals in the compendium (**Fig. 1b** and **Supplementary Table 1**), automatically up-weighting data sets from relevant tissues and prioritizing tissue-relevant signals over other data. According to fivefold cross-validation, our method outperformed a Bayesian integration limited to only tissue-related data sets identified on the basis of the experimental description (for 62 of 64 tissues; $P = 3.2 \times 10^{-12}$; **Supplementary Fig. 1**). Our approach also substantially increased the number of tissues for which networks could be constructed. Only 64 tissues had sufficient labeled data to construct networks, but we were able to construct networks for 144 tissues by extracting tissue-specific information from hundreds of data sets. For example, our method constructed a network for the dentate gyrus (a tissue with limited data) by taking advantage of curated dentate gyrus–specific knowledge to extract relevant signals from those for other tissues and cell types in the nervous system. Networks for tissues with no or a very limited amount of data had accuracies comparable to those for tissues with abundant tissue-specific data (**Supplementary Fig. 1**). Our approach generated diverse networks that reflect the tissue-specific connectivity of genes and pathways (**Supplementary Table 2**).

Tissue-specific networks predict IL-1β response

Our networks provided experimentally testable hypotheses about tissue-specific gene function and responses to pathway perturbations. We examined and experimentally verified the tissue-specific molecular response of blood vessel cells to stimulation by IL-1β (*IL1B*), a proinflammatory cytokine. We anticipated that the genes most tightly connected to *IL1B* in the blood vessel network would be among those responding to IL-1β stimulation in blood vessel cells (**Fig. 2a**). We tested this hypothesis by profiling the gene expression of human aortic smooth muscle cells (HASMCs; the predominant cell type in blood vessels) stimulated with IL-1β. Examination of the genes whose expression was significantly upregulated at 2 h after stimulation showed that 18 of the 20 *IL1B* network neighbors were among the top 500 most upregulated genes in the experiment ($P = 2.07 \times 10^{-23}$; **Fig. 2b**). The blood vessel network was the most accurate tissue network in predicting this experimental outcome; none of the other 143 tissue-specific networks or the tissue-naive network performed as well when evaluated by each network's ability to predict the result of IL-1β stimulation on the cells (**Fig. 2a**). Nine of the top 20 highest–edge weight neighbors of *IL1B* in the blood vessel network were not top neighbors in the tissue-naive network, and each has a key vasculature-specific role (**Supplementary Table 3** and **Supplementary Note**). Networks of other cardiovascular system tissues also captured IL-1β response better than the tissue-naive network, and this was consistent across a range of thresholds for top *IL1B* network neighbors as well as upregulated genes in the experiment (**Supplementary Fig. 2b**,**c**).

We also evaluated nine additional publicly available data sets that used modern genome-wide platforms to measure cellular response to IL-1β stimulation in a diverse set of tissues. In all ten experiments, the appropriate tissue network identified a set of genes that significantly responded to IL-1β, and randomly selected control sets of genes did not show a significant response to treatment (**Supplementary Fig. 3**).

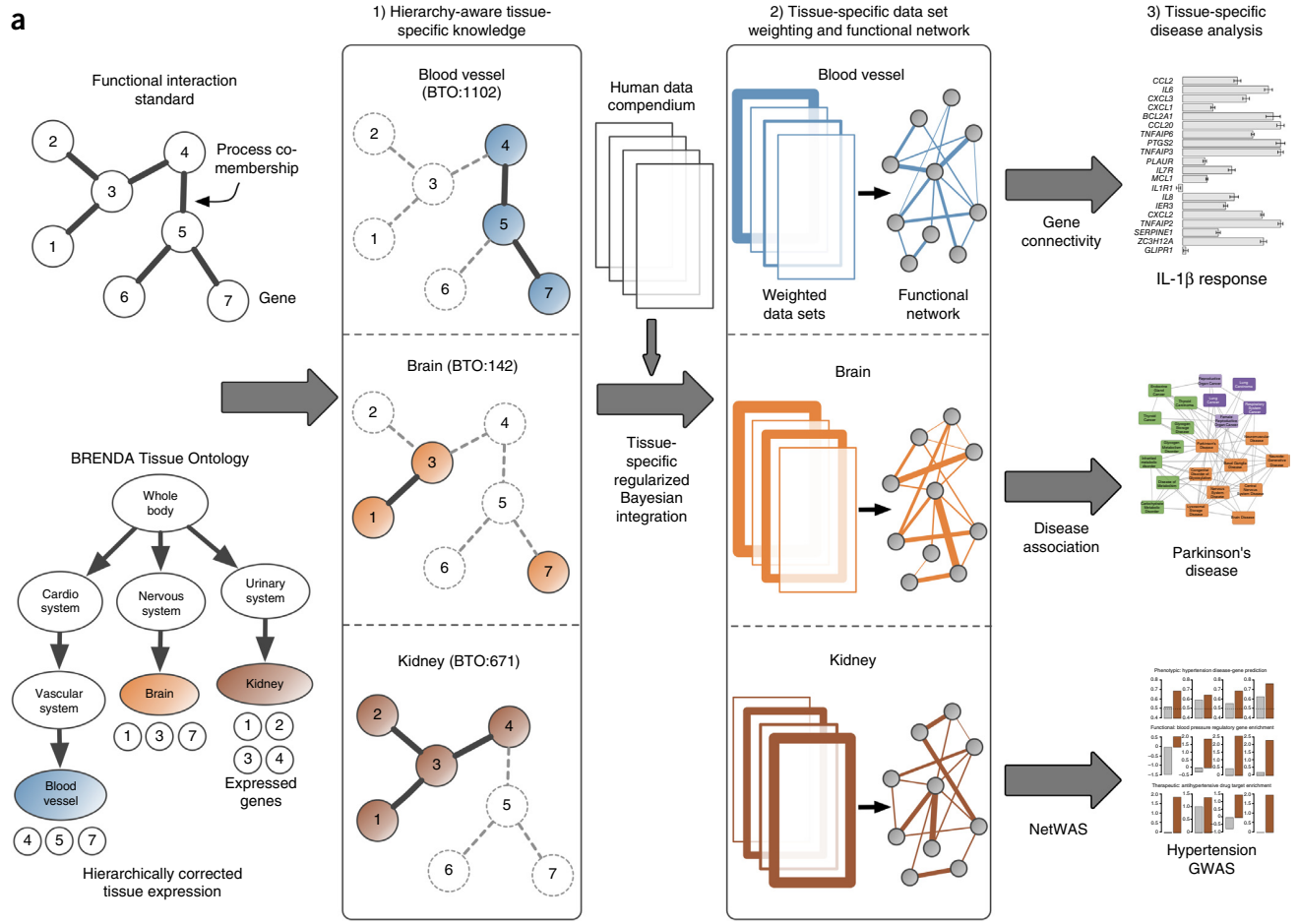
Tissue-specific network rewiring of multifunctional genes

Complex multicellular organisms have multifunctional genes that participate in distinct cellular processes according to developmental and anatomical context. For example, developmental programs are known to be controlled by broadly expressed transcription factors that, in specific combinations, regulate cell type–specific gene expression[5,22,23] and have the potential to force cell lineage conversions[24,25]. Multifunctional genes have also been implicated in pleiotropic disease phenotypes[26,27]. Such effects are likely to arise when a gene is 'rewired' to associate with different functional partners in different tissues. Our genome-wide functional network maps of human tissues could potentially delineate the tissue-specific wiring of multifunctional genes.

We focused on the transcription factor *LEF1*, which has a key role in mediating the tissue-specific response to Wnt signaling[28], a fundamental and highly conserved pathway known to elicit diverse cell type–specific developmental responses[29]. Because very little is known about the activity of *LEF1* in human tissues, we probed the

tissue-specific functional role of this transcription factor by examining the neighbors of *LEF1* across different networks (**Fig. 3a** shows neighbors in B lymphocytes, hypothalamus, osteoblasts and trachea; **Supplementary Fig. 4** shows a detailed view of *LEF1* in the B-lymphocyte network). Analysis of *LEF1* network partners (Online Methods) showed that, in 12 tissues, *LEF1* was significantly associated with processes relevant to each tissue (area under the receiver operating characteristic curve (AUC) ≥ 0.8; **Fig. 3b**), reflecting highly accurate representation of tissue-specific wiring of *LEF1*.

In addition to recapitulating current knowledge (represented by solid blue edges in **Fig. 3b**), we identified several new associations between *LEF1* and tissue-specific processes in humans that have



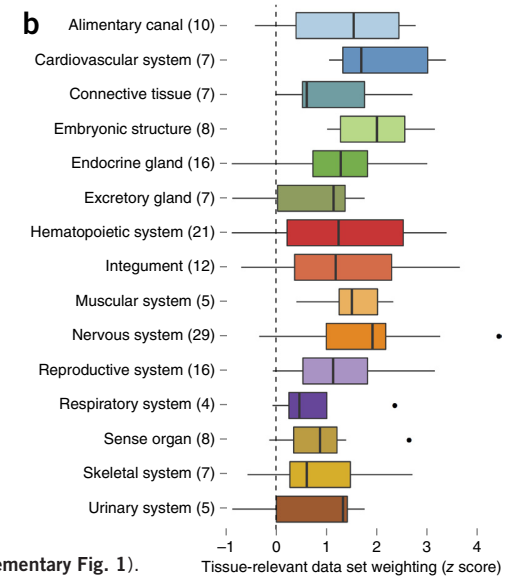


Figure 1 Regularized Bayesian integration based on tissue ontology. (**a**) Our integration pipeline constructs tissue-specific functional interaction networks by (1) using tissue-specific knowledge to (2) identify and weight data sets by their tissue-relevant signal. (3) We demonstrate the capabilities of the networks by experimentally validating the gene connectivity scores (top), demonstrating that they identify disease associations (middle) and reprioritizing GWAS results (bottom). (**b**) Bayesian integration using tissue-specific knowledge automatically identifies and weights tissue-relevant data sets. We validate our approach by evaluating the weights in a set of data sets with clear tissue specificity. We calculate a *z* score for each tissue that measures how much the 'relevant' data sets are up-weighted relative to all data sets in the compendium for that tissue. Plotted here per organ system (*y* axis) is the distribution of *z* scores (*x* axis) of tissues within that system in the form of a box plot. The thick line within each box indicates the median tissue *z* score for that system; the lower and upper ends of the box indicate the first and third quartiles of the distribution; and the extended lines on either side denote the limits of the distribution, with the outliers (dots) further away. Beyond automatically identifying relevant data sets, our method of automatic weighting constructed higher-quality networks than an identical approach limited to only curation-identified tissue-relevant data sets (**Supplementary Fig. 1**).

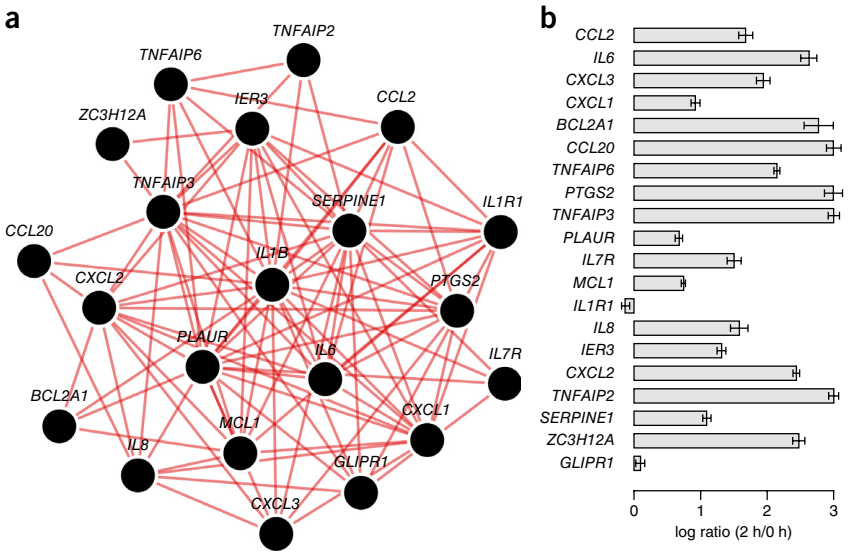


Figure 2 Predicted *IL1B* functional interaction partners from the blood vessel network are significantly upregulated after stimulation of blood vessel cells with IL-1β. (**a**) The 20 genes most tightly connected to *IL1B* in the blood vessel network are shown. These genes are predicted to respond to IL-1β stimulation in blood vessel. (**b**) The bar plot shows the differential expression levels of the 20 *IL1B* neighbors measured in a microarray experiment at 0 h and 2 h after IL-1β stimulation in HASMCs, which constitute most of the blood vessel. Each bar represents the gene's log ratio of mean expression at 2 h over its mean expression at 0 h. Error bars represent regularized pooled standard errors estimated by LIMMA ($n = 4$ biological replicates). Eighteen of the 20 *IL1B* network neighbors (labeled in bold) were found to be among the most significantly differentially expressed genes at 2 h relative to 0 h ($P = 1.95 \times 10^{-23}$).

experimental support in model organisms (represented by dotted red edges in **Fig. 3b**). As for the four tissues in **Figure 3a**, we highlighted predicted functional associations of *LEF1* in B lymphocytes, hypothalamus, osteoblasts and trachea (red, orange, green and purple, respectively, in **Fig. 3b**). The role of *Lef1* in B cell activation has already been characterized in mouse[30,31]; further, *LEF1* has been strongly linked to chronic lymphocytic leukemia (CLL)[32–34]. Lef1-mediated Wnt signaling has been shown to be critical for hypothalamic neurogenesis in zebrafish[35,36]. Numerous studies point to a pivotal role for *LEF1* in osteoblast proliferation, maturation, function and regeneration[37–39] and to its potential involvement in bone disease[40,41]. Finally, several animal models support a clear association of *LEF1* with the development of submucosal glands[42–44], which are epithelial secretory structures in the human tracheobronchial airways, involved in hypersecretory lung diseases such as asthma, chronic bronchitis and cystic fibrosis[45]. Thus, tissue-specific networks can unravel the distinct functions of multifunctional genes such as *LEF1* and provide opportunities to probe the tissue-specific pleiotropic effects of disease-associated mutations.

Tissue networks can capture disease-disease associations

Most human diseases are syndromes with complex origins and manifestations in multiple tissues[3,4,26]. Diseases with common causative pathways or that are connected through crosstalk between pathways are expected to exhibit high levels of functional association in their relevant tissues[26]. We used the tissue-specific networks to quantify molecular interactions between diseases to derive a map of tissue-specific disease relationships. These were data-driven maps discovered from tissue-specific functional associations inferred from an integration of high-throughput data, making them relatively unbiased with respect to prior knowledge of disease associations. Here we focused on Parkinson's disease, a neurodegenerative disorder caused by progressive neuronal loss in the substantia nigra and subsequent reduction in dopamine production[46]. We created a functional disease map of Parkinson's disease based on the substantia nigra network (**Fig. 4**). Several documented disease associations were observed in the disease map: for example, Parkinson's disease was connected to 'neurodegenerative disease' and 'basal ganglion disease', classes of disease that include Parkinson's disease. The disease map also contained more subtle connections. For instance, Parkinson's disease was strongly connected to both lung and reproductive organ cancers, likely through the ubiquitin-protein ligase gene *PARK2*, which has been implicated in Parkinson's disease as well as brain, colorectal, lung and ovarian cancers[47,48]. We observed additional undocumented connections to thyroid cancer, driven by functional interactions involving the Parkinson's disease–associated genes *PARK2*, *PARK7* and *HTRA2*. A blinded literature evaluation showed that this disease map was significantly enriched (Fisher's exact test, $P = 0.001228$) for associations strongly supported by the literature as compared to a control set of associations (**Supplementary Fig. 5**). Thus, modeling complex diseases in humans using tissue-specific networks provided several insights into disease genetics and crosstalk and highlighted avenues for the discovery of new molecular disease associations. We generated additional disease maps for Alzheimer's disease, glomerulonephritis and glycogen metabolism disorder (**Supplementary Fig. 6**).

Tissue networks are tools for data-driven analysis of GWAS

In the last decade, quantitative genetics—particularly GWAS—has emerged as a powerful approach to catalog heritable and *de novo* sequence variation associated with a wide range of human traits and diseases[49]. However, owing to the lack of statistical power to detect low-frequency mutations, small genetic effects and epistatic interactions, GWAS findings usually only account for a small proportion of the observed heritability[50]. Because most complex diseases have tissue-specific origins and manifestations, we hypothesized that tissue-specific networks could complement GWAS data in discovering disease-gene associations. Top GWAS hits, even those below a reasonable genome-wide significance cutoff, should be enriched with relevant (even 'real', causal) genes. Consequently, by identifying the functional signatures associated with these top genes in the appropriate tissue-specific networks, we can further enrich for phenotype-associated genes in a genome-wide re-ranking of the GWAS results. Thus, we developed the network-wide NetWAS approach consisting of a tissue network classifier that learns the network connectivity patterns associated with the phenotype of interest (using the top GWAS hits) and makes predictions for genes across the genome. The NetWAS approach is discovery driven, as the genes used to identify connectivity patterns are derived from the GWAS itself rather than from potentially biased or limited prior disease knowledge. This attribute allows NetWAS to be applied to any GWAS, even

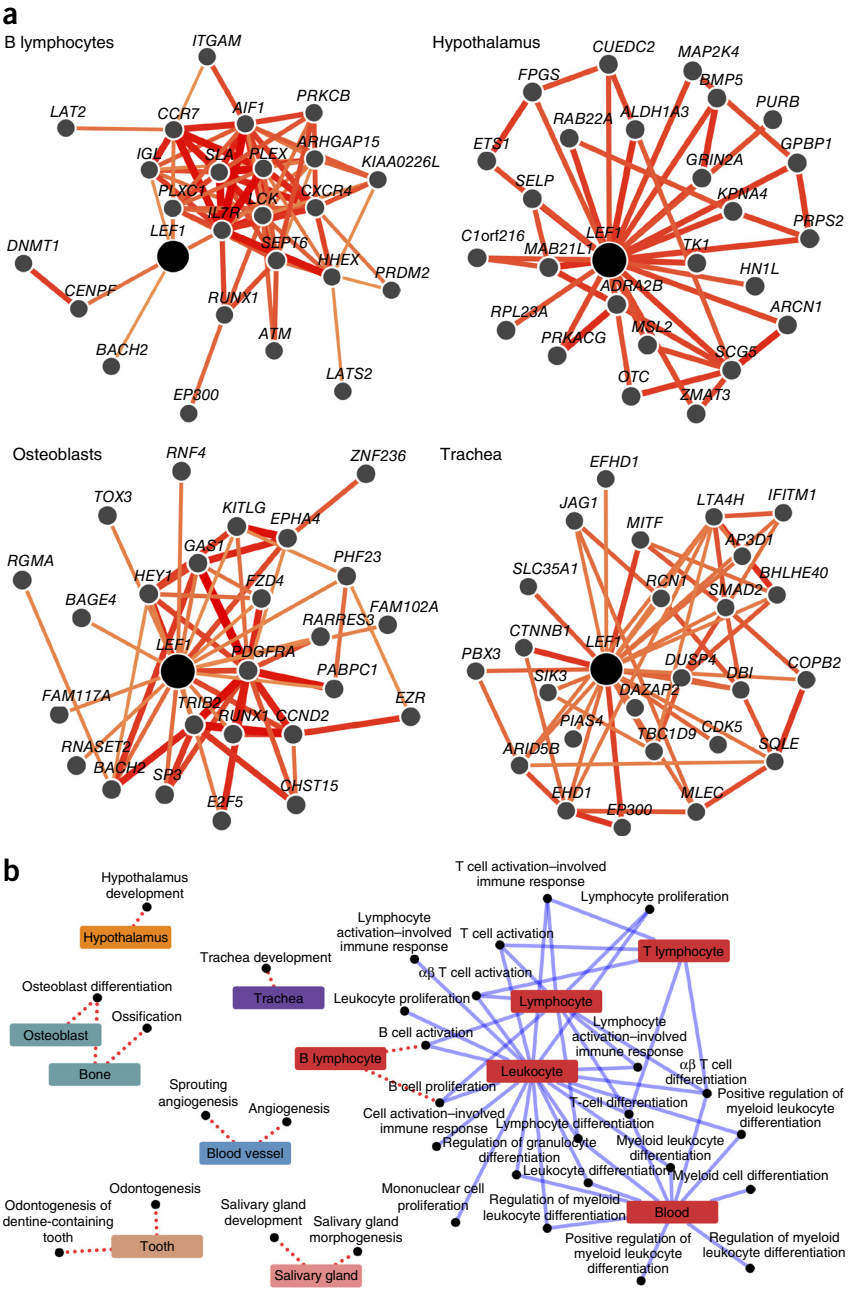


Figure 3 Tissue networks capture tissue-specific functional rewiring. (**a**) Multi-tissue view of *LEF1* retrieved from GIANT, a web interface to our tissue-specific networks that facilitates user-directed analysis of human tissue networks. Using the advanced functionality for comparing functional interactions across tissues, we queried GIANT with the multifunctional gene *LEF1* (encoding lymphoid enhancer-binding factor 1) in four tissues: B lymphocytes, hypothalamus, osteoblasts and trachea. The retrieved functional neighbors of *LEF1* were indeed notably different across the four networks, leading us to the hypothesis that the tissue networks could recapitulate tissue-specific gene wiring. Edge thickness and color correspond to edge strength (color scale in **Supplementary Fig. 4**). (**b**) The diverse tissue-specific functional rewiring of *LEF1*. This bipartite graph of tissues (colored rectangles) and processes (black circles) shows how *LEF1* participates in different processes in distinct tissues. For example, in the blood vessel, *LEF1* is most closely associated with angiogenesis, but in hypothalamus it is closely associated with hypothalamus development. In addition to displaying prior knowledge about the tissue-specific associations of *LEF1* (solid blue edges), the tissue networks also aid in the discovery of several new tissue associations that have experimental support in model organisms (dotted red edges).

those probing currently uncharacterized or minimally characterized diseases and phenotypes as well as those for which no associations reached genome-wide significance.

We applied the NetWAS strategy to a GWAS of hypertension[51,52] (Online Methods). Hypertension is a major cardiovascular risk factor and a complex trait involving a large number of genetic variants[53]. We converted SNP-level association statistics into gene-level statistics[19] for each of three recorded phenotypes—diastolic blood pressure (DBP), systolic blood pressure (SBP) and hypertension. We applied support vector machines, using genes with nominally significant *P* values as positive examples and randomly selected genes as negative examples. Using the tissue-specific network for kidney, a tissue that has a central role in blood pressure control[54], the approach constructed a classifier that identified tissue-specific network connectivity patterns associated with the phenotype of interest. Genes annotated to hypertension phenotypes in the Online Mendelian Inheritance in Man (OMIM) database[55] were more highly ranked by this classifier than by the initial GWAS (**Fig. 5a**). We hypothesized that the distinct endpoints might reflect different aspects of the disease. We evaluated whether combining the three phenotypes using a rank-sum approach resulted in predictions with a stronger association with known hypertension genetics than the individual phenotypes and found that it did. This result was specific to the relevant tissues—when performing the same analysis on all tissue-specific networks, we found that the kidney, heart and liver networks showed stronger performance than the tissue-naive network (**Supplementary Fig. 7a**).

In addition to well-documented hypertension-associated genes such as *MTHFR* and *PPARG*, NetWAS identified several additional candidates (**Supplementary Table 4**). Many lines of evidence in the literature link the top predicted genes to hypertension via mechanistic relationships to known disease genes and pathways or associations with hypertension risk factors. Several such examples are tabulated in **Supplementary Table 5**. Using functional annotations from Gene Ontology (GO) to subsequently interpret NetWAS-prioritized genes, we observed that NetWAS-ranked genes were annotated to the GO term 'regulation of blood pressure' at a significantly higher frequency than the top-ranked GWAS genes (**Fig. 5b**). Because NetWAS provides a useful reprioritization of the genome in terms of phenotypic and functional association, we explored whether the re-ranking was also helpful in discovering appropriate therapeutics. Using drug target

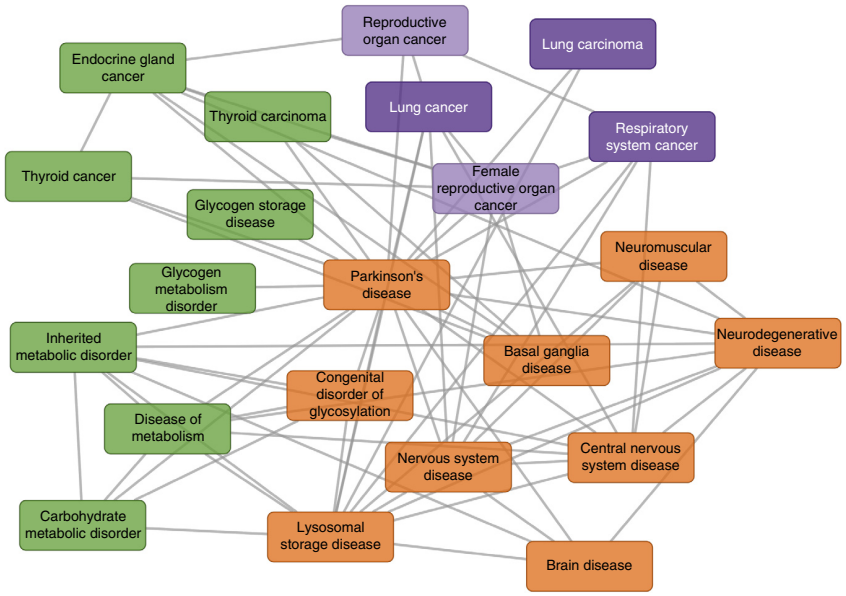


Figure 4 A disease map centered on Parkinson's disease summarizing its molecular associations with other diseases in substantia nigra. The disease map effectively identifies the connection of Parkinson's disease to both documented nervous system diseases and several cancers, through the PARK gene family. Parkinson's disease is characterized by the death of dopaminergic neurons in substantia nigra. Associations between the genes associated with Parkinson's disease and other diseases were tested by calculating the connectivity across the disease gene sets relative to background connectivity in the substantia nigra network. All significant (z score > 2.5) connections (edges) between diseases (nodes) are shown in this disease map.

data from DrugBank[56], we found that targets of antihypertensive drugs were significantly enriched among the top genes from NetWAS more than among the top genes from GWAS (**Fig. 5c**). We found similar results for drug targets from three other databases (PharmGKB[57], TTD[58] and CTD[59]; **Supplementary Fig. 7b**).

We evaluated NetWAS on four additional GWAS spanning diverse disease and tissue contexts and found that the approach consistently ranked documented disease-associated genes higher than the GWAS (**Supplementary Fig. 8**). Thus, NetWAS builds from nominally significant associations from GWAS to identify candidates by their connectivity in tissue-specific networks that are valuable for guiding research into disease mechanisms and therapy.

A dynamic, interactive interface for biomedical researchers

To facilitate broad use of these networks by biomedical researchers, we have developed GIANT—a dynamic, interactive web interface. Researchers can query by individual genes or by gene sets of interest to analyze tissue-specific gene functions and interactions. For example, GIANT can provide tissue-specific functional maps and predictions of tissue-specific gene function and disease association. Multi-tissue view allows for rapid examination of the tissue-specific rewiring of functional connections across diverse tissues (**Fig. 3a**). Custom gene set functionality is implemented using the Tribe web service and is integrated into user analyses, such as biological process enrichment and querying by gene set. GIANT also provides a full NetWAS implementation, allowing users to upload gene-based association P values to receive NetWAS association scores. Visualizations in the user-friendly, dynamic web interface are implemented using the D^3 library[60], which enables use on any modern web browser without plugin installation. In addition to the interface, all of the underlying networks are provided for download, and the full list of input data sets and their sources is available through the webserver.

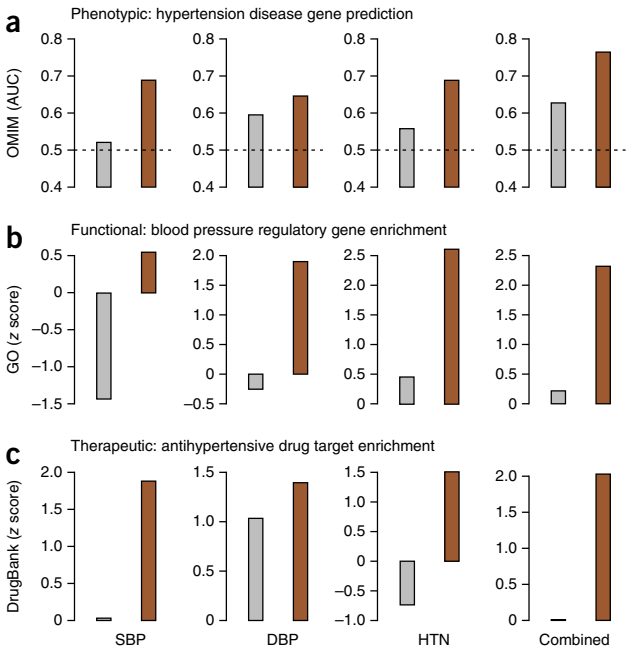


Figure 5 Network reprioritization of hypertension GWAS identifies hypertension-associated genes. Genes ranked using GWAS (gray) and genes reprioritized using NetWAS (brown) were assessed for correspondence to genes known to be associated with hypertension phenotypes, regulatory processes and therapeutics. We compared individual (systolic blood pressure, SBP; diastolic blood pressure, DBP; hypertension, HTN) as well as combined hypertension endpoints. (**a**) Gene rankings were compared to OMIM-annotated hypertension genes using AUC. The AUC for the tissue-specific NetWAS is consistently higher than that for the original GWAS for all hypertension endpoints. Merging the network-based predictions for the three hypertension-related endpoints into a combined phenotype results in the best performance (AUC = 0.77; original GWAS AUC = 0.62; the dashed line at 0.5 denotes the AUC of a baseline random predictor). (**b**,**c**) Gene rankings were also assessed for enrichment of genes involved in the regulation of blood pressure (GO) (**b**) and targets of antihypertensive drugs (DrugBank) (**c**). The top NetWAS results were significantly enriched for genes involved in blood pressure regulation as well as for genes that are targets of antihypertensive drugs. Enrichment was calculated as a z score (Online Methods), with higher scores indicating a greater shift from the expected ranking toward the top of the list. In nearly all cases, the NetWAS ranking was both significantly enriched with the respective gene sets (z score > 1.645 ≈ P value < 0.05) and more enriched than in the original GWAS ranking.

DISCUSSION

Genes with tissue-specific expression and function have key roles in the physiological processes of complex organisms, and such genes are expected to underlie many human diseases[2,3]. Recent advances

now allow for high-throughput discovery of genes expressed in specific lineages in solid tissues[9,61]. The next challenge is to understand the tissue-specific function of genes. This remains difficult because the precise functions of genes in multicellular organisms such as humans are defined by the context in the cell lineage where the genes are expressed. Tissue-specific interactions are not well characterized because high-throughput interaction measurements are largely infeasible in solid tissues and their cell lineages. For direct studies of human genes, the tools available to assess tissue-specific function are generally confined to cell lines, many of which have diverged phenotypically from normal tissues. Moreover, many low-throughput experiments are highly skewed toward well-studied genes[7,62].

We developed a data-driven approach that identifies tissue-specific interactions by integrating heterogeneous publicly available data using a tissue-specific regularized Bayesian framework. Our learned networks complement the tools of modern molecular genetics by more precisely predicting tissue-specific relationships to generate hypotheses about tissue-specific gene action. These lineage-specific networks also effectively connect the roles of genes in cell lineages to common diseases. We leverage this power in NetWAS, which uses GWAS as the starting point for network analysis and provides a way to increase the value of existing GWAS. Other methods[63] that reprioritize GWAS findings using networks are also expected to benefit substantially from tissue-specific networks. Analysis of genetic association data presents a key opportunity to apply tissue-specific networks to understand common human diseases. Because these networks accurately weigh and integrate diverse molecular data, they provide a more complete picture of the relationships between genes, phenotypes and tissues and a clearer understanding of the etiology of complex disease. This is particularly important in the domain of phenome-wide association studies that rely on endpoints gleaned from electronic health records (EHRs)[64]. Tissue-specific networks can provide the necessary gene and tissue context to analyze such data and will help in scaling methods to the repositories that we expect to see in the coming era of widespread EHR and genetic data.

Healthy and disease states in humans are the result of the interplay of genes within specific cell lineages and tissues, modulated by environmental exposures. Many of the key challenges in medicine involve tissue specificity. For example, identifying off-target effects for therapeutics requires an understanding of the therapeutics' effect not just in the target tissue but also in all tissues. By disentangling the functions of genes in specific tissues, integrated tissue-specific networks learned from large data compendia present a means to address these challenges.

URLs. GIANT, a web portal for tissue-specific functional networks, http://giant.princeton.edu/; Sleipnir, an open source library for functional genomics, http://libsleipnir.bitbucket.org/; Tribe, a web service that provides cross-server analysis of gene sets, http://tribe.greenelab.com/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene expression measurements of HASMCs with and without IL-1β stimulation are available in the Gene Expression Omnibus (GEO) database under accession GSE59671.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The first three authors are co-first authors and are listed alphabetically.

We sincerely thank Y. Lee and D. Gorenshteyn for help in curating disease associations and L. Bongo and M. Homilius for help in processing expression data. We are grateful to all members of the Troyanskaya laboratory for help in curating specific GO biological processes and for valuable discussions.

This work was primarily supported by US National Institutes of Health (NIH) grants R01 GM071966 and R01 HG005998 to O.G.T. and U54 HL117798 to G.A.F. C.S.G. was supported in part by US NIH grants T32 CA009528 and P20 GM103534. A.K.W. was supported in part by US NIH grant T32 HG003284. This work was supported in part by US NIH grant P50 GM071508 and by US NIH contract HHSN272201000054C. O.G.T. is a senior fellow of the Genetic Networks program of the Canadian Institute for Advanced Research (CIFAR).

AUTHOR CONTRIBUTIONS

C.S.G., A.K., A.K.W. and O.G.T. conceived and designed the research. C.S.G., A.K. and A.K.W. performed computational analyses with contributions from D.S.H. and R.Z., and E.R. performed the molecular experiments. A.K.W., R.A.Z. and C.S.G. developed the web interface. D.I.C., B.M.H., E.Z., S.C.S. and K.D. provided data. C.S.G., A.K., A.K.W. and O.G.T. wrote the manuscript with input from E.R., T.G., G.A.F. and K.D. and revisions from all co-authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. D'Agati, V.D. The spectrum of focal segmental glomerulosclerosis: new insights. *Curr. Opin. Nephrol. Hypertens.* **17**, 271–281 (2008).
2. Cai, J.J. & Petrov, D.A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol. Evol.* **2**, 393–409 (2010).
3. Winter, E.E., Goodstadt, L. & Ponting, C.P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* **14**, 54–61 (2004).
4. 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. USA* **105**, 20870–20875 (2008).
5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
6. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
7. Pandey, A.K., Lu, L., Wang, X., Homayouni, R. & Williams, R.W. Functionally enigmatic genes: a case study of the brain ignorome. *PLoS ONE* **9**, e88889 (2014).
8. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
9. Ju, W. *et al.* Defining cell-type specificity at the transcriptional level in human disease. *Genome Res.* **23**, 1862–1873 (2013).
10. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. USA* **100**, 8348–8353 (2003).
11. Myers, C.L. & Troyanskaya, O.G. Context-sensitive data integration and prediction of biological networks. *Bioinformatics* **23**, 2322–2330 (2007).
12. Hibbs, M.A. *et al.* Directing experimental biology: a case study in mitochondrial biogenesis. *PLoS Comput. Biol.* **5**, e1000322 (2009).
13. Park, C.Y. *et al.* Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput. Biol.* **9**, e1002957 (2013).
14. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
15. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
16. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9** (suppl. 1), S4 (2008).
17. Hwang, S., Rhee, S.Y., Marcotte, E.M. & Lee, I. Systematic prediction of gene function in *Arabidopsis thaliana* using a probabilistic functional gene network. *Nat. Protoc.* **6**, 1429–1442 (2011).
18. Kofler, S., Nickel, T. & Weis, M. Role of cytokines in cardiovascular diseases: a focus on endothelial responses to inflammation. *Clin. Sci.* **108**, 205–213 (2005).
19. Liu, J.Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
20. Keshava Prasad, T.S. *et al.* Human Protein Reference Database—2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
21. Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39**, D507–D513 (2011).
22. Britten, R.J. & Davidson, E.H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).

23. Spitz, F. & Furlong, E.E.M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
24. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462**, 587–594 (2009).
25. Stadtfeld, M. & Hochedlinger, K. Induced pluripotency: history, mechanisms, and applications. *Genes Dev.* **24**, 2239–2263 (2010).
26. Goh, K.-I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690 (2007).
27. Brunner, H.G. & van Driel, M.A. From syndrome families to functional genomics. *Nat. Rev. Genet.* **5**, 545–551 (2004).
28. Arce, L., Yokoyama, N.N. & Waterman, M.L. Diversity of LEF/TCF action in development and disease. *Oncogene* **25**, 7492–7504 (2006).
29. van Amerongen, R. & Nusse, R. Towards an integrated view of Wnt signaling in development. *Development* **136**, 3205–3214 (2009).
30. Reya, T. *et al.* Wnt signaling regulates B lymphocyte proliferation through a LEF-1 dependent mechanism. *Immunity* **13**, 15–24 (2000).
31. Park, S.-K., Son, Y. & Kang, C.-J. A strong promoter activity of pre–B cell stage-specific *Crlz1* gene is caused by one distal LEF-1 and multiple proximal Ets sites. *Mol. Cells* **32**, 67–76 (2011).
32. Gutierrez, A. *et al.* LEF-1 is a prosurvival factor in chronic lymphocytic leukemia and is expressed in the preleukemic state of monoclonal B-cell lymphocytosis. *Blood* **116**, 2975–2983 (2010).
33. Erdfelder, F., Hertweck, M., Filipovich, A., Uhrmacher, S. & Kreuzer, K.-A. High lymphoid enhancer-binding factor-1 expression is associated with disease progression and poor prognosis in chronic lymphocytic leukemia. *Hematol. Rep.* **2**, e3 (2010).
34. Gandhirajan, R.K. *et al.* Small molecule inhibitors of Wnt/β-catenin/lef-1 signaling induces apoptosis in chronic lymphocytic leukemia cells *in vitro* and *in vivo*. *Neoplasia* **12**, 326–335 (2010).
35. Lee, J.E., Wu, S.-F., Goering, L.M. & Dorsky, R.I. Canonical Wnt signaling through Lef1 is required for hypothalamic neurogenesis. *Development* **133**, 4451–4461 (2006).
36. Wang, X., Lee, J.E. & Dorsky, R.I. Identification of Wnt-responsive cells in the zebrafish hypothalamus. *Zebrafish* **6**, 49–58 (2009).
37. Kahler, R.A. *et al.* Lymphocyte enhancer-binding factor 1 (Lef1) inhibits terminal differentiation of osteoblasts. *J. Cell. Biochem.* **97**, 969–983 (2006).
38. Hoeppner, L.H. *et al.* Runx2 and bone morphogenic protein 2 regulate the expression of an alternative *Lef1* transcript during osteoblast maturation. *J. Cell. Physiol.* **221**, 480–489 (2009).
39. Noh, T. *et al. Lef1* haploinsufficient mice display a low turnover and low bone mass phenotype in a gender- and age-specific manner. *PLoS ONE* **4**, e5438 (2009).
40. Westendorf, J.J., Kahler, R.A. & Schroeder, T.M. Wnt signaling in osteoblasts and bone diseases. *Gene* **341**, 19–39 (2004).
41. Cohen, M.M. Biology of RUNX2 and cleidocranial dysplasia. *J. Craniofac. Surg.* **24**, 130–133 (2013).
42. Duan, D. *et al.* Submucosal gland development in the airway is controlled by lymphoid enhancer binding factor 1 (LEF1). *Development* **126**, 4441–4453 (1999).
43. Driskell, R.R. *et al.* Wnt-responsive element controls Lef-1 promoter expression during submucosal gland morphogenesis. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **287**, L752–L763 (2004).
44. Driskell, R.R. *et al.* Wnt3a regulates Lef-1 expression during airway submucosal gland morphogenesis. *Dev. Biol.* **305**, 90–102 (2007).
45. Verkman, A.S., Song, Y. & Thiagarajah, J.R. Role of airway surface liquid and submucosal glands in cystic fibrosis lung disease. *Am. J. Physiol. Cell Physiol.* **284**, C2–C15 (2003).
46. Forno, L.S. Neuropathology of Parkinson's disease. *J. Neuropathol. Exp. Neurol.* **55**, 259–272 (1996).
47. Veeriah, S. *et al.* Somatic mutations of the Parkinson's disease–associated gene *PARK2* in glioblastoma and other human malignancies. *Nat. Genet.* **42**, 77–82 (2010).
48. Denison, S.R. *et al.* Alterations in the common fragile site gene *Parkin* in ovarian and other cancers. *Oncogene* **22**, 8370–8378 (2003).
49. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
50. O'Seaghdha, C.M. & Fox, C.S. Genome-wide association studies of chronic kidney disease: what have we learned? *Nat. Rev. Nephrol.* **8**, 89–99 (2012).
51. Ridker, P.M. *et al.* Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000 initially healthy American women. *Clin. Chem.* **54**, 249–255 (2008).
52. Ho, J.E. *et al.* Discovery and replication of novel blood pressure genetic loci in the Women's Genome Health Study. *J. Hypertens.* **29**, 62–69 (2011).
53. Oldham, P.D., Pickering, G., Roberts, J.A. & Sowry, G.S. The nature of essential hypertension. *Lancet* **1**, 1085–1093 (1960).
54. Guyton, A.C. Blood pressure control—special role of the kidneys and body fluids. *Science* **252**, 1813–1816 (1991).
55. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
56. Wishart, D.S. *et al.* DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).
57. Thorn, C.F., Klein, T.E. & Altman, R.B. PharmGKB: the Pharmacogenomics Knowledge Base. *Methods Mol. Biol.* **1015**, 311–320 (2013).
58. Qin, C. *et al.* Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.* **42**, D1118–D1123 (2014).
59. Davis, A.P. *et al.* The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.* **41**, D1104–D1114 (2013).
60. Bostock, M., Ogievetsky, V. & Heer, J. D3: Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **17**, 2301–2309 (2011).
61. Forrest, A.R.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
62. Hoffmann, R. & Valencia, A. Life cycles of successful genes. *Trends Genet.* **19**, 79–81 (2003).
63. Köhler, S., Bauer, S., Horn, D. & Robinson, P.N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82**, 949–958 (2008).
64. Denny, J.C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

ONLINE METHODS

Data download and processing. We collected and integrated 987 genome-scale data sets encompassing approximately 38,000 conditions from an estimated 14,000 publications including both expression and interaction measurements. We downloaded interaction data from BioGRID[65], IntAct[66], MINT[67] and MIPS[68]. BioGRID edges were discretized into five bins, labeled 0 to 4, where the bin number reflected the number of experiments supporting the interaction. For the remaining databases, edges were discretized into the presence or absence of an interaction.

Predicting transcriptional regulation on the basis of DNA sequence is a major challenge in understanding transcription at a systems level. To estimate shared transcription factor regulation, binding motifs were downloaded from JASPAR[69]. Genes were scored for the presence of transcription factor binding sites using the MEME software suite[70]. FIMO[71] was used to scan for each transcription factor profile within the 1-kb sequence upstream of each gene[72]. Motif matches were treated as binary scores (present if $P < 0.001$). The final score for each gene pair was obtained by calculating the Pearson correlation between the motif association vectors for the genes.

Chemical and genetic perturbation (c2:CGP) and microRNA target (c3:MIR) profiles were downloaded from the Molecular Signatures Database (MSigDB[73]). Each gene pair's score was the sum of shared profiles weighted by the specificity of each profile (1/len(genes)). The resulting scores were converted to *z* scores and discretized into bins ((−infinity, −1.5), [−1.5, −0.5), [−0.5, 0.5), [0.5, 1.5), [1.5, 2.5), [2.5, 3.5), [3.5, 4.5), [4.5, infinity)).

We downloaded all gene expression data sets from NCBI's Gene Expression Omnibus[74] (GEO) and collapsed duplicate samples. GEO contains 980 human data sets representing 20,868 conditions. Genes with more than 30% of values missing were removed, and remaining missing values were imputed using ten neighbors[75]. Non-log-transformed data sets were log transformed. Expression measurements were summarized to Entrez[76] identifiers, and duplicate identifiers were merged. The Pearson correlation was calculated for each gene pair, normalized with Fisher's z transform, mean subtracted and divided by the standard deviation. The resulting z scores were discretized into bins ((−infinity, −1.5), [−1.5, −0.5), [−0.5, 0.5), [0.5, 1.5), [1.5, 2.5), [2.5, 3.5), [3.5, infinity)).

Hierarchically aware knowledgebase construction via ontological pruning with functional knowledge transfer.

Functional knowledge extraction. We constructed a tissue-naive functional relationship gold standard from a set of 564 expert-selected GO biological process terms and experimentally derived gene annotations (GO evidence codes: EXP, IDA, IPI, IMP, IGI and IEP). Curators identified processes testable through specific molecular experiments (**Supplementary Table 6**). Pairs of genes that were co-annotated to expert-selected terms after propagation were treated as positive (i.e., functionally related) examples. Gene pairs not co-annotated to any of these terms were considered as negative examples, except in the following cases: (i) if two genes were annotated to two different GO terms with a significant number of shared genes (hypergeometric P value < 0.05) and (ii) if two genes were co-annotated to a set of 'negative' GO terms that defined minimal relatedness[77]. Gene pairs that met either condition were excluded from the set of negative examples and treated as neither related nor unrelated.

Functional knowledge transfer. To increase the coverage of functional interactions, we transferred experimentally confirmed mouse GO annotations to human functional analogs identified by FKT[13], a high-specificity annotation transfer method, for the 520 GO terms with mouse annotations. This resulted in a tissue-naive gold standard of 604,038 functionally related gene pairs (positive examples) and 12,425,713 potentially unrelated pairs (negative examples).

Ontology pruning. Gene-to-tissue annotations were obtained from the Human Protein Reference Database (HPRD)[20]. HPRD tissues were mapped to the BRENDA Tissue Ontology[21] (BTO) using direct matching where possible and manual curation where direct matches were unavailable (**Supplementary Table 7**). Tissues with fewer than ten directly annotated genes were pruned as non-informative from a molecular standpoint (for example, BTO:0001493, trunk). Pruning resulted in an ontology containing functional, as opposed to structural, divisions of tissues and cell lineages (**Supplementary Table 8**). We defined 'tissue categories' from generic BRENDA terms, for example, nervous system, to categorize tissues into organ systems for evaluation and analysis. For each tissue, we termed the set T as those genes directly annotated to that tissue or any of its descendants in the ontology. We used tissue categories to define unrelated tissues (those not associated with the same category as the tissue of interest). We defined T′ for each tissue as genes specifically annotated to unrelated tissues.

Annotation of ubiquitously expressed genes. Genes ubiquitously expressed across tissues frequently carry out core biological processes and interact with tissue-specific genes to perform specialized functions[78]. We identified ubiquitous genes from a multi-tissue RNA sequencing experiment[79] and added 'widely expressed' genes from a multi–cell line mass spectroscopy experiment[80], genes for proteins expressed in >75% of the tissues assayed in the human protein atlas[81] and curated 'ubiquitous genes' from HPRD[20]. These 8,475 ubiquitous genes (U) were considered expressed in all tissues and cell types, in addition to the curated tissue-specific genes (T). Sets T and U were made disjoint by retaining only genes in T genes that were not in U.

Integration of tissue-specific and functional knowledge. We combined the curated gene-to-tissue annotations with the tissue-naive functional gold standard to construct a hierarchical tissue-specific knowledgebase. We labeled each gene pair (positive or negative) in the functional relationship standard as specifically coexpressed in a tissue if both genes were tissue specific (T, T) or one was tissue specific and the other was ubiquitous (T, U). Interactions between ubiquitous gene pairs were deemed not tissue specific and were ignored. After labeling specifically coexpressed gene pairs (edges) across all tissues, we considered four classes of edges—C1, C2, C3 and C4—to constitute each tissue standard.

- C1: positive functional edges between genes specifically coexpressed in the tissue [T–T and T–U].
- C2: positive functional edges between a gene expressed in the tissue and another specifically expressed in an unrelated tissue [T–T′ and U–T′].
- C3: negative functional edges between genes specifically coexpressed in the tissue [T–T and T–U].
- C4: negative functional edges between one gene expressed in the tissue and another specifically expressed in an unrelated tissue [T–T′ and U–T′].

Among the four tissue classes, C1 represented tissue-specific functional relationships. To identify tissue-specific relationships, we constructed a specific gold standard for each tissue by labeling edges in C1 as positives and edges in the other classes as negatives. Because C3 is defined on the basis of tissue-expressed genes and C2 and C4 are defined on the basis of non-expressed genes, the number of edges in these classes varied across tissues according to how specific (cell type, tissue, organ or system), well studied (or easily studied) and well curated (literature bias) they are. To construct comparable networks across tissues, we used a negative set composed of equal proportions of edges from C2, C3 and C4. We limited all integrations to the set of 144 tissues (**Supplementary Table 8**) that contained at least ten C1 edges between tissue-specific genes (T–T). This method incorporates the hierarchical relationships of tissues, allowing supervised methods to leverage these relationships.

Data integration. We constructed functional networks from genome-scale data by performing a tissue-specific Bayesian integration. We trained one naive Bayesian classifier for each tissue using the tissue-specific standards described above. We also trained a classifier limited to only functional interactions to generate a tissue-naive network. In each case, we constructed a class node, i.e., the presence or absence of a functional relationship between a pair of genes that is conditioned on nodes for each data set. For large-scale genomics data sets, the assumption of conditional independence required for a naive Bayes classifier is often not met, so we calculated and corrected for non-biological conditional dependency[13].

Each tissue model trained on the hierarchy-aware tissue-specific knowledge was used to make genome-wide predictions by estimating the probability of tissue-specific functional interaction between all pairs of genes. We also estimated the probability of global functional interactions for the tissue-naive

network. We assigned a prior probability of a functional relationship of 0.01 for all models, allowing edge probabilities to be compared across tissues.

Code availability. Integrations were performed with C++ naive Bayesian learning implementations from the open source Sleipnir library for functional genomics[82].

Evaluation of tissue-specific functional relationships. We evaluated tissue-naive and tissue-specific functional networks using fivefold cross-validation. The 6,062 genes represented in the tissue-specific knowledge-base were randomly partitioned into 5 sets. For each cross-validation run, gene pairs where neither gene was present in the holdout interval were used for training. Any gene pair where both genes were present in the holdout was used for evaluation of the AUC. The estimated performance of each of the 144 functional networks was summarized as the median AUC of the five cross-validation runs (**Supplementary Table 8**).

Mapping data sets to tissues. We mapped data sets to tissues to compare with an integration of only tissue-specific data. On the basis of previous work[83] that annotated samples from biological text, we extracted the title and description for each GDS data set and annotated each using MetaMap[84]. This resulted in a mapping of GDS data sets to Unified Medical Language System (UMLS) terms. We applied the same process for the title and description of each BRENDA tissue and merged the two mappings by shared UMLS terms.

Network-based prediction. Top genes functionally connected to *IL1B* in a network (tissue specific or naive) were identified by ranking all genes on the basis of the edge weight to *IL1B* normalized by their connection to all the genes. More precisely, in a network with V genes and interaction probabilities p_{uv}, the specific connection (s_v) of each gene $v \in V$ in the network to a query gene u (*IL1B* in this case) is

$$s_v = \frac{p_{uv}}{d_v};\ d_v = \frac{\sum_{t \in V} p_{ut}}{|v|}$$

This measure identified genes specifically connected to *IL1B*, which were compared to genes identified from the validation experiment described below.

Cell culture. HASMCs (Cambrex) were maintained in smooth muscle cell growth medium with the manufacturer's additives (SM-GM, Cambrex) and 10% FCS in 5% CO_2 at 37 °C. Cells were expanded to subconfluent cultures and split onto 100-mm culture dishes, where they were grown to confluence. Subsequently, cells were rendered quiescent, by incubation for 24 h in serum-free medium, before stimulation with 10 ng/ml IL-1β (Sigma) for 2 h ($n = 4$). The cells tested negative for mycoplasma, bacteria, yeast and fungi. All donors tested negative for HIV-1, hepatitis B and hepatitis C.

Gene expression analysis. Total RNA was isolated from HASMCs using the Qiagen RNeasy Mini kit. Samples were prepared in one batch using the Nugen sample preparation protocol and hybridized to Affymetrix HG U-133A v2. CEL files were background corrected, normalized and summarized using RMA[85] on the basis of a custom CDF[86]. Differential expression analysis was carried out using LIMMA[87], and genes induced at 2 h post-stimulation compared to at 0 h were identified by ranking genes by their reported t statistic. These data have been submitted to the GEO database (accession GSE59671).

Evaluation in publicly available data. In addition to our validation experiment (GSE59671), we curated all series in GEO that included treatment of cells with IL-1β and controls. This resulted in nine data sets: GSE13168 (airway smooth muscle), GSE26315 (amnion mesenchymal cells), GSE31679 (trophoblast cells), GSE40007 (endometrial stromal cells), GSE49604 (osteoarthritis), GSE7216 (keratinocytes), GSE37624 (umbilical vein endothelial cells), GSE40560 (fibroblasts) and GSE40838 (peripheral blood mononuclear cells). Of these data sets, only GSE7216 was included in the data compendium used for integration. The rest were independent of the integration. To assess our networks' ability to identify gene sets that would respond to IL-1β treatment, we contrasted IL-1β treatments with controls using GEO2R[74]. We queried the GIANT webserver for neighbors of *IL1B* in the tissue network that best corresponded to each data set. In each tissue, genes were ranked on the basis of the connectivity measure described above. We evaluated the mean fold change of the top 20 returned results. We evaluated randomly selected matched size sets of genes from each data set as controls.

Evaluation of tissue-specific processes, gene-level rewiring and disease-disease association. *Mapping GO biological processes to tissues.* To evaluate tissue-specific functional rewiring in our networks, we needed associations between tissues and tissue-specific processes. We used text matching followed by manual curation to map biological process (BP) terms in GO to tissue terms in the BRENDA Tissue ontology (**Supplementary Table 9**).

Network connectivity of tissue-specific processes. For each tissue, we constructed a tissue-minus-naive network by subtracting edge probabilities of the naive network from those of the tissue network. Negative weights were set to zero. In this subtracted network, positive scores corresponded to edges with a tissue network interaction probability greater than the naive network probability. We expected relevant tissue-specific processes to be more connected in the tissue network than the naive network and over processes that are not. For instance, for T lymphocytes, 'T cell receptor signaling pathway' is a relevant process, whereas 'neuron projection development' is not. Within each subtracted tissue network, we ranked all tissue-specific processes by their edge density in the network and evaluated the extent to which relevant processes (positives) were ranked above processes specific to other tissues (negatives). Edge density for each process (with n genes) was calculated as the sum of weights divided by the total number of possible ($n \times (n - 1)/2$) edges between genes in that process. We measured the performance of the ranking using AUC and calculated a 'best AUC' as the relative rank of the densest relevant process.

Gene-level rewiring across tissue networks. Because tissue specificity emanates from specialization of gene function, we identified genes with distinct functional neighborhoods in different tissue networks. We curated genes annotated to tissue-specific processes associated with at least two widely different tissues (descendants of different tissue categories). Using this gene process–tissue mapping, we identified gene-tissue pairs, each with a set of relevant tissue-specific processes labeled positive and other processes annotated to the gene labeled negative. For example, the gene *LEF1* is annotated to both the 'blood vessel' and 'osteoblast' tissues. In 'blood vessel', the term 'angiogenesis' would be a positive and 'osteoblast differentiation' would be a negative. Then, for a given gene-tissue pair (for example, *LEF1* and blood vessel), we calculated a z score for the connectivity of a process (for example, angiogenesis) using the formulation

$$z_{\text{process}} = \frac{m_{\text{process}} - \mu}{\left(\sigma / \sqrt{n}\right)}$$

where μ and σ are the mean and standard deviation of the interaction probabilities of all genes to the query gene; n is the number of genes annotated to the process; and m is the mean of the interaction probabilities of the process genes to the query gene.

We ranked all processes by decreasing z score and quantified the separation between positively and negatively labeled processes using AUC. A high AUC for a gene across multiple associated networks showed that tissue networks reflected the gene's annotation to multiple tissue-specific processes through preferential connectivity to the appropriate tissue-specific process in the matched tissue.

Disease-association map. We constructed a disease association map, which represents a high-level view of functionally related diseases. As in Huttenhower *et al.*[8], we calculated an association score between each disease pair using functional interactions between two diseases' constituent genes. The score compared the means of two edge distributions: the edges between disease gene sets (between) and the edges that were incident to the disease gene sets across the genome (background). We calculated a t statistic as follows for disease pair i and j:

$$t_{i,j} = \frac{X_w - X_b}{s_x}$$

$$s_x = \sqrt{\frac{s_w^2}{n_w} + \frac{s_b^2}{n_b}}$$

doi:10.1038/ng.3259

where X_w is the mean weight of edges between the two disease gene sets, X_b is the mean weight of all genome-wide edges incident to either gene set, and s and n are the respective standard deviation and sizes of the distributions.

We generated a bootstrapped null distribution for each disease pair by sampling 10,000 random gene set pairs of the same size and recalculated the above t statistic. With this null distribution, we calculated the final disease association score for each disease pair as follows:

$$z_{i,j} = \frac{t_{i,j} - \mu}{s}$$

where $t_{i,j}$ is the calculated t statistic for a disease pair, and μ and s are the mean and standard deviation of the null distribution. We applied a z-score cutoff of 2.5 to produce the Parkinson's disease map in **Figure 4**.

Blinded literature evaluation of the disease association map. To rigorously evaluate these maps, we constructed and shuffled a list of putative associations for Parkinson's disease that combined associations from the disease map with ten randomly selected control associations. We provided this list to a researcher with no previous exposure to our manuscript or results. This researcher categorized disease associations from the literature as 'strong', indicating there was clear evidence, 'weak', indicating that there existed co-mentions but that the available evidence was limited, or 'none', indicating that there were no publications with co-mentions.

Network-based reprioritization of genome-wide association study. We used tissue-specific networks to reprioritize gene candidates associated with hypertension endpoints in a GWAS. We hypothesized that disease-relevant genes would be enriched among the nominally significant genes, which would allow reprioritization through modern machine learning methods. We trained a support vector machine classifier using nominally significant ($P < 0.01$) genes as positive examples and 10,000 randomly selected non-significant ($P \geq 0.01$) genes as negatives. The classifier was constructed using the tissue network specific to kidney, a tissue associated with hypertension[54], where the features of the classifier were the edge weights of the labeled examples to all the genes in the network. Genes were re-ranked using their distance from the hyperplane, which represented a network-based prioritization of a GWAS, termed NetWAS.

We applied NetWAS to a GWAS from the Women's Genome Health Study to identify additional genes involved in hypertension[51]. The study focused on three hypertension-related endpoints: systolic blood pressure, diastolic blood pressure and hypertension diagnosis. To calculate per-gene P values for each endpoint, we used the versatile gene-based association study (VEGAS) system[19]. To generate a combined list across phenotypes, we combined results from each hypertension-related endpoint using summed ranks. Performance was assessed by evaluating the ranking of genes annotated to 'hypertension' in OMIM. We performed functional evaluation by comparing NetWAS results to genes annotated to the term 'regulation of blood pressure' in GO. We performed an analogous calculation for therapeutics with targets of antihypertensive drugs from four different databases: DrugBank, TTD, PharmGKB and CTD.

Evaluation of additional GWAS data. We performed NetWAS on four additional GWAS: C-reactive protein levels (lnCRP)[51], type 2 diabetes (T2D)[88], body mass index (BMI)[89] and advanced age-related macular degeneration (advanced AMD)[90]. Publicly available studies were obtained from their respective websites (BMI, advanced AMD) or the database of Genotypes and Phenotypes (dbGaP)[91] (T2D, phs000007–pha000418). NetWAS was applied as described for the hypertension NetWAS analysis. The relevant OMIM diseases were used to evaluate NetWAS results in relevant tissues.

65. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* **41**, D816–D823 (2012).
66. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40**, D841–D846 (2012).
67. Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.* **40**, D857–D861 (2012).
68. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **27**, 44–48 (1999).
69. Portales-Casamar, E. *et al.* JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38**, D105–D110 (2010).
70. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
71. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
72. Huber, B.R. & Bulyk, M.L. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics* **7**, 229 (2006).
73. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
74. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
75. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
76. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39**, D52–D57 (2011).
77. Myers, C.L., Barrett, D.R., Hibbs, M.A., Huttenhower, C. & Troyanskaya, O.G. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7**, 187 (2006).
78. Bossi, A. & Lehner, B. Tissue specificity and the human protein interaction network. *Mol. Syst. Biol.* **5**, 260 (2009).
79. Ramsköld, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, e1000598 (2009).
80. Burkard, T.R. *et al.* Initial characterization of the human central proteome. *BMC Syst. Biol.* **5**, 17 (2011).
81. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).
82. Huttenhower, C., Schroeder, M., Chikina, M.D. & Troyanskaya, O.G. The Sleipnir library for computational functional genomics. *Bioinformatics* **24**, 1559–1561 (2008).
83. Schmid, P.R., Palmer, N.P., Kohane, I.S. & Berger, B. Making sense out of massive data by going beyond differential expression. *Proc. Natl. Acad. Sci. USA* **109**, 5594–5599 (2012).
84. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* **2001**, 17–21 (2001).
85. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
86. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).
87. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
88. Meigs, J.B. *et al.* Genome-wide association with diabetes-related traits in the Framingham Heart Study. *BMC Med. Genet.* **8** (suppl. 1), S16 (2007).
89. Randall, J.C. *et al.* Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet.* **9**, e1003500 (2013).
90. Fritsche, L.G. *et al.* Seven new loci associated with age-related macular degeneration. *Nat. Genet.* **45**, 433–439 (2013).
91. Mailman, M.D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).

Molecular networks in context

Andrew M Gross & Trey Ideker

Network biology is beginning to tackle the complexities of multicellular systems and disease associations.

As a young science, network biology still lacks the tools to capture the rich diversity of complex biological phenomena such as multicellular organization and disease processes. But recent developments in the field point to the emergence of more-sophisticated approaches to network reconstruction. In a new study in *Nature Genetics*, Greene *et al.*[1] build a library of tissue-specific networks from diverse data sets—including tissue-specific gene-expression profiles, general and tissue-specific protein interactions, and maps of tissue development—and apply the library to predict how gene function changes across tissues. The tissue-specific networks provided in the study are the most comprehensive to date. They will aid efforts to model interactions of different cell types in tissues and whole organisms and to understand how these interactions are altered in disease.

The past two decades have seen an explosion in high-throughput 'omics' data sets on different biological systems. These data have been integrated and assembled into large gene networks through various computational methods that identify interactions among genes and correlations among gene profiles[2]. Many online resources for such analyses are now available, including FunctionalNet, STRING, GeneMania and bioPIXIE, and these sites make available large networks of gene interactions for humans and model species. However, the networks are typically presented as generic models without contextual information on the particular tissue, cell type, disease state, developmental stage or time point of the system or its external stresses and stimuli—all of which may have a strong influence on gene and protein interactions[3].

A handful of studies have attempted to address this limitation by constructing tissue-specific networks. Magger *et al.*[4] generated networks for 60 tissues and applied the networks to prioritize disease-gene associations. Cahan *et al.*[5] assembled expression data from 56 papers into networks for 22 tissues, which were used to guide cellular engineering[6]. The scale of the work of Greene *et al.*[1] far exceeds that of these earlier studies. Combining data from more than 14,000 publications, they construct networks for an unprecedented 144 human tissues and cell types (**Fig. 1**). This large number of networks allows gene function to be inferred even in tissues for which little experimental data exists, because biological information from functionally related tissues can be used to enhance the signal.

Greene *et al.*[1] use their networks to predict how certain cell lineages respond to perturbations and to investigate changes in gene function across different tissues. For example, they identify interactions that are specifically active in blood vessels in response to interleukin-1B. This result reflects that protein neighbors of interleukin-1B in the blood-vessel network are more likely to be perturbed than in other tissue types or in a tissue-naive network. In another example, the authors explore the ability of the LEF1 transcription factor to mediate signal transduction through interactions with various different factors, depending on the tissue. These cases show the potential of tissue-dependent networks to reveal diverse context-dependent functions for a protein.

In a very different application, Greene *et al.*[1] reinterpret the data from a published genome-wide association study (GWAS). These findings add to a growing body of work demonstrating the utility of network-based prior knowledge in improving the search for disease-associated genes[2,4,7]. GWAS often reveal marginally significant associations between various single-nucleotide polymorphisms (SNPs) and a disease of interest. However, if several identified genes are found to interact within the same network neighborhood, this event can be more significant than any of the individual SNPs on which it is based. The result not only improves statistical power but also helps home in on the mechanisms underlying the SNPs by isolating the pathways involved.

Using such an approach to reanalyze the genes identified in a GWAS of hypertension, Greene *et al.*[1] identify known disease-causing genes as well as new candidate disease-associated

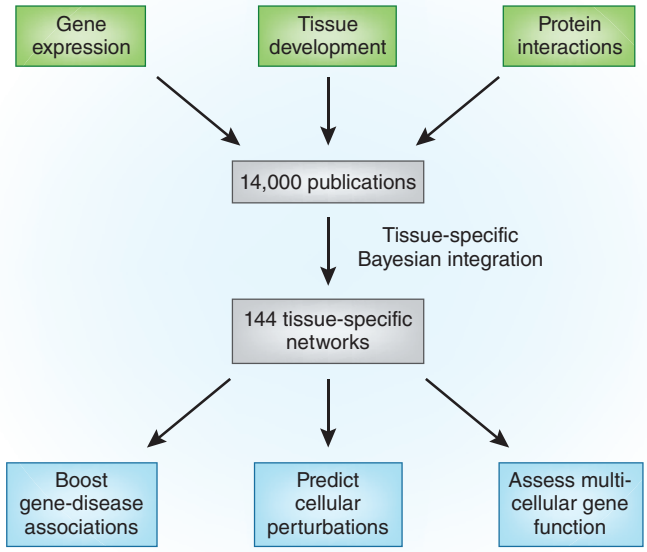


Figure 1 Overview of the approach by Greene *et al.*[1] to construct tissue-specific networks. Data from diverse experiments (green) from over 14,000 publications were used to construct interaction networks for 144 human tissues and cell types. Tissue-specific networks were applied to answer a range of biological questions (blue).

Andrew M. Gross and Trey Ideker are at the Bioinformatics Program and Department of Medicine, University of California, San Diego, San Diego, California, USA.
e-mail: tideker@ucsd.edu

genes that may control blood pressure. The availability of such a large number of tissue-specific maps will expand the phenotypic scope for investigating candidate-gene function in disease-relevant tissues.

Although the tissue-specific networks in Greene *et al.*[1] represent an important advance, improving accuracy and including more tissues and conditions remain key goals for the field. As in all data-driven approaches, differences in data quality and size can have a large impact on the ability to predict interactions. In the present study, some tissues had higher predictive power than others, and cross-validation of even the best network maps showed plenty of room for improvement (area under the receiver operating curve <0.65). Nonetheless, such networks may still prove extremely valuable in understanding how genes contribute to disease.

Beyond its technical achievements, the work of Greene *et al.*[1] is notable for distilling large and diverse data sets into a form accessible to a wide scientific audience. The authors' tissue-specific maps are available for download and can be interrogated directly through a web portal, which will facilitate rapid adoption. As studies such as these are extended to an even wider range of cell types and conditions, network biology may play a greater role in personalized medicine by aiding the interpretation of patients' genetic and phenotypic information. Only by defining the plasticity of the interactome in its many contexts can we truly begin to understand the functioning of complex organisms in health and disease.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Greene, C.S. *et al. Nat. Genet.* **47**, 569–576 (2015).
2. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. & Marcotte, E.M. *Genome Res.* **21**, 1109–1121 (2011).
3. Ideker, T. & Krogan, N.J. *Mol. Syst. Biol.* **8**, 565 (2012).
4. Magger, O., Waldman, Y.Y., Ruppin, E. & Sharan, R. *PLoS Comput. Biol.* **8**, e1002690 (2012).
5. Cahan, P. *et al. Cell* **158**, 903–915 (2014).
6. Morris, S.A. *et al. Cell* **158**, 889–902 (2014).
7. Ganegoda, G., Wang, J., Wu, F.-X. & Li, M. *BMC Syst. Biol.* **8**, S3 (2014).

Research Highlights

Papers from the literature selected by the Nature Biotechnology editors. (Follow us on Twitter, @NatureBiotech #nbtHighlight)

Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets
Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015)

Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells
Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015)

Effect of predicted protein-truncating genetic variants on the human transcriptome
Rivas, M.A. *et al. Science* **348**, 666–669 (2015)

Comprehensive serological profiling of human populations using a synthetic human virome
Xu, G.J. *et al. Science* doi:10.1126/science.aaa0698 (5 June 2015)

A microfluidic device for label-free, physical capture of circulating tumor cell clusters
Sarioglu, A.F. *et al. Nat. Methods* doi:10.1038/nmeth.3404 (18 May 2015)