Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases

Daniel Marbach[1,2], David Lamparter[1,2], Gerald Quon[3,4], Manolis Kellis[3,4], Zoltán Kutalik[2,5] & Sven Bergmann[1,2]

Mapping perturbed molecular circuits that underlie complex diseases remains a great challenge. We developed a comprehensive resource of 394 cell type– and tissue-specific gene regulatory networks for human, each specifying the genome-wide connectivity among transcription factors, enhancers, promoters and genes. Integration with 37 genome-wide association studies (GWASs) showed that disease-associated genetic variants—including variants that do not reach genome-wide significance—often perturb regulatory modules that are highly specific to disease-relevant cell types or tissues. Our resource opens the door to systematic analysis of regulatory programs across hundreds of human cell types and tissues (http://regulatorycircuits.org).

GWASs have successfully identified thousands of genetic loci associated with complex traits and diseases. However, translating these findings into a functional understanding of disease processes remains a major challenge, notably because the effects of individual trait-associated variants are typically minute, underlying mechanisms are generally cell-type specific[1–3], and most variants lie in poorly understood noncoding regions of the genome[4].

Integration of regulatory genomics data is emerging as a promising strategy to address these challenges: it has been shown that GWAS variants are enriched in regulatory regions of cell types that are relevant to the pathophysiological basis of a given trait[5,6], and regulatory annotations have been used to prioritize and fine-map GWAS loci[7–9]. However, such studies do not consider the interplay between variants at the pathway and network levels. Pathway- and network-based approaches, in contrast, have been successful at identifying relevant pathways or modules on the basis of connectivity between trait-associated genes, but they typically rely on protein-protein interaction[10,11], coexpression[12] or functional association networks[13] lacking fine-grained regulatory and, with few exceptions[14–16], tissue-specific information. Indeed, a comprehensive compendium of tissue-specific regulatory circuits was previously not available, as most studies focused on building gene regulatory networks either globally[17–19] or for a single tissue or condition of interest[20–22].

Here we introduce a unique resource of 394 cell type– and tissue-specific gene regulatory networks for human. We infer networks by integrating transcription factor (TF) sequence motifs with promoter and enhancer activity data from the FANTOM5 project[23,24] (**Fig. 1**). All networks and tools are freely available at http://regulatorycircuits.org.

RESULTS

Cell type– and tissue-specific gene regulatory circuits

Our pipeline for reconstructing transcriptional regulatory circuits involves (1) genome-wide mapping of promoters and enhancers, (2) linking TFs to promoters and enhancers and (3) linking enhancers and promoters to target genes (**Fig. 1a**,**b** and Online Methods). Here we applied this approach to data from the FANTOM5 consortium[23,24], which has performed cap analysis of gene expression (CAGE) for ~1,000 human primary cell, tissue and cell line samples. CAGE maps regions of transcription initiation with high sensitivity, enabling the identification of both active promoters and enhancers (active enhancers are often transcribed at low levels[24]). To identify regulatory inputs of promoters and enhancers, we used a curated collection of sequence binding motifs for 662 TFs[25,26]. We found greater than tenfold enrichment of these TF motifs in CAGE-defined promoters and distal enhancers, supporting the FANTOM5 data (Online Methods and **Supplementary Figs. 1–3**).

We used this pipeline to infer 394 weighted, transcriptional regulatory circuits, including 146 different cell types, 111 tissues and 137 cell lines (**Supplementary Table 1** and **Supplementary Figs. 4** and **5**). Each circuit has up to 662 TFs, 41,455 enhancers, 59,126 promoters and 43,352 gene isoforms from 19,125 protein-coding genes (**Supplementary Figs. 6–8**).

ChIP-seq, eQTL and RNA-seq data support regulatory edges

Although systematic validation of our regulatory circuits is difficult because the ground truth is not known[27], independent data sets that support regulatory edges are available for a subset of cell lines and tissues. First we assessed the inferred TF-enhancer and TF-promoter edges using TF binding data (chromatin immunoprecipitation followed by sequencing

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [3]Broad Institute, MIT, Cambridge, Massachusetts, USA. [4]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts, USA. [5]Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland. Correspondence should be addressed to D.M. (daniel.marbach@unil.ch).

RECEIVED 23 JULY 2015; ACCEPTED 26 JANUARY 2016; PUBLISHED ONLINE 7 MARCH 2016; DOI:10.1038/NMETH.3799

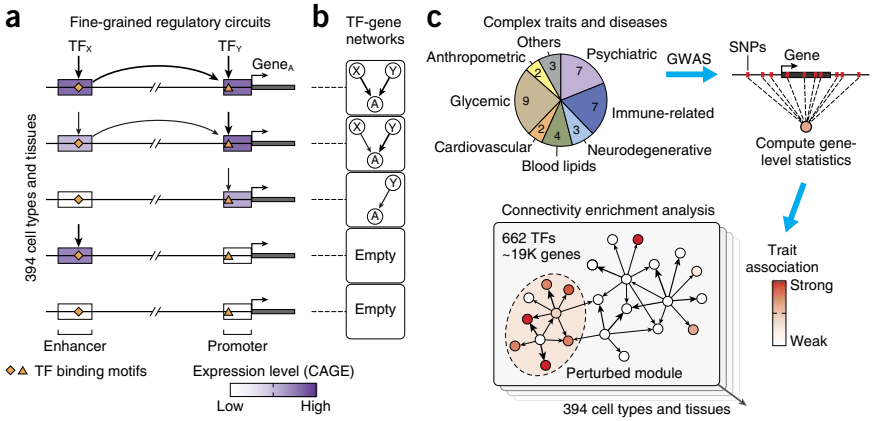

Figure 1 | Regulatory circuit inference and GWAS analysis. (**a**) Schematic showing how cell type– and tissue-specific regulatory circuits are derived from expression profiles of CAGE-defined enhancers and promoters from the FANTOM5 project[23,24]. Weighted, tissue-specific links between TFs and regulatory elements (enhancers and promoters) are inferred using TF binding motifs and tissue-specific expression of target elements. Links between regulatory elements and target genes are inferred on the basis of genomic distance and joint expression in the given tissue. Different circuit configurations are shown schematically for five tissues. For clarity, only one enhancer and promoter are shown, but genes typically have multiple tissue-specific enhancers and promoters. (**b**) Coarse-grained TF-gene networks encapsulate the fine-grained circuitry of enhancers and promoters. (**c**) Schematic showing how the interconnectivity of genes that are perturbed by trait-associated variants was systematically assessed for a large panel of 37 GWASs.

(ChIP-seq)) from ENCODE (Encyclopedia of DNA Elements) for 59 TFs in five cell lines (**Supplementary Table 2**). Overall, about half of the TF edges were confirmed by a corresponding ChIP-seq peak, which is not far from the reproducibility of the ChIP-seq data themselves (**Fig. 2a** and **Supplementary Figs. 9** and **10**).

Second, we assessed enhancer-gene edges of our circuits in 13 tissues for which expression quantitative trait locus (eQTL)

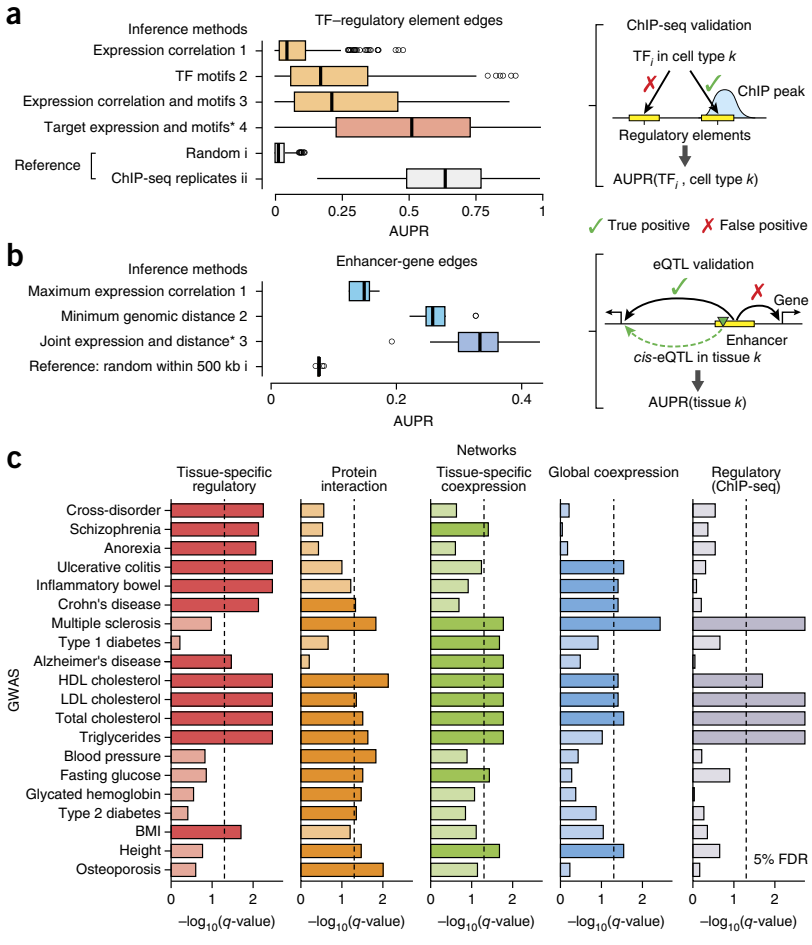

Figure 2 | Assessment of regulatory circuits. (**a**) Evaluation of different approaches for inferring edges between TFs and regulatory elements (enhancers and promoters): (1) standard network inference based on expression correlation between TFs and regulatory elements across samples; (2) presence of TF motifs within regulatory elements; (3) presence of TF motifs, weighted by expression correlation; and (4) presence of TF motifs, weighted by target-element expression in the given cell type (Online Methods). For each TF and cell type for which ChIP-seq data were available (159 samples, including 59 TFs and five cell lines), the area under the precision-recall curve (AUPR) was computed. As a reference, AUPR values were also computed for (i) random data and (ii) replicates of ChIP-seq experiments. Box plots show the distribution of AUPR values for each method. Right and left hinges correspond to the first and third quartiles; whiskers extend to the highest and lowest values within 1.5× the interquartile range of the hinges. (**b**) Assessment of different approaches for linking enhancers to target genes: (1) maximum expression correlation across tissues; (2) minimum genomic distance; and (3) joint tissue-specific activity weighted by genomic distance (Online Methods). The AUPR was evaluated for each method along with random predictions in 13 tissues for which eQTL data were available. (Of note, AUPR values are not comparable across panels (**a**,**b**) because the underlying gold standards are different.) Asterisks in **a** and **b** indicate retained methods. (**c**) Evaluation of whether trait-associated genes tend to cluster in modules for different types of networks and GWAS traits. Five types of networks are compared from left to right: cell type– and tissue-specific regulatory networks (the 32 high-level networks), four protein-protein interaction networks, 35 tissue-specific coexpression networks[15], a global coexpression network inferred from the FANTOM5 data, and a global regulatory network based on ChIP-seq[17] (Online Methods). The plot summarizes whether trait-associated genes are more densely interconnected than expected (maximum connectivity enrichment score) for each network type (column) and trait (row). False discovery rate (FDR) correction was performed separately for each network type to allow for fair comparison. Only traits passing the 5% FDR threshold (dashed lines) for at least one network are shown (full results are presented in **Supplementary Fig. 31**). Tissue-specific regulatory networks showed the strongest connectivity enrichment overall.

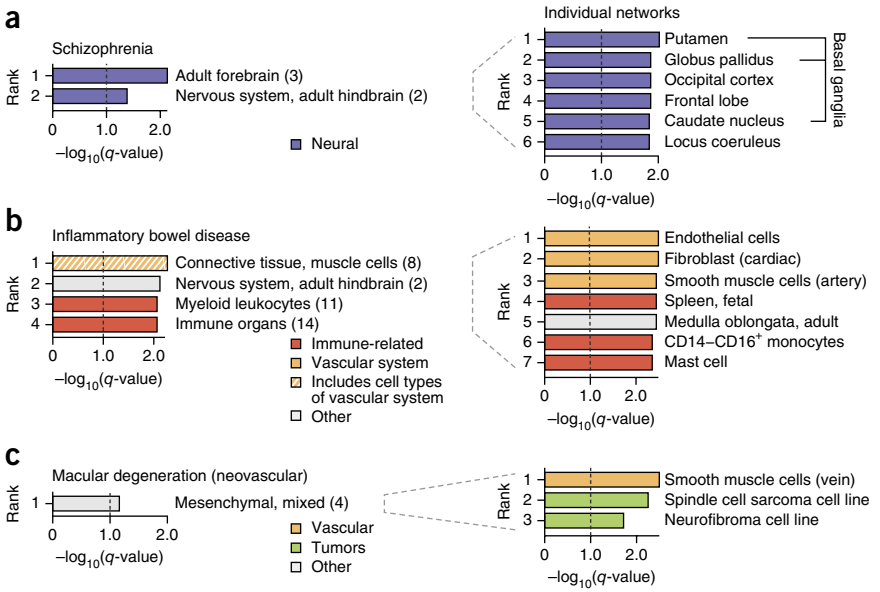

Figure 3 | Network-connectivity enrichment reveals disease-relevant cell types and tissues. Shown are connectivity enrichment scores across the 32 high-level networks (left; numbers in parentheses correspond to cluster indexes in **Supplementary Fig. 14**) and corresponding individual networks (right) for three GWAS traits. Similar results were obtained for the remaining traits (discussed in the text). Only top-ranking networks are shown (full results are presented in **Supplementary Figs. 34, 37** and **47**). (**a**) Schizophrenia showed the strongest clustering of associated genes in high-level regulatory networks of neural tissues (left), with top-ranking individual networks pinpointing specific, disease-relevant brain structures (right). (**b**) Inflammatory bowel disease (IBD) displayed connectivity enrichment in immune-related networks. IBD also showed enrichment in other high-level networks, including vascular cells that are involved in the inflammatory response and drive the signal, as shown to the right. (**c**) Age-related macular degeneration of the neovascular type showed the strongest connectivity enrichment in regulatory networks of vascular smooth muscle cells, followed by diverse tumors, which often induce vascularization to achieve growth.

data were available from GTEx[28] (**Supplementary Table 3** and Online Methods). Overall, about one-third of the enhancer-gene edges were confirmed by a *cis*-eQTL, which is an improvement on results obtained by alternative approaches such as assigning enhancers to the closest or most strongly correlated target gene (**Fig. 2b** and **Supplementary Figs. 11** and **12**).

Lastly, we evaluated whether the regulatory circuits are predictive of gene expression levels using RNA-seq data for 40 matching tissues from the Roadmap Epigenomics project[29] (**Supplementary Table 4**). We indeed found a strong correlation between the TF input of a gene (defined as the sum of its incoming edge weights) and its expression level in a given tissue (Online Methods and **Supplementary Fig. 13**).

Network architecture differs between lineages

To compare regulatory programs of genes across lineages, we defined for each regulatory circuit a corresponding coarse-grained network that summarized TF inputs at the level of genes (**Fig. 1b** and Online Methods). We hierarchically clustered these networks on the basis of edge overlap (Online Methods). As expected, developmentally and functionally related lineages were consistently grouped together, indicating that they share regulatory components (**Supplementary Figs. 14–24**). Analysis of network architecture showed that the highest numbers of TFs per gene were found in immune cells, followed by cells and tissues of the nervous system, whereas other network properties were comparable across lineages (**Supplementary Figs. 25** and **26**).

GWAS variants perturb tissue-specific regulatory modules

To systematically evaluate the relevance of tissue-specific regulatory networks for a broad range of traits and diseases, we compiled a large panel of 37 GWASs (**Supplementary Table 5**) and assessed whether trait-associated variants perturb genes that cluster in network modules (connectivity-enrichment analysis; **Fig. 1c** and **Supplementary Figs. 27–30**). In our pipeline, first we integrated GWAS summary statistics at the level of genes using Pascal[30], a tool that accurately corrects for confounders such as linkage disequilibrium, and then we evaluated whether top-ranked genes tend to cluster in network modules using a random-walk graph kernel (Online Methods). To reduce the burden of multiple testing (394 networks × 37 GWASs), we first considered 32 high-level networks derived from the hierarchical clustering described in the preceding section. (Networks in the same cluster were merged.) We found that the majority of GWAS traits showed stronger connectivity enrichment in tissue-specific regulatory networks than in protein interaction, coexpression and ChIP-based networks (**Fig. 2c**). Moreover, connectivity enrichment often extended to weakly associated genes that did not pass the GWAS significance threshold (**Supplementary Fig. 31**). Some traits did not show significant enrichment, possibly because the signal was too weak, the right tissues were not profiled, or other types of networks are more relevant for these traits.

Next we asked whether the observed clustering of perturbed genes was specific to regulatory networks of trait-relevant tissues. We examined all GWAS traits that showed connectivity enrichment in at least one network and ranked the 32 high-level networks by enrichment score (Online Methods). Networks of disease-relevant cell types and tissues consistently ranked at the top, thus showing the strongest evidence for perturbed modules. For example, schizophrenia showed the strongest clustering of associated genes precisely in the two high-level networks specific to adult brain (**Fig. 3a**). Going one step further, for all high-level networks that showed connectivity enrichment, we also assessed the individual cell type– and tissue-specific networks pertaining to the corresponding clusters. For example, in the two adult brain clusters that showed signal for schizophrenia, we found the strongest clustering of perturbed genes in components of the basal ganglia, which modulate motor, cognitive and emotional behavior (**Fig. 3a**). The dopaminergic system of the basal ganglia manifests pathological anomalies in schizophrenia patients and is the primary target of current antipsychotic drugs[31]. This example

demonstrates that perturbed regulatory modules can pinpoint disease-relevant tissues with remarkable precision.

We made similar observations for the remaining traits (**Supplementary Figs. 32–47**). The psychiatric cross-disorder study showed strong clustering of perturbed genes in regulatory networks of the caudate nucleus, thalamus, locus coeruleus and other brain structures underlying cognitive and emotional functions that are impaired in psychiatric disorders (**Supplementary Fig. 33**). For bipolar disorder, the amygdala, a group of brain nuclei key to memory, experienced emotions and mood that has been consistently implicated in the illness[32], ranked first (**Supplementary Fig. 36**). Inflammatory bowel disease showed strong connectivity enrichment in endothelial cells, which are gatekeepers of inflammatory processes targeted by current drugs[33], as well as in immune cells and tissues involved in pathogenesis (spleen cells, monocytes and mast cells[34,35]; **Fig. 3b**). For rheumatoid arthritis, the strongest clustering of perturbed genes was found in neutrophils, which have an activated phenotype in people with the condition and contribute to pathogenesis[36] (**Supplementary Fig. 40**). For Alzheimer's disease, the strongest clustering of associated genes occurred in regulatory networks of adult forebrain, followed by endothelial cells, suggesting that neurovascular dysregulation might be implicated[37] (**Supplementary Fig. 41**). Narcolepsy, a rare sleep disorder caused by autoimmune targeting of hypocretin-producing neurons[38], showed connectivity enrichment in neural stem cells (**Supplementary Fig. 42**). Body mass index showed the strongest clustering of perturbed genes in regulatory networks of the intestinal tract and immune system, suggesting a potential link with the gut microbiome, which has recently been proposed to have a heritable component that affects body mass index[39] (**Supplementary Fig. 45**). Finally, of the two types of age-related macular degeneration, only the neovascular type caused by abnormal blood vessel growth showed network-connectivity enrichment in vascular smooth muscle cells, consistent with their function in neovascularization (**Fig. 3c**).

DISCUSSION

Recently there has been growing interest in tissue-specific gene networks[14–16], and our compendium of regulatory circuits further expands the number of cell types and tissues for which networks are available. These resources are enabling increasingly fine-grained analysis of molecular circuits underlying human physiology and disease.

The FANTOM5 enhancers and promoters have been validated previously[23,24], and the observed enrichment for the independently curated TF motifs provides further support. In addition, we validated the inferred tissue-specific edges using independent ChIP-seq, eQTL and RNA-seq data, in each case finding good support. Of note, TF motifs alone are known to have low specificity for predicting TF binding, as we confirmed. Key to our approach is that we require both a TF motif and activity of the corresponding regulatory element to infer an edge in a given tissue. Implicitly, this approach also accounts for cofactors on which the TF may be dependent for binding at a specific regulatory element: if cofactors or other conditions necessary to activate the TF are not present, the regulatory element is not active, and thus no edge is added in that tissue, explaining the high specificity of our circuits.

Genetic association data provide an orthogonal means of validation; our systematic analysis across 37 GWASs showed that perturbed regulatory modules often pinpoint cell types and tissues that are known to be involved in pathophysiological processes, and in some cases that are targets of current medications, with remarkable precision. For most traits, evidence of increased connectivity between perturbed genes extended to variants that did not pass the genome-wide significance threshold, indicating that regulatory network information will be useful for prioritizing candidate variants.

Our analysis of network architecture showed that similarity between regulatory networks closely reflects developmental or functional relationships between cell types and tissues, suggesting that phenotypic relatedness results from shared circuit components. We also found variability in network architecture across different lineages, with regulatory networks of the immune and nervous system having by far the most TF inputs per gene. Both immune cells and neurons perform highly adaptive functions; our networks suggest that the orchestration of transcriptional responses in these cells involves intricate regulatory programs. Interestingly, the high number of TF inputs (i.e., binding sites) also makes these genes more likely to be perturbed by regulatory variants, which suggests that immune and neural processes may be particularly prone to genetic dysregulation. This proposition is further supported by the strong network-connectivity enrichment that we observed for immune-related, psychiatric and neurodegenerative disorders. Tissue-specific regulatory networks may thus be key to understanding the etiology of these disorders.

We have not yet attempted to infer causal variants in the context of our circuits. This is a difficult task because a given GWAS locus often spans multiple genes and regulatory regions, harboring many potentially causal variants. Another challenge is the fine-grained characterization of perturbed pathways and their role in disease processes, which will be extremely valuable for biomarker and drug discovery. We make our networks freely available, along with user-friendly software tools, and we hope that this resource will spur the development of additional methods leveraging large network compendia to dissect the tissue-specific regulatory architecture of complex traits.

METHODS

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

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

ACKNOWLEDGMENTS

We thank P. Kheradpour (MIT) for providing the collection of curated TF binding motifs. This work was supported by the Swiss National Science Foundation (grant FN 310030_152724/1 to S.B. and grant FN 31003A-143914 to Z.K.), SystemsX.ch (grant SysGenetiX to S.B. and grant AgingX to Z.K.), the Swiss Institute of Bioinformatics (Z.K. and S.B.) and the Leenaards Foundation (Z.K.).

AUTHOR CONTRIBUTIONS

D.M. designed the study, performed analyses and prepared the manuscript. D.L. performed gene scoring and phenotype-label permutation. D.M., D.L., G.Q., M.K., Z.K. and S.B. conceived methods, discussed the results and implications, and commented on the manuscript at all stages.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Marstrand, T.T. & Storey, J.D. Identifying and mapping cell-type-specific chromatin programming of gene expression. *Proc. Natl. Acad. Sci. USA* **111**, E645–E654 (2014).
2. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
3. Roy, S. *et al.* A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.* **43**, 8694–8712 (2015).
4. Ward, L.D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
5. Parker, S.C.J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. USA* **110**, 17921–17926 (2013).
6. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
7. Faye, L.L., Machiela, M.J., Kraft, P., Bull, S.B. & Sun, L. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS Genet.* **9**, e1003609 (2013).
8. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
9. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
10. Navlakha, S. & Kingsford, C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics* **26**, 1057–1063 (2010).
11. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7**, e1001273 (2011).
12. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
13. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. & Marcotte, E.M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).
14. Mäkinen, V.-P. *et al.* Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet.* **10**, e1004502 (2014).
15. Pierson, E., Koller, D., Battle, A., Mostafavi, S. & the GTEx Consortium. Sharing and specificity of co-expression networks across 35 human tissues. *PLoS Comput. Biol.* **11**, e1004220 (2015).
16. Greene, C.S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
17. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
18. Marbach, D. *et al.* Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.* **22**, 1334–1349 (2012).
19. Karczewski, K.J., Snyder, M., Altman, R.B. & Tatonetti, N.P. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. *PLoS Genet.* **10**, e1004122 (2014).
20. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
21. Ciofani, M. *et al.* A validated regulatory network for Th17 cell specification. *Cell* **151**, 289–303 (2012).
22. Chen, J.C. *et al.* Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell* **159**, 402–414 (2014).
23. The FANTOM Consortium & the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
24. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
25. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).
26. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* **42**, 2976–2987 (2014).
27. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
28. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
29. Roadmap Epigenomics Consortium. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
30. Lamparter, D., Marbach, D., Rico, R., Kutalik, Z. & Bergmann, S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.* **12**, e1004714 (2016).
31. Perez-Costas, E., Melendez-Ferro, M. & Roberts, R.C. Basal ganglia pathology in schizophrenia: dopamine connections and anomalies. *J. Neurochem.* **113**, 287–302 (2010).
32. Garrett, A. & Chang, K. The role of the amygdala in bipolar disorder development. *Dev. Psychopathol.* **20**, 1285–1296 (2008).
33. Cromer, W.E., Mathis, J.M., Granger, D.N., Chaitanya, G.V. & Alexander, J.S. Role of the endothelium in inflammatory bowel diseases. *World J. Gastroenterol.* **17**, 578–593 (2011).
34. Swirski, F.K. *et al.* Identification of splenic reservoir monocytes and their deployment to inflammatory sites. *Science* **325**, 612–616 (2009).
35. Chichlowski, M., Westwood, G.S., Abraham, S.N. & Hale, L.P. Role of mast cells in inflammatory bowel disease and inflammation-associated colorectal neoplasia in IL-10-deficient mice. *PLoS ONE* **5**, e12220 (2010).
36. Wright, H.L., Moots, R.J. & Edwards, S.W. The multifactorial role of neutrophils in rheumatoid arthritis. *Nat. Rev. Rheumatol.* **10**, 593–601 (2014).
37. Iadecola, C. Neurovascular regulation in the normal brain and in Alzheimer's disease. *Nat. Rev. Neurosci.* **5**, 347–360 (2004).
38. Hor, H. *et al.* Genome-wide association study identifies new HLA class II haplotypes strongly protective against narcolepsy. *Nat. Genet.* **42**, 786–789 (2010).
39. Goodrich, J.K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).

ONLINE METHODS

Inference of cell type– and tissue-specific regulatory networks. We represent transcriptional regulatory networks as graphs including four different types of nodes: TFs, enhancers, promoters and gene isoforms. Directed, weighted edges connect TFs to enhancers and promoters, and enhancers and promoters to gene isoforms. Inference of regulatory circuits involves defining the nodes and the edges for each of these layers. To this end, we developed a pipeline consisting of four steps.

The first step is to map regulatory elements (promoters and enhancers) and their tissue-specific activity. Both chromatin-state and CAGE-defined maps can be used for this purpose. Here we used the CAGE-defined maps from the FANTOM5 project because they currently cover the greatest number of human cell and tissue types (~1,000 samples, compared to 127 epigenomes available from the ENCODE[40] and Roadmap Epigenomics[29] projects). First, we obtained the FANTOM5 set of 184,827 robust CAGE peaks and their activity levels across all samples (relative log expression (RLE)-normalized expression profiles)[23]. These CAGE peaks represent high-confidence regions of transcription initiation; we refer to them as CAGE-defined promoters (note that the full promoter region where TFs bind extends beyond these peaks, as described below). Second, we obtained the FANTOM5 set of 43,012 CAGE-defined enhancers and their activity levels across all samples (RLE-normalized expression profiles)[24]. These enhancers have been identified on the basis of enhancer RNAs that are detectable through a robust signature of weak, bidirectional transcription in CAGE data[24]. After discarding samples that were not present in either promoter or expression data, we were left with 808 cell and tissue samples, which formed the basis for regulatory circuit reconstruction (**Supplementary Table 1**).

The second step is to link TFs to promoters and enhancers. To this end, we used a curated collection of sequence binding motifs (position weight matrices) for 662 TFs, with each motif occurrence in the genome (referred to as a motif instance) further assigned a confidence score based on its evolutionary conservation across mammals[25,26,41]. On the basis of our positional enrichment analysis (described in the next section), we determined that motif instances are enriched more than tenfold in a window 400 bp upstream to 50 bp downstream of CAGE-defined promoters (**Supplementary Fig. 1**). We thus linked TFs to CAGE-defined promoters on the basis of the occurrence of TF motif instances in these windows (of note, as the bulk of the motif instances are located around the transcription start site (TSS; distribution in **Supplementary Fig. 1a**), the cutoff at tenfold enrichment is not a critical parameter; using a less stringent cutoff expands the windows but adds only a small fraction of edges to the networks). The weight of TF-promoter edges was defined as the confidence score of the corresponding TF motif instance (if multiple motif instances of the same TF were found in a given window, the maximal confidence score was taken). Almost the exact same approach was used to link TFs to enhancers, the only difference being that the window where motif instances were more than tenfold enriched coincided with the CAGE-defined enhancers (i.e., it did not extend up- or downstream). Edge weights were defined in the same way as for TF-promoter links.

The third step is to link CAGE-defined promoters to known gene isoforms. We used the Ensembl genome annotation, comprising 53,449 isoforms from 19,125 protein-coding genes (accessed May 31, 2014). Note that the isoforms, not the genes, constitute the nodes of our circuits. This is because different isoforms of the same gene often have distinct TSSs that are under the control of independent promoters, which may have different regulatory inputs and tissue-specific activity. We linked CAGE-defined promoters to TSSs of gene isoforms using the same approach as described above: we determined that CAGE-defined promoters enrich more than tenfold in a window 250 bp upstream to 500 bp downstream around TSSs and linked promoters within these windows to the corresponding gene isoforms. The weight of promoter-gene edges was defined as the normalized activity level of the promoter across all samples (normalization was done per regulatory element because expression levels of diverse enhancers and promoters might not be on the same scale). Thus, if the promoter is not active in a given cell type, the edge weight is 0 (i.e., the edge is not present), and if the promoter is maximally active, the edge weight equals 1.

The fourth step is to link enhancers to their target gene isoforms, which is more challenging because enhancers are often distal and their targets may be cell-type specific[1,3]. We opted for a parsimonious approach weighting potential enhancer-isoform links on the basis of just two factors: their genomic distance and their activity level in the given tissue. The weighting function for the distance was derived directly from the distance distribution of known *cis*-eQTLs from their target genes. To this end, we obtained 38,935 high-confidence *cis*-eQTLs from RegulomeDB[42], computed their distance from the TSSs of their target genes, and defined the weighting function as a smooth fit of the resulting distance distribution (we used local polynomial regression fitting). The weighting function was defined for the range (1 kb, 500 kb) and scaled to give a maximum weight of 1; enhancers at a distance of <1 kb and >500 kb were assigned weights of 1 and 0, respectively (**Supplementary Fig. 1d**). This implies that enhancer-isoform interactions over more than 500 kb were not included in our networks. However, the *cis*-eQTL distribution shows that this is only a minor fraction of all interactions (**Supplementary Fig. 1d**; less than 6% of *cis*-eQTLs are located more than 500 kb away from their target gene), and our correlation analysis suggests that these long-range interactions would be difficult to identify from CAGE data alone (**Supplementary Figs. 2** and **3**). Given the distance-based weight d_{ij} for a given enhancer i and gene isoform j, we defined the weight of the corresponding edge in cell type k as

$$w_{ij}(k) = d_{ij} \times \sqrt{x_i(k) y_i(k)}$$

where the second term is the geometric mean of the normalized activity levels of the enhancer and the gene isoform. The activity level of isoforms was defined as the maximum activity level of their promoters (which are usually few—the majority of isoforms have only one or two alternative promoters; **Supplementary Fig. 6**). Accordingly, confidence that an enhancer regulates a gene isoform increases as the distance between them decreases and their joint activity in the given cell type increases.

Application of this pipeline to the FANTOM5 panel of samples gave rise to 808 regulatory circuits. Next, we merged regulatory circuits of closely related samples (**Supplementary Table 1**) (e.g., samples of the same cell or tissue type from different donors,

doi:10.1038/nmeth.3799

or cell lines from the same cancer subtype) by taking the graph union (this amounts to taking the union of the node and edge sets while retaining the maximum weight for each edge). The resulting 394 cell type– and tissue-specific regulatory circuits constituted the basis of our analysis (corresponding sample annotations are provided in **Supplementary Table 1**). Note that the samples being merged were typically very similar to each other: if one of them was left out, on average 94% of edges in the final network remained the same (**Supplementary Figs. 4** and **5**).

We kept our reconstruction approach for regulatory circuits deliberately simple, making it generally applicable; here we used CAGE-defined enhancers and promoters from FANTOM, but the same approach could be applied directly to chromatin-defined enhancers and promoters, for instance. Future work will show whether performance can be improved by the use of more sophisticated inference algorithms[27] for specific steps—for example, to infer the target regulatory elements of TFs or to infer the target genes of enhancers.

Positional enrichment of motif instances and regulatory elements. Positional enrichment of TF motif instances near CAGE-defined promoters (**Supplementary Fig. 1a**) was computed as follows. First, a window of 10 kb was defined around each promoter, and the distance of each motif instance from the promoter within this window was evaluated. Distance was defined as the number of base pairs separating the motif from the promoter, with the sign indicating whether the motif was located up- or downstream from the promoter (negative, upstream; positive, downstream; zero, the two elements overlap). The empirical distribution of motif-promoter distances was computed using a bin size of 50 bp. Note that only promoters of genes with a single TSS were considered (to avoid potential bias due to nearby TSSs of isoforms). Second, we computed a background distribution by shuffling motif instances in the region defined by the union of all promoter-centric windows. The resulting motif-promoter distances and corresponding distance distribution were evaluated as described. This procedure was repeated 100 times, leading to a very precise estimate of the background distribution owing to the large number of motif instances and promoters. Finally, positional enrichment of motifs was defined as the ratio of the actual distance distribution to the background distance distribution. The entire analysis was performed independently for motif instances of distinct confidence scores (0.1, 0.2, …, 1.0).

The same approach was used to compute positional enrichment of TF motif instances near CAGE-defined enhancers (**Supplementary Fig. 1b**) and of CAGE-defined promoters near TSSs (**Supplementary Fig. 1c**). For the former, there was a slight difference because enhancers have no orientation (i.e., no negative distance values for upstream location). Note that only enhancers located more than 10 kb away from any known TSS in Ensembl were considered, so as to avoid potential bias due to nearby promoters.

Validation of TF-enhancer and TF-promoter edges using ChIP-seq data. We assessed edges between TFs and regulatory elements (enhancers and promoters) using ChIP-seq data from ENCODE[40] in five cell lines that are also present in our library (**Supplementary Table 2**). Besides our method to infer edges described above, which weights edges on the basis of motif confidence and regulatory element expression, we also assessed three alternative inference approaches: (1) standard network inference based on expression correlation[27] (edges are weighted using Spearman correlation between TF expression and regulatory element expression across samples), (2) TF motifs alone (edges are weighted on the basis of motif confidence), and (3) both TF motifs and expression correlation (edges are weighted on the basis of correlation and are filtered for those containing at least one motif instance of the given TF at the target regulatory element).

For each TF and cell line for which ChIP-seq data were available, we assessed inferred TF-target edges using the area under the precision-recall curve[27,43] (AUPR; this choice is discussed below). We defined edges as positives (true edges) if they were supported by TF binding (i.e, if there was a ChIP-seq peak of the TF overlapping the target regulatory element) and as negatives (absent edges) otherwise. It should be kept in mind that ChIP-seq is an imperfect gold standard, and some edges may thus have been incorrectly labeled as negatives (binding was not detected) or positives (nonspecific or nonfunctional binding). The AUPR was computed separately for TF-enhancer and TF-promoter edges.

As a reference, we further computed the AUPR for (i) random data, obtained by shuffling the ChIP peaks of a given TF within the bound regions (the union of ChIP peaks from all TFs) and (ii) ChIP-seq replicates (edges defined by the first replicate were assessed using the second replicate and vice versa). The AUPR values for random data and ChIP-seq replicates represent lower and upper bounds for the expected performance of inference methods.

We chose to report the AUPR for TF-target edges (and for enhancer-gene edges, as discussed in the next section) in the main text because it is a standard metric in the field of network inference; for example, it is used in the DREAM network inference challenges to assess predictions[27,43]. Note that because of class imbalance—there are few positives compared to negatives—the AUPR is better suited as performance metric than the area under the receiver operating characteristic (ROC) curve. We obtained similar results when we used the *F*-score as a performance metric (**Supplementary Fig. 12**).

Validation of enhancer-gene edges using eQTL data. Edges between enhancers and genes were assessed using *cis*-eQTLs from GTEx[28] in 13 tissues that are also present in our library (**Supplementary Table 3**; GTEx data were not used to construct the regulatory circuits; see discussion below). In addition to our method described above, which links enhancers to target genes on the basis of their genomic distance and joint activity in the given tissue, we also assessed two alternative approaches: (1) linking enhancers to the closest gene (minimum genomic distance) and (2) linking enhancers to the most strongly correlated gene (maximum Spearman correlation of enhancer and gene expression across samples). For each of these methods, only genes within 500 kb from the enhancer were considered. As a reference, we further assessed the performance for random predictions (linking enhancers to randomly selected genes within the given window of 500 kb).

For each tissue for which eQTLs were available, we assessed the inferred enhancer-gene edges using the AUPR and *F*-score (the same approach as for TF edges; see previous section). Only enhancers that contained at least one eQTL were considered. We defined edges as positives if they were confirmed by an eQTL (i.e., the enhancer contained an eQTL for that target gene) and

as negatives otherwise. As mentioned above, both ChIP-seq and eQTLs are imperfect gold standards, and thus their use may lead to incorrectly labeled edges. In particular, the eQTLs from GTEx are known to be largely incomplete, as they were called using only ~100 samples per tissue[28].

Of note, the tissue-specific eQTLs from GTEx used for validation are independent from our regulatory circuits, in which enhancers were linked to target genes solely on the basis of their distance and joint tissue-specific activity (possible eQTLs of a given enhancer were never used to determine its targets). The weighting function used for the distance (**Supplementary Fig. 1d**) was the same for all enhancers and tissues and was derived from a different eQTL data set (RegulomeDB[42]) that does not include GTEx tissues.

Correlation between regulatory edge strength and target gene expression level. We assessed whether incoming regulatory edges of a gene are predictive of its expression level using RNA-seq data from the Roadmap Epigenomics project[29] in 40 cell types and tissues that are also present in our library (**Supplementary Table 4**). First, we defined the gross TF input of a gene as its weighted in-degree in the given tissue, that is, the sum of the weights of all incoming TF edges (TF-gene edges were given by the coarse-grained networks, as defined in the next section). For each gene, the correlation coefficient between its TF input and expression level (RPKM value) was computed across the 40 tissues (**Supplementary Fig. 13**).

Hierarchical clustering of regulatory networks. First, we derived a coarse-grained TF-gene network from each cell type– or tissue-specific regulatory circuit. TF-gene networks are useful for high-level analyses in which one is most interested in the TFs regulating each gene, and not the detailed and often redundant wiring of these interactions at the level of enhancers, promoters and isoforms. They also have the advantage of being smaller, and thus computationally more amenable, than the fine-grained circuits. TF-gene networks were defined as follows. First, we created a network encapsulating all regulatory interactions via promoters. For each pair of edges forming a chain that connects a TF to a promoter to an isoform—that is, edges (TF_i, $promoter_j$) and ($promoter_j$, $isoform_k$) with weights w_{ij} and w_{jk}—a corresponding edge (TF_i, $gene_l$) with weight $w_{il} = w_{ij}w_{jk}$ was added to the TF-gene network (where $isoform_k$ belongs to $gene_l$). If several redundant edges between the same TF and gene were found (via different promoters or isoforms), they were merged and the maximum edge weight was retained. A separate TF-gene network encapsulating all regulatory interactions via enhancers was created using the same approach. Both TF-gene networks thus had edge weights ranging from 0 (absent edge) to 1 (highest confidence), which were added to form a combined TF-gene network including evidence from both promoters and enhancers.

Pairwise similarity of regulatory networks was defined on the basis of the similarity of the two edge sets. For unweighted networks, the Jaccard index can be used for this purpose (the size of the intersection divided by the size of the union). Here we used an extension of the Jaccard index defined as

$$f(\mathbf{a},\mathbf{b}) = \frac{\mathbf{ab}}{|\mathbf{a}|^2 + |\mathbf{b}|^2 - \mathbf{ab}}$$

where **a** and **b** are the edge-weight vectors of the two networks (elements a_i and b_i giving the weights of corresponding edges (connecting the same TF and gene) in the two networks; if the edge is not present in one of the networks, the weight is zero). Note that for unweighted networks (i.e., where the edge weight is 0 or 1), this definition is equivalent to the Jaccard index. We used the difference function $1 - f$ for the hierarchical clustering of the 394 regulatory networks because we empirically found that it resulted in more homogeneous clusters than Euclidean distance, for instance. We also systematically tested clusters of networks defined by a given dendrogram cutoff for enriched sample annotations (**Supplementary Figs. 14–24**). To this end, FANTOM5 sample annotations from the Cell Ontology (CL), Anatomical Ontology (UBERON) and Disease Ontology (DOID) were propagated to the corresponding networks, and each cluster was tested for enriched terms (hypergeometric test; the false discovery rate (FDR) was controlled via the Benjamini-Hochberg procedure).

We formed the high-level networks corresponding to the 32 clusters defined in **Supplementary Figure 14** by taking the union of the individual networks within each cluster, always retaining the maximum edge weights (i.e., the same approach as described above for merging regulatory circuits of closely related samples). We used the same approach to create a global regulatory network by taking the union of all 394 cell type– and tissue-specific networks.

GWAS summary statistics and computation of gene-level *P* values. We obtained SNP-phenotype association summary statistics for a comprehensive collection of 37 GWASs including psychiatric[44–47], neurodegenerative[38,48,49], immune-related[50–54], cardiovascular[55,56], blood lipid[57,58], glycemic[59–64], anthropometric[65,66] and other traits[67,68] (**Supplementary Table 5**). Most of these studies are well-powered meta-analyses; the average sample size is over 50,000 individuals, with about one-third of the studies including over 100,000 individuals. The average number of SNPs is over 2.2 million.

Because we did not have access to genotype data for most studies, we used linkage disequilibrium (LD) information from a reference population (the European panel of the 1000 Genomes Project[69] (1KG)) to summarize SNP-association *P* values at the level of genes using the Pascal tool[30]. (We used the summary statistics as provided by the original studies; i.e., some studies are not 1KG-imputed, which is not critical to summarize the signal at the level of genes.) Briefly, SNPs in the vicinity of a gene were aggregated using either the maximum or the sum of χ^2 statistics, which measure the strongest and the average association signal per gene, respectively. (Multiple isoforms of the same gene were merged because, owing to LD, the resolution of a GWAS is typically too low to differentiate between individual isoforms of a gene.) Whereas previous methods relied on costly Monte Carlo simulations to estimate *P* values for these statistics, Pascal leverages analytic solutions, offering a dramatic increase in both speed and precision[30]. We applied Pascal to define gene-level *P* values for all 37 GWASs using default parameters. Results reported in this paper are based on the maximum of the χ^2 statistic; similar results were obtained using the sum of χ^2 statistics.

Network connectivity–enrichment analysis. Given a network and summary statistics from a GWAS, our aim is to evaluate

doi:10.1038/nmeth.3799

whether genes perturbed by trait-associated variants are more densely interconnected than expected. We refer to such groups of densely interconnected nodes as network modules. (Here we evaluated only the degree to which trait-associated genes cluster in modules; we did not identify discrete modules or pathways.) An overview of the four steps of the approach is given in **Supplementary Figure 27**.

The first step is to aggregate GWAS summary statistics at the level of genes, as described in the previous section.

The second step is to define how 'close' any two genes are in the network. The directionality of links is not considered for this purpose; for example, two genes that are coregulated by a TF gene may be considered close because they are connected through that TF, and two TF genes that regulate the same target gene could be considered close because they are connected through that target gene. Shortest path length is sometimes used to define 'closeness', but it is not very informative for biological networks because the shortest path between any two nodes is typically short owing to the presence of hubs. An alternative approach is to use diffusion kernels on graphs[70], with closeness (hereinafter referred to as connectivity) between two genes defined on the basis of the probability that a random walk on the graph will lead from one gene to the other. This approach naturally captures the hierarchical modular structure of biological networks (genes in the same module are more likely to be connected by a random walk). Here we used a weighted p-step random-walk kernel[71] to define the pairwise connectivity between nodes.

$$K = (I + \tilde{W})^p, \text{ with } \tilde{W} = D^{-1/2}WD^{-1/2}$$

where I is the identity matrix, W is the weighted adjacency matrix of the graph (entry w_{ij} is the weight of the edge between nodes i and j; entries on the diagonal are set to zero), D is the degree matrix of the graph (a diagonal matrix where entry d_{ii} is the degree of node i), and p is the number of steps in the random walk (we used $p = 4$ because biological networks are typically shallow and we expect few meaningful interactions over paths longer than four steps). Note that K can be computed cheaply because W is sparse. As mentioned above, the directionality of links was not considered for the purpose of defining pairwise connectivity; that is, W and K are both symmetric (thus, coregulated genes are connected through their shared regulators and tend to have high pairwise connectivity, particularly if they are part of a regulatory module with multiple shared regulators (i.e, many possible random walks connecting them)).

The third step is to compute connectivity-enrichment curves. To this end, we ranked genes by their GWAS P value from most to least significant. For each position n in the ranked list ($n = 1, 2, \ldots, N$, where N is the number of genes), the connectivity between the top n genes was defined as the mean of their pairwise connectivity values k_{ij} (gene pairs in LD were excluded; see below). Next, the connectivity between the top n genes was computed in the same way for 10,000 permutations of the ranked gene list. Importantly, only labels of genes with similar network centrality were permuted among one another (the centrality of a gene was defined as its mean pairwise connectivity with all other genes). Specifically, genes were separated into 100 bins according to their centrality, and only labels in the same bin were shuffled (a similar approach, within-degree gene label permutation, is commonly used in network-based GWAS analysis[11,72]; **Supplementary Fig. 30**). Finally, the connectivity-enrichment curve was computed as the ratio of the observed connectivity between the top n genes to the median connectivity between the top n genes across the 10,000 permutations, for each position n in the ranked list.

The fourth step is to summarize the connectivity-enrichment curves by the signed area under the curve. This is done both for the original data and the 10,000 permutations, enabling the computation of a corresponding empirical P value. Finally, the connectivity-enrichment score for the given GWAS and network is defined as the negative $\log_{10}$ of the empirical P value.

Note that the connectivity between neighboring genes on the genome was excluded from this analysis (**Supplementary Fig. 29**). This is important because the GWAS association signal of neighboring genes is often correlated because of LD, and neighboring genes are often also functionally related or coregulated (i.e., they may also be neighbors at the network level). To ensure that such groups of correlated and functionally related genes did not inflate connectivity enrichment, we took a conservative approach and completely excluded the connectivity between all neighboring genes (distance of <1 Mb) from the network connectivity–enrichment analysis (the corresponding entries in the connectivity matrix K were set to NA (not available), i.e., they were ignored in all calculations). Because the human leukocyte antigen (HLA) genes form an exceptionally large cluster of related genes in LD, which also show strong association with many immune-related traits, we further completely excluded all genes in the HLA region. Taken together, this ensured that the observed network connectivity enrichment was not driven by the HLA genes or similar gene clusters.

We used phenotype-label permutations to confirm that our method corrects for LD structure and other potential confounders (**Supplementary Fig. 30**).

Assessment of other network types. Aside from our cell type– and tissue-specific regulatory networks, we also assessed connectivity enrichment for five other types of networks. We considered only molecular networks derived from experimental data, not functional networks that were mined from the literature (e.g., co-citation networks).

First, we generated a global coexpression network from the FANTOM5 data[23] (the same data set that was used to construct the regulatory circuits). To this end, we defined the gene expression level as the sum of the corresponding promoter expression levels. TF-gene edge weights were computed using Spearman correlation, and the top 100,000 edges were retained.

Second, we obtained 35 tissue-specific coexpression networks from Piersen *et al.*[15]. These networks were inferred from GTEx data[28] using an algorithm that shares information between related tissues.

Third, we collected four well-established protein-protein interaction networks: (1) the InWeb protein interaction database developed by Lage *et al.*[73] and used by the popular DAPPLE tool[11], which probabilistically integrates evidence from diverse sources, including MINT, BIND, IntAct, KEGG annotated protein-protein interactions (PPrel), KEGG enzymes involved in neighboring steps (ECrel), Reactome and others; (2) Entrez GeneRIF (gene reference into function), which includes protein interactions from BIND, BioGRID, HPRD and other sources[74]; (3) the

BioGRID database, which provides literature-curated protein interactions[75]; and (4) the Human Interactome project, which is systematically screening the human proteome for interactions using yeast two-hybrid and other assays[76].

Fourth, we obtained a global regulatory network from Gerstein *et al.*[17]. The network is based on ChIP-seq data from ENCODE.

Fifth, we considered 41 tissue-specific regulatory networks based on DNaseI footprints from Neph *et al.*[77], but we did not find any significant connectivity enrichment. This is probably because these networks contain only TF-TF edges—that is, edges between TFs and target genes that are not TFs themselves were not included.

Identification of trait-relevant regulatory networks. A possible way to identify trait-relevant cell type– and tissue-specific networks would be to perform connectivity-enrichment analysis exhaustively for all 37 traits and 394 regulatory networks in our library (i.e., ~15,000 trait-network combinations). However, this is not practical because of the resulting multiple-testing burden and computation time (a single run with 10,000 permutations takes about 15 min, depending on the network size). Thus, we leveraged the fact that related networks are often very similar and exhaustively tested connectivity enrichment only for the 32 high-level networks across all traits. The six high-level networks specific to cancer cell lines were excluded for GWAS traits that are not cancer related. FDR was controlled using the Benjamini-Hochberg procedure. Subsequently, for each trait, we tested connectivity enrichment only for individual cell type– and tissue-specific networks belonging to high-level networks that had shown connectivity enrichment with a score of >1.0 for this trait (i.e., 10% FDR).

Availability of networks, data, tools and source code. We provide all networks, supplementary data and software tools on our website (http://regulatorycircuits.org). In addition, the networks and supplementary data have been deposited in an external repository for biomedical data (Synapse[78]), and code has been deposited in GitHub. Links to these resources are available on our website and in **Supplementary Table 6**.

40. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
41. Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* **17**, 1919–1931 (2007).
42. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
43. Marbach, D. *et al.* Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA* **107**, 6286–6291 (2010).
44. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
45. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371–1379 (2013).
46. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
47. Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol. Psychiatry* **19**, 1085–1094 (2014).
48. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
49. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).
50. International Multiple Sclerosis Genetics Consortium. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
51. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
52. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
53. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
54. Rauch, A. *et al.* Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. *Gastroenterology* **138**, 1338–1345 (2010).
55. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
56. The International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).
57. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
58. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
59. Prokopenko, I. *et al.* A central role for GRB10 in regulation of islet function in man. *PLoS Genet.* **10**, e1004235 (2014).
60. The DIAbetes Genetxzics Replication and Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
61. Scott, R.A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet.* **44**, 991–1005 (2012).
62. Soranzo, N. *et al.* Common variants at 10 genomic loci influence hemoglobin A1C levels via glycemic and nonglycemic pathways. *Diabetes* **59**, 3229–3239 (2010).
63. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).
64. Strawbridge, R.J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2. *Diabetes* **60**, 2624–2634 (2011).
65. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
66. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
67. The AMD Gene Consortium. Seven new loci associated with age-related macular degeneration. *Nat. Genet.* **45**, 433–439 (2013).
68. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44**, 491–501 (2012).
69. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).
70. Kondor, R.I. & Lafferty, J.D. Diffusion kernels on graphs and other discrete input spaces. in *Proc. Nineteenth International Conference on Machine Learning* 315–322 (Morgan Kaufmann, 2002).
71. Smola, A.J. & Kondor, R. Kernels and regularization on graphs. in *Learning Theory and Kernel Machines* (eds. Schölkopf, B. & Warmuth, M.K.) 144–158 (Springer Berlin Heidelberg, 2003).
72. Erten, S., Bebek, G., Ewing, R.M. & Koyutürk, M. DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Min.* **4**, 19 (2011).
73. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
74. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39**, D52–D57 (2011).
75. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43**, D470–D478 (2015).
76. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
77. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274–1286 (2012).
78. Derry, J.M.J. *et al.* Developing predictive molecular maps of human disease through community-based modeling. *Nat. Genet.* **44**, 127–130 (2012).

doi:10.1038/nmeth.3799