Selecting causal genes from genome-wide association studies via functionally coherent subnetworks

Murat Taşan[1–5,10], Gabriel Musso[6,7,10], Tong Hao[4,8], Marc Vidal[4,8], Calum A MacRae[6,7] & Frederick P Roth[1–5,9]

Genome-wide association (GWA) studies have linked thousands of loci to human diseases, but the causal genes and variants at these loci generally remain unknown. Although investigators typically focus on genes closest to the associated polymorphisms, the causal gene is often more distal. Reliance on published work to prioritize candidates is biased toward well-characterized genes. We describe a 'prix fixe' strategy and software that uses genome-scale shared-function networks to identify sets of mutually functionally related genes spanning multiple GWA loci. Using associations from ~100 GWA studies covering ten cancer types, our approach outperformed the common alternative strategy in ranking known cancer genes. As more GWA loci are discovered, the strategy will have increased power to elucidate the causes of human disease.

Although simple (i.e., Mendelian) traits can be explained by a few strong-effect loci, the modest effects at many complex trait loci complicate the precise identification of causal variants[1]. GWA studies in large cohorts help address this issue by being powered to detect modest associations at multiple loci simultaneously[2]. GWA studies have to date detected thousands of robust associations between genomic loci and disease-related traits. However, rather than identifying causal genes or variants directly, these associations generally identify 'tag' single-nucleotide polymorphisms (tagSNPs), each representing many linked variants. Moving from these genomic landmarks to individual causal genes within these loci remains challenging, and precise understanding of the genotype-to-phenotype relationship for most traits remains elusive[3].

To address this gap, orthogonal evidence can help prioritize candidate genes at disease-associated loci[3,4]. Co-occurrence of gene names within PubMed abstracts, for example, has been used to identify connections between candidate genes at implicated loci[5]. However, many genes are poorly characterized in the literature, and restricting analyses to 'popular' genes diminishes the opportunity for novel discovery of gene-disease associations. Likewise, protein-protein interactions have informed our mechanistic understanding of disease[6–8], but interaction evidence alone is limited in scope; much of the human proteome is underrepresented in high-quality databases[9] (**Supplementary Fig. 1**), and only a small fraction of the complete interactome has been mapped[10]. Additionally, nearly half of all current human protein-protein interaction knowledge comes from small-scale targeted studies, which, like literature text-mining, limits the opportunity for new discovery[11].

Gene-function annotations (for example, pathway membership) can help identify causal genes within disease-associated loci. For example, groupwise disease associations can be sought for sets of single-nucleotide polymorphisms (SNPs) mapping to a given functional category[7,12]. Assigning SNPs to functional sets, however, requires (i) existing assignments of SNP effects to specific genes and (ii) complete knowledge of function—both of which remain problematic[13].

Shared-function or 'cofunction' networks (CFNs) augment curated functional annotations by connecting pairs of genes that share, or are likely to share, biological function[14] (for example, by sharing protein domain annotations). Guilt-by-association methods[15] have used CFNs to assign function to uncharacterized genes for *Saccharomyces cerevisiae*[14], *Arabidopsis thaliana*[16], *Mus musculus*[17] and *Homo sapiens*[18–20], among other species. CFNs have also contributed to fine-scale mapping of Mendelian disorder associations[21] and can prioritize genes not located at disease-associated loci (for example, by connectivity to known 'seed' causal genes[8,22]).

Here we used CFNs to prioritize groups of candidate genes from multiple disease-associated loci on the basis of mutual functional relatedness. We framed the problem as a constrained optimization task, analogous to choosing mutually compatible items from a prix fixe restaurant menu, with one dish from each course (cocktail, appetizer, entree, dessert, etc.). Combinations of genes, with one gene from each locus, were evaluated for their collective extent of shared function within the CFN. Although each solution is initially constrained to a single gene per locus, analysis of many top solutions can point to multiple strong candidates within a locus. We found that the prix fixe strategy improves upon the ubiquitous approach of ranking candidate causal genes

[1]Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [3]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [4]Center for Cancer Systems Biology (CCSB), Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [5]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [6]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA. [7]Cardiovascular Division, Brigham and Women's Hospital, Boston, Massachusetts, USA. [8]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [9]Canadian Institute for Advanced Research, Toronto, Ontario, Canada. [10]These authors contributed equally to this work. Correspondence should be addressed to M.T. (murat@tasanlab.org) or F.P.R. (fritz.roth@utoronto.ca).

RECEIVED 1 MAY; ACCEPTED 24 NOVEMBER; PUBLISHED ONLINE 22 DECEMBER 2014; DOI:10.1038/NMETH.3215

by their genetic distance to trait-associated tagSNPs. Mutually connected gene groups can reveal disease-relevant pathways and prioritize candidate disease genes. This method is freely available online at http://llama.mshri.on.ca/prixfixe and as an R package (**Supplementary Software**).

RESULTS

Candidate genes within GWA loci are often considered only if they overlap or flank reported tagSNPs, excluding other potentially causal nearby genes (see, for example, the "mapped genes" field in the US National Human Genome Research Institute (NHGRI) GWA studies catalog[23]). Moreover, candidate causal genes are typically examined in the context of existing literature, which may be subject to substantial confirmation bias. For example, the rate of new publications is substantially higher for earlier-characterized genes than for genes discovered more recently (**Supplementary Fig. 2**). This 'rich get richer' phenomenon lures us from novel discoveries toward already well-characterized genes.

To prioritize candidate genes from disease-associated loci while minimizing bias toward well-studied genes, we integrated genome-scale data and analyzed published GWA studies spanning 23 diverse complex diseases and traits, including autoimmune disorders, cognition levels, cardiovascular and metabolic traits, and ten distinct cancer types (**Table 1** and **Supplementary Tables 1–4**).

After tagSNPs associated with a given trait are identified, nearby genes found by linkage disequilibrium (LD) are consolidated into disjoint gene sets (**Fig. 1**). A stochastic optimization strategy then identifies 'prix fixe menu selections': sets of genes (one gene per locus) that correspond to dense subnetworks of functional relationships. Finally, we measure each gene's contribution to the top-scoring subnetworks. These dense subnetworks yield sets of genes working in concert and highlight particular processes that may contribute to disease etiology. Although we performed this analysis for all 23 traits, below we highlight prostate cancer susceptibility as a case study.

Defining locus boundaries

We systematically defined genomic boundaries for trait-associated loci using pairwise LD correlations (r^2) between each associated tagSNP and nearby SNPs (**Fig. 1**). Genes (defined to capture *cis*-regulatory elements via up- and downstream 'padding') within these boundaries were then identified (Online Methods).

Inferring a human cofunction network

To aggregate information about functional relationships between human genes, we constructed a CFN covering most of the human genome (**Fig. 1** and **Supplementary Fig. 3**). For this, we used (i) a method based on gene-pair features (for example, shared protein domain signatures)[19] and (ii) a graph-walking strategy to find pairs of genes that are likely to share function[24] (Online Methods). The two networks cover nearly the full human genome but are complementary (**Supplementary Fig. 3**), echoing earlier findings that different methods often excel at inferring different functions[17]. We merged these two networks into a single CFN, providing ~10^7 cofunction links involving ~19,000 genes (covering 94% of the human genome).

Identifying mutually connected subnetworks

To connect trait-associated loci, we searched the CFN for groups of candidate genes that appeared to work in concert (**Fig. 1**).

Table 1 | Traits and diseases analyzed in this work

Trait group	Trait	# pubs	# loci	# PF combinations
Cancers	ALL	7	44	2.4×10^8
	Breast cancer	1	58	7.5×10^{12}
	CLL	5	23	1.8×10^7
	Colorectal cancer	13	36	1.7×10^{10}
	Gastric cancer	3	7	1.7×10^3
	Glioma	4	8	2.6×10^2
	Lung cancer	18	29	2.6×10^{12}
	Ovarian cancer	4	11	8.7×10^4
	Pancreatic cancer	4	31	1.9×10^7
	Prostate cancer	19	73	7.8×10^{28}
Autoimmune diseases	Type 1 diabetes	1	38	9.9×10^{19}
	Multiple sclerosis	1	75	5.7×10^{35}
	Crohn's disease	1	70	6.7×10^{34}
	Ulcerative colitis	1	47	1.1×10^{21}
	IBD	1	110	1.3×10^{53}
Cardiovascular traits	Cholesterol, total	1	52	1.4×10^{24}
	Cholesterol, HDL	1	47	7.3×10^{19}
	Cholesterol, LDL	1	37	2.0×10^{18}
	Triglycerides	1	32	6.1×10^{15}
	QT interval	1	27	7.3×10^5
Metabolic traits	Height	1	183	1.7×10^{73}
	Type 2 diabetes	1	26	2.2×10^8
Cognition	Cognitive performance (1)	1	57	5.7×10^{12}
	Cognitive performance (2)	1	53	3.2×10^9

"# pubs" is the number of distinct GWA studies publications used in this study. "# loci" gives the number of loci found after mapping the associated SNPs to nonredundant genomic windows. "# PF combinations" is the number of unique prix fixe subnetworks that can be derived from the associated loci and their constituent genes. ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia; IBD, irritable bowel disease. Full details of all GWA studies (including all associated SNPs) are available in **Supplementary Tables 1** and **4**.

Specifically, we sought densely connected subnetworks such that each locus contributes a single gene to the subnetwork (thereby implementing a prix fixe constraint). In graph-theoretic terms, these are dense L-partite subgraphs for L loci, where density is a measure of mutual connectivity among the genes.

For most complex traits, the large number of associated loci and candidate genes (**Table 1**) make enumeration of all possible prix fixe subnetworks infeasible. The 73 loci implicated in prostate cancer, for example, have ~10^{29} potential prix fixe subnetworks, and for height this number exceeds 10^{73}. To circumvent this, we used a genetic algorithm[25] seeded with a 'population' of random prix fixe gene sets ('individuals'), each subjected to 'mutation'—a low-probability swap of two genes at a given locus. Each subnetwork individual was evaluated for fitness (here, edge density), and pairs of individuals were randomly mated (preferring fitter pairs) to create new subnetworks (Online Methods). After repeated generations of selection, the population was enriched for dense prix fixe subnetworks (**Fig. 1** and **Supplementary Fig. 4**). To measure significance, we compared the final population's average edge density to the same measure from 1,000 trials with random input sets matching the true input set in terms of number of genes and connectivity (**Supplementary Fig. 4** and Online Methods).

The importance of each gene at each locus was estimated by the difference in edge densities in subnetworks with and without that gene. For example, a gene with no connections yields the same density whether included or not, implying zero importance to that subnetwork. We averaged the importance measurements of each gene over the final fittest population of prix fixe subnetworks, obtaining a prix fixe score (**Fig. 1**) for each candidate gene (Online Methods).

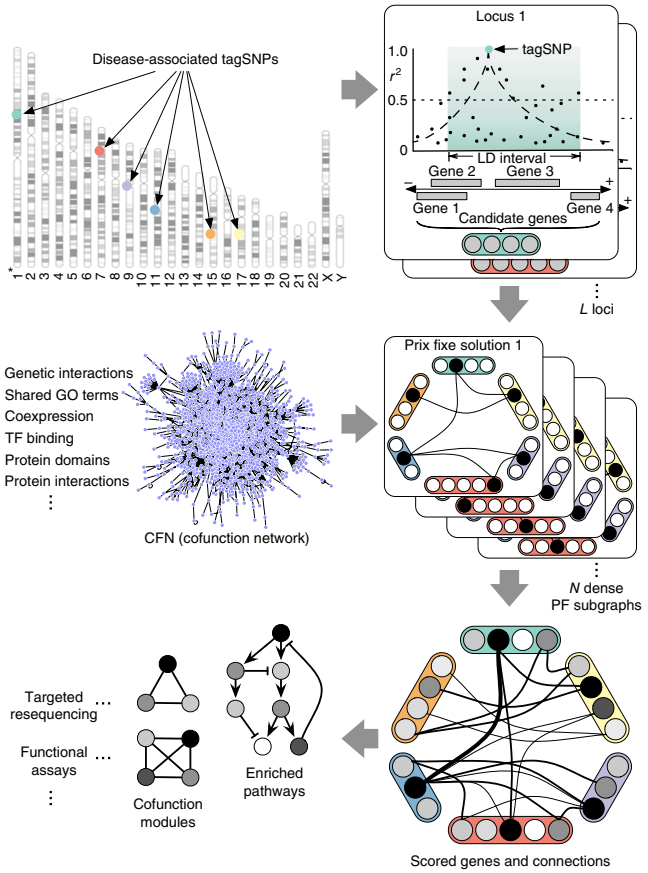

Figure 1 | Overview of the prix fixe strategy. tagSNPs associated with disease are used to define linkage-disequilibrium (LD) windows. A CFN is then used to identify dense 'prix fixe' (PF) subnetworks. Dense PF subnetworks are aggregated, and genes are scored to reflect their importance in the subnetworks. High-scoring genes are then used to find causal pathways, processes and additional candidate genes.

Both the gene scores and the frequencies with which edges appear in the subnetworks provide clues about how candidate genes work together. Among the 73 prostate cancer–associated loci, for example, at locus 6p21.33 the candidate gene *POU5F1* (also known as *OCT4*) was highlighted, along with one of its frequent subnetwork partners, *HNF1B* (at locus 17q12; **Fig. 2**). Despite being previously linked to prostate cancer[26], *POU5F1* might have otherwise been overlooked given that four other genes are closer to the associated tagSNP. But both *POU5F1* and *HNF1B* have important roles in embryonic development and boost each other's importance. *HNF1B* has also recently been shown to modulate the effects of growth hormones and tumor progression[27].

Cancer-susceptibility gene prioritization

To broadly evaluate prix fixe–based gene prioritization, we analyzed 78 published GWA studies spanning ten types of cancer (**Table 1**). For nine types (all but chronic lymphocytic leukemia (CLL)), at least one associated multigenic locus contained a known cancer-linked gene, as defined by the Sanger cancer gene census (SCGC)[28].

Prioritization success was measured at multigenic loci by ranking the SCGC gene by its prix fixe score within each locus and rescaling this rank from 0% to 100% (Online Methods). The prix fixe score successfully identified the SCGC gene as the highest-ranked gene (100% relative rank) for 21 out of 34 loci, with an average relative rank of 80% for SCGC genes (**Fig. 3**), which is significantly higher than expected for non-informative random rankings ($P = 4.9 \times 10^{-7}$, one-sided one-sample Student's *t*-test). The prix fixe approach also outperformed the common alternative LD-based closest-gene strategy of ranking genes by tagSNP proximity (average relative rank, 58%; $P = 0.015$, one-sided paired Wilcoxon signed-rank test; **Fig. 3** and Online Methods).

Note that a gold-standard set of cancer genes would include only genes for which germline susceptibility alleles have been observed, given that cancer susceptibility is the GWA trait under study. This more stringent reference standard yielded a similar effect size (average relative rank of 91%), but with only eight qualifying loci, it had insufficient statistical power ($P = 0.14$). However, there is strong overlap between somatically mutated cancer genes and those associated with germline susceptibility, with half (43/81) of germline SCGC genes showing evidence of somatic mutation. The high gene rankings within the more complete SCGC set suggest that many of the cancer-linked genes at these 34 loci previously known only through somatic mutations may also harbor germline predisposition alleles.

To further investigate our rankings, we used mRNA expression data from The Cancer Genome Atlas (TCGA) for both breast (BRCA) and prostate (PRAD) cancer (which were the only cancer types with sufficient matched tumor-versus-normal RNA-seq data at the time of this study). The prix fixe scoring method ranked differentially expressed genes significantly higher than genes without a marked expression difference between matched tumor and healthy tissues for both cancers ($P = 0.03$ for PRAD and $P = 0.01$ for BRCA; Wilcoxon rank-sum test; **Fig. 4a**, **Supplementary Fig. 5** and Online Methods). Closest-gene rankings did not show correlation with cancer-dependent expression ($P = 0.17$ and $P = 0.59$; Wilcoxon rank-sum test; **Fig. 4a** and **Supplementary Fig. 5**).

Identifying causal pathways

Commonalities among high-scoring candidate genes can provide insight into the processes contributing to disease (**Fig. 1**), and so for each trait, we searched for Gene Ontology (GO) terms that were over-represented among the highest-scoring genes[29] (Online Methods). Prix fixe–ranked prostate cancer candidate genes yielded significant enrichment for 163 GO terms (permutation tests with multiple-testing FWER < 0.05; **Supplementary Table 3** and Online Methods). The maximal enrichment for most (75%) of these terms was found using just the top 23 genes, indicating a high concentration of shared function between these highest-scoring candidates. By contrast, functional enrichment analysis with the complete set of genes from prostate cancer–associated loci (i.e., an unordered search) yielded no enriched terms. More surprisingly, no terms were found in an ordered functional enrichment analysis of prostate cancer genes when ranked by the closest-gene approach. For all traits examined, prix fixe scoring provided more enriched terms than the closest-gene approach, with the latter method providing nearly the same amount of term enrichment as the unranked approach (**Supplementary Table 3**).

Many enriched terms in our prostate cancer analysis have clear links to prostate function and development, including "androgen receptor activity," "male genitalia morphogenesis" and "prostate gland morphogenesis" (**Fig. 4b**). The high-scoring candidate genes

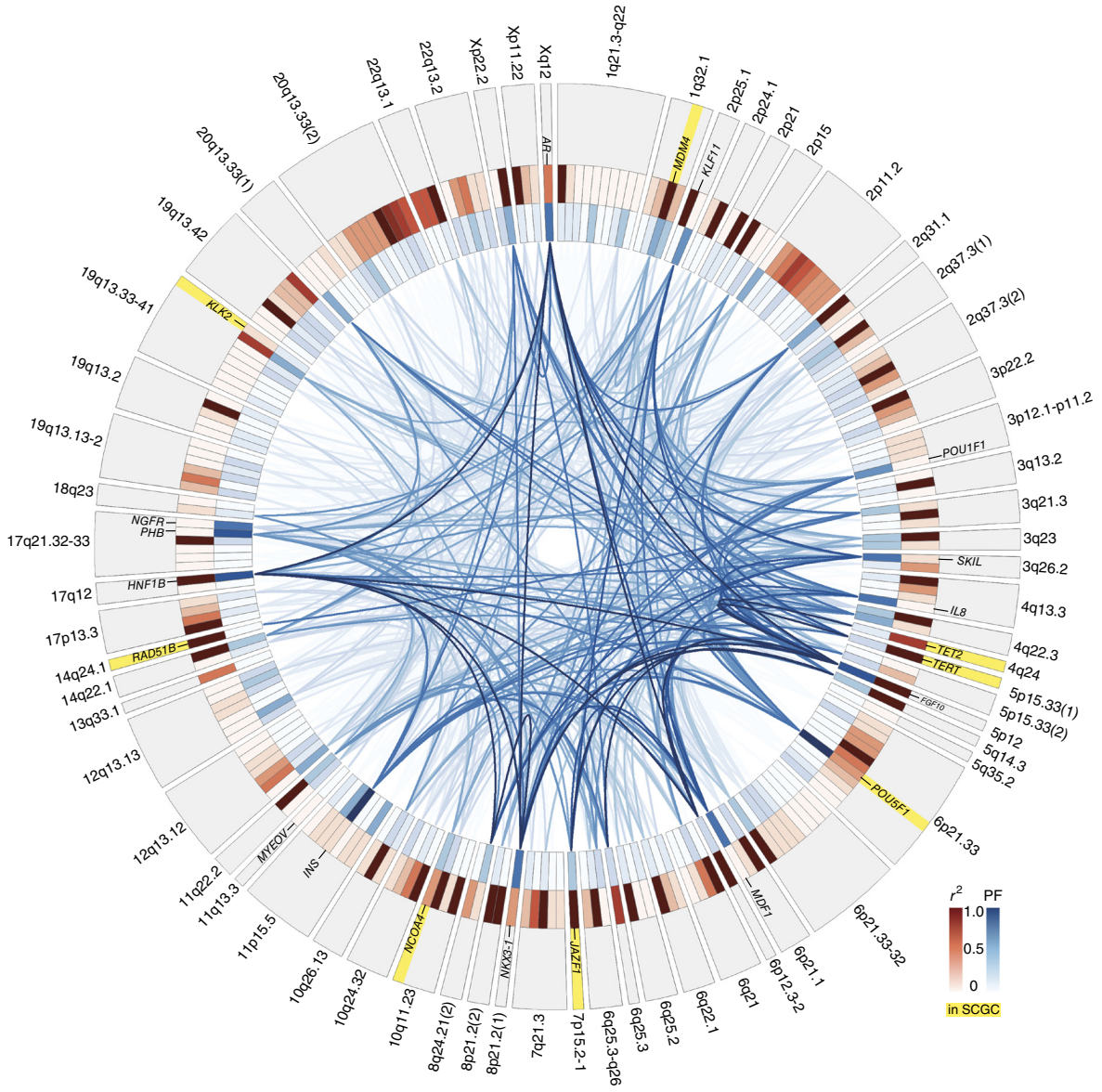

Figure 2 | Functional connectivity patterns in prostate cancer. Candidate genes are organized by locus in genomic order. Genes highlighted in yellow are members of the Sanger cancer gene census (SCGC). Red intensity indicates the LD (r^2) value between the gene and tagSNP for that locus. Blue intensity reflects the prix fixe (PF) score. Edges represent presence in the final collection of dense subnetworks, with blue edge intensity reflecting the proportion of final dense subnetworks containing that edge. Only select gene names are shown here owing to space constraints. A full-sized plot with all gene names is available on the supporting website (http://llama.mshri.on.ca/prixfixe).

AR (androgen receptor), *FGF10* (fibroblast growth factor 10), and *NKX3-1* (NK3 homeobox 1)—found at Xq12, 5p12 and 8p21, respectively—have all received considerable attention for their probable role in prostate cancer development despite not yet being declared causal in the SCGC[30–32]. In particular, *AR* and receptors of the *FGF* family (i.e., FGFRs) are considered possible therapeutic targets for inhibiting prostate tumor progression[33] and tumorigenesis more generally[34]. *AR* was among the few significantly mutated genes (SMGs) detected by a recent pan-cancer whole-genome sequencing analysis of diverse tumor samples[35] (**Fig. 4b**).

Enriched terms provide the opportunity to identify new candidate genes outside trait-associated loci. "Prostate gland morphogenesis," for example, was associated with 30 genes not found within any prostate cancer–associated locus (**Fig. 4b**). Four of these genes are known to have causal roles in tumorigenesis (*FGFR2*, *NOTCH1*, *HOXD13* and *HOXA13*), and four (in addition to *AR*) appear in the pan-cancer SMG list[35] (**Fig. 4b**). Thus, the prix fixe method systematically prioritizes candidate genes that can then serve as seeds to find additional candidates (for example, through guilt-by-association techniques).

For nearly every trait examined, we found enriched terms that were highly relevant to that trait (**Supplementary Table 3**): for example, "learning or memory" for cognitive performance and "plasma lipoprotein particle assembly" for cholesterol levels. The processes highlighted by our candidate gene rankings also helped identify non-obvious or environmental factors contributing to complex traits. For example, top prix fixe–ranked lung cancer candidate genes were highly enriched for associations with "behavioral response to nicotine," capturing the role of smoking in lung cancer and possible gene-environment interactions.

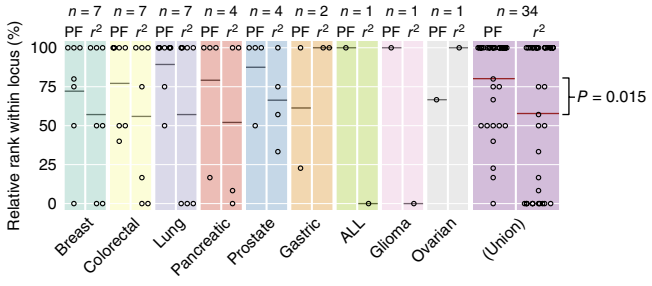

Figure 3 | Rank-based analysis of SCGC prioritization. Genes are ranked within each cancer-associated locus, and normalized ranks of SCGC genes are shown as dots for prix fixe–based (PF) and LD-based (r^2) rankings (100% is highest ranked; 0% is lowest). The average relative rank of SCGC genes (for both methods) within each locus is identified by horizontal bars; the number of multigenic loci is shown above as *n*. The rightmost plot ("Union") shows pooled results across all cancer-associated loci. PF SCGC ranks significantly outperform LD-based SCGC ranks ($P = 0.015$, one-sided paired Wilcoxon signed-rank test). ALL, acute lymphoblastic leukemia. Chronic lymphocytic leukemia contained no SCGC-harboring loci in our primary analysis and is thus not displayed here.

Finally, we found terms that were commonly enriched across a subset of traits, indicating diseases with shared etiology. High-scoring genes in CLL, type 1 diabetes, Crohn's disease, ulcerative colitis, inflammatory bowel disease and multiple sclerosis, for example, were all associated with functions of immunity. More generally, "response to stress" was over-represented for nearly half of the traits examined in this work, underscoring commonalities of diverse diseases and disorders. Complete results for all traits can be found in **Supplementary Table 3**.

DISCUSSION

Genes contributing to the same trait often share functional relationships[36]. Here we have exploited this phenomenon to prioritize candidate causal genes without specifying a priori which functions contribute to the phenotype. We found limitations in the naïve (but commonly used) closest-gene approach, which provided almost no advantage over ranking genes within loci uniformly at random. The extensive haplotype block structures found in human populations limit the utility of the closest-gene strategy. Furthermore, the use of CFNs built from genome-scale data permits scoring for nearly all candidate genes in implicated loci, reducing the knowledge bias that is coupled with literature-mining approaches.

The importance-scoring step of the prix fixe strategy provides flexibility when aggregating results across many dense prix fixe subnetworks. As not all loci are multigenic, this scoring method can measure the contributions of genes even at monogenic loci. Those genes with strong connections to other candidate genes achieve high scores (for example, *AR* at Xq12; **Fig. 2**), whereas the weakly connected genes tend to score poorly (for example, *MYEOV* at 11q13.3; **Fig. 2**). The use of multiple top-scoring subnetworks followed by importance scoring also allows for similarly connected genes within a multigenic locus to obtain similar scores. For example, *NGFR* and *PHB* at 17q21.32-33 are both strong prostate cancer candidates (**Fig. 2**), and selecting one at the expense of the other by selecting only a single top-scoring subnetwork might have conferred false confidence in a single recommended gene. Individual SNP effect sizes may in the future be included to augment network-based prioritization methods (for example, by placing prior probability weights on candidate genes[37]); however, such analyses at large scale will require a (currently unavailable) catalog of annotated effect sizes for markers across all tested traits.

To better understand mechanisms underlying a given phenotype, researchers must view candidate genes in the context of biological processes and pathways[3]. Ranking candidate gene sets by their level of collective cooperation within the cell is a principled way to simultaneously identify causal genes and explanatory causal pathways. In addition to those enriched functional annotations found for each trait, the enriched functions shared by different traits point to shared etiologies that might underlie

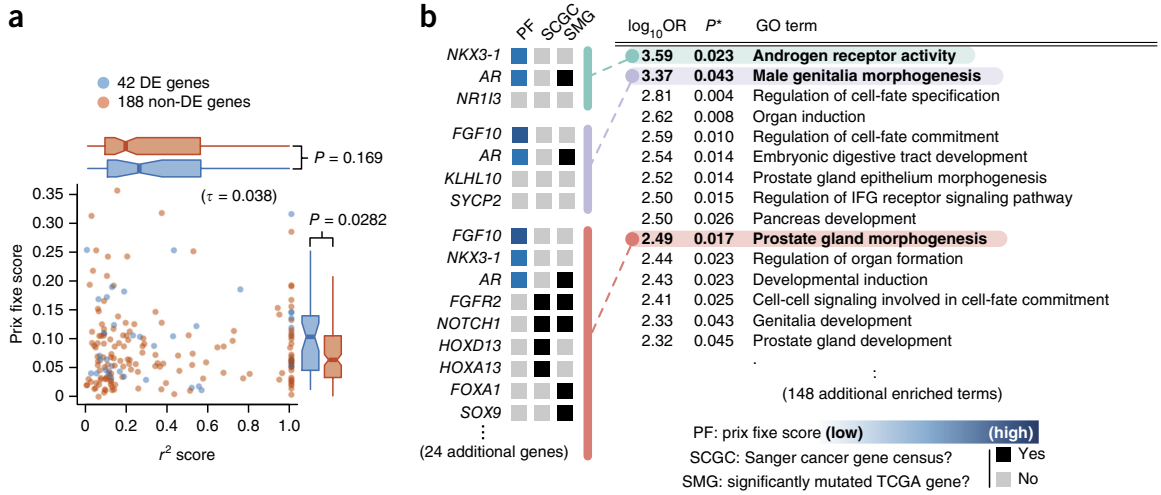

Figure 4 | Prix fixe gene-score distribution and functional enrichment. (**a**) Prix fixe (PF) scores are uncorrelated with LD (r^2) values. Each scatter plot point is a candidate breast cancer gene. Correlation is computed using Kendall's τ rank coefficient. Blue indicates significantly differentially expressed (DE) mRNA levels in matched case-control TCGA prostate adenocarcinoma (PRAD) sample; red indicates no evidence of cancer-dependent differential expression. Flanking box plots indicate score distributions of DE and non-DE genes. Box-plot whiskers extend to 1.5× the interquartile range; outliers not shown. Box plots were compared by one-sided Wilcoxon rank-sum tests. (**b**) PF rankings identify disease-relevant Gene Ontology (GO) terms for prostate cancer, with no a priori knowledge of disease etiology. The top 15 (by odds-ratio (OR)) GO terms by ordered functional enrichment analysis with significance (P^*) corrected for multiple testing[29] are shown. Three GO terms are expanded to show constituent genes with (if available) PF score, SCGC status and SMG[35] status. Full functional enrichment analysis for all traits is provided in **Supplementary Table 3**.

comorbidity patterns[38] and may help identify therapies for one disease that might be repurposed for another.

Using CFNs and connectivity measures to prioritize large candidate lists can be extended beyond GWA studies. It could also be applied, for example, to candidate disease-related variants found by sequencing-based mutational burden studies. Incorporating prior functional knowledge about these candidates will help to prioritize subsets of genes, possibly even in mutually exclusive combinations[39]. Resulting gene sets can then be fed back to GWA prioritization results, increasing our power to identify the underlying causal pathways. The inclusion of large-effect rare variants may help solve the 'missing heritability' problem[40].

Thus, the use of unbiased genomic data sets and a prix fixe–constrained optimization procedure can identify mutual functional similarity among genes in trait-associated loci to prioritize loci, genes and trait-associated pathways.

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 members of the Roth lab and the Center for Cancer Systems Biology (CCSB) at the Dana-Farber Cancer Institute (DFCI) for helpful comments and discussion; the lab of Q. Morris for assistance with GeneMANIA data; and M. Çokol and J. Mellor for useful conversations and advice during manuscript preparation. This work was primarily supported by Center of Excellence in Genomic Science (CEGS) grant P50 (HG004233) from the NHGRI awarded to M.V. and F.P.R. F.P.R. is additionally supported by US National Institutes of Health (NIH) grants (HG003224 and HL107440), the Krembil and Avon Foundations, a Canadian Ontario Research Fund Research Excellence Award and the Canada Excellence Research Chairs Program. C.A.M. was supported in this work by an NIH grant (HL098938), the Leducq Foundation and the Harvard Stem Cell Institute. M.T. was supported by an NIH grant (HG004098).

AUTHOR CONTRIBUTIONS

M.T., G.M., C.A.M. and F.P.R. conceived of the project. M.T., G.M. and T.H. performed computational analyses. M.T., G.M., C.A.M. and F.P.R. wrote the manuscript. M.V., C.A.M. and F.P.R. oversaw and guided the research effort.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**, 695–701 (2008).
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
3. Chakravarti, A., Clark, A.G. & Mootha, V.K. Distilling pathophysiology from complex disease genetics. *Cell* **155**, 21–26 (2013).
4. Gilman, S.R. *et al.* Rare *de novo* variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70**, 898–907 (2011).
5. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
6. 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).
7. Han, S. *et al.* Integrating GWASs and human protein interaction networks identifies a gene subnetwork underlying alcohol dependence. *Am. J. Hum. Genet.* **93**, 1027–1034 (2013).
8. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.* **6**, e1000641 (2010).
9. Das, J. & Yu, H. HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).
10. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6**, 83–90 (2009).
11. Rolland, T. *et al.* A Proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
12. Hirschhorn, J.N. Genomewide association studies—illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).
13. Cantor, R.M., Lange, K. & Sinsheimer, J.S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).
14. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
15. Wang, P.I. & Marcotte, E.M. It's the machine that matters: predicting gene function and phenotype from protein networks. *J. Proteomics* **73**, 2277–2289 (2010).
16. 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).
17. Peña-Castillo, L. *et al.* A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. *Genome Biol.* **9** (suppl. 1), S2 (2008).
18. Mostafavi, S. & Morris, Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* **26**, 1759–1765 (2010).
19. Taşan, M. *et al.* A resource of quantitative functional annotation for *Homo sapiens* genes. *G3 (Bethesda)* **2**, 223–233 (2012).
20. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19**, 1093–1106 (2009).
21. Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* **78**, 1011–1025 (2006).
22. 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).
23. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
24. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).
25. Goldberg, D.E. *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, 1989).
26. de Resende, M.F. *et al.* Prognostication of OCT4 isoform expression in prostate cancer. *Tumour Biol.* **34**, 2665–2673 (2013).
27. Hu, Y.L. *et al. HNF1b* is involved in prostate cancer risk via modulating androgenic hormone effects and coordination with other genes. *Genet. Mol. Res.* **12**, 1327–1335 (2013).
28. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
29. Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. & Roth, F.P. Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).
30. Memarzadeh, S. *et al.* Enhanced paracrine FGF10 expression promotes formation of multifocal prostate adenocarcinoma and an increase in epithelial androgen receptor. *Cancer Cell* **12**, 572–585 (2007).
31. Heinlein, C.A. & Chang, C. Androgen receptor in prostate cancer. *Endocr. Rev.* **25**, 276–308 (2004).
32. Bhatia-Gaur, R. *et al.* Roles for Nkx3.1 in prostate development and cancer. *Genes Dev.* **13**, 966–977 (1999).
33. Gao, W. Androgen receptor as a therapeutic target. *Adv. Drug Deliv. Rev.* **62**, 1277–1284 (2010).
34. Katoh, M. & Nakagama, H. FGF receptors: cancer biology and therapeutics. *Med. Res. Rev.* **34**, 280–300 (2014).
35. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
36. King, O.D. *et al.* Predicting phenotype from patterns of annotation. *Bioinformatics* **19** (suppl. 1), i183–i189 (2003).
37. Liu, J.Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
38. Lee, D.-S. *et al.* The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. USA* **105**, 9880–9885 (2008).
39. Vandin, F., Upfal, E. & Raphael, B.J. *De novo* discovery of mutated driver pathways in cancer. *Genome Res.* **22**, 375–385 (2012).
40. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

ONLINE METHODS

Cofunction network derivation. We derived a human cofunction network (CFN) from two existing CFN resources and published methods. The first CFN was constructed as described in Taşan *et al.*[19] but with the exclusion of Online Mendelian Inheritance of Man (OMIM)[41] data. OMIM data were removed specifically for this study to limit any potential source of circular logic while we evaluated our methods. The remaining predictive data types are briefly summarized below, each chosen with the intent of being as free of survey bias as possible such that combinations of their features retained low bias while providing increased power for discovery.

Protein domain signatures for all genes were downloaded from InterPro[42] and represented as a binary matrix (i.e., presence or absence of each signature for each gene), and scores were computed for each gene pair using the PhenoBlast method[43]. Transcription factor binding site (TFBS) information was acquired as UCSC Genome Browser[44] hg19 tracks for TRANSFAC and ENCODE ChIP-seq data. To assign TFBSs to genes, we defined gene boundaries by expanding RefSeq transcripts (also mapped to UCSC hg19 coordinates) upstream by 5,000 bp and downstream by 500 bp, and any TFBS overlapping a gene was then assigned to that gene. A single binary matrix was created for all TFBS data and all genes, and similarity between gene pairs was scored using the PhenoBlast method[43]. Similarity between phylogenetic profiles (downloaded from Inparanoid[45]) were also scored using the PhenoBlast method[43]. Normalized and summarized gene expression profiles covering normal human tissues were downloaded from BioGPS[46]. These expression data were then log-transformed, and Kendall rank correlation coefficients were computed for each gene pair. Finally, a catalog of literature-curated protein-protein interactions between human ORFs was separated into "binary" and "all" interactions, creating two features (where binary interactions must come from experiments specifically testing pairs of proteins, whereas the complete data set includes interactions derived from co-complex methods, such as affinity purification and mass spectrometry (AP-MS) experiments).

As positive training examples of gene pairs sharing function, we used gene pairs sharing Gene Ontology (GO) Biological Process (BP) terms. To ensure specificity in our definition of cofunction, we limited the terms used to those with fewer than 300 non-electronic (i.e., excluding RCA and IEA GO evidence codes) gene associations. These data were then used to train a Random Forest ensemble classifier[47], and the top 1% of scored gene-pairs were used as our predicted CFN. Note that gene-pair scores were 'out of bag' in that the random forest used to score each gene pair excluded any tree that made use of that gene pair.

The second CFN we used was generated using a different prediction strategy also shown to produce high-quality inferences of shared function between genes using a label-propagation method[18]. Prescored data were downloaded from GeneMANIA[24], and as disease annotations were not included as a source data set, we performed no additional pruning of these data.

Both strategies have been demonstrated to provide high-quality gene-function predictions for (amongst others) *H. sapiens*[19,24], *M. musculus*[17,48,49], *D. rerio*[50] and *S. cerevisiae*[51]. The union of these two CFNs were then taken as the single CFN used for this study, noting that although gene coverage overlap was high between the two networks, the gene-pair predictions were largely complementary (**Supplementary Fig. 3**).

Gene, SNP and LD positional data. All gene definitions used in this study were acquired from the NCBI Gene database. Transcripts corresponding to these genes were mapped to UCSC hg19 coordinates[44]. Variation data from dbSNP 137 were also mapped to UCSC hg19 coordinates, and linkage-disequilibrium (LD) data for SNP pairs within 500 kbp of each other were downloaded from the International HapMap Project (Phase III, CEU population)[52].

GWA study data and gene-set construction. All GWA study data used in this work were acquired from the NHGRI GWAS catalog[23]. As some publications report on associations to multiple distinct traits, we took each publication-trait pair and treated it as a distinct 'study'. Studies were then ranked by their number of significantly associated loci, and we chose to focus on complex and/or heterogeneous traits, generally with at least 20 reported loci per study. For our cancer analyses, we preferentially selected recent meta-analyses where available but otherwise took the union of reported SNPs for studies addressing the same type of cancer. For our noncancer traits, we treated each study independently. Many of the traits we analyzed were associated with more than 20 loci each, indicating substantial complexity in the underlying biology (**Table 1**). Prostate cancer, for example, has been associated with 73 loci, whereas height has been associated with nearly 200 loci[53].

We then processed each analyzed set of associated SNPs by first finding all other SNPs in LD with the associated markers. To this end, a genomic window was defined by taking the positions of the physically farthest upstream and downstream SNPs in LD, such that $r^2 \geq 0.5$ between each boundary SNP and the associated SNP. Genes were defined by RefSeq transcript boundaries, but extended 100 kbp upstream and 10 kbp downstream to include *cis*-regulatory regions. Overlapping windows (which may occur owing to multiple SNPs in close proximity being reported for the same locus) were merged to create a set of disjoint genomic windows. The transcripts within these windows were mapped back to unique NCBI Gene IDs, creating a disjoint collection of gene sets. All PubMed IDs, dbSNP IDs, window coordinates and candidate genes are available in **Supplementary Table 1**.

LD decay score. For each trait-associated locus, we derived a score for each candidate gene based solely on that locus's local LD properties (the r^2 score). In cases where the locus was defined by a single tagSNP, we used that SNP for the procedure below. When a locus had been identified by multiple tagSNPs (leading to locus merging, as described above), the SNP with the strongest reported effect size was chosen as the representative SNP for that locus. (In cases where no effect size was available, the SNP with the smallest reported *P* value was chosen.)

An LD decay model for each locus was then learned using the r^2 correlations between the representative SNP and all other in-LD SNPs in the locus. The decay was modeled using beta regression[54] with an inverse link function of $r^2 = 1/(1 + x)$, where x is the distance (in bp) between the two SNPs. This follows the theoretical relationship between LD and genetic distance described as

$r^2 = 1/(1 + 4N_e c)$, where N_e is the effective population size and c is the recombination fraction between the two loci[55,56].

Each transcript in the locus was then given an r^2 score according to this model, where the r^2 decay value was computed for the point along the transcript closest to the representative SNP (i.e., the maximal predicted r^2 value along the length of the transcript). The transcripts were collapsed into unique genes, with the maximal score for these collapsed transcripts taken to represent the gene. Note that transcripts overlapping the representative SNP itself are assigned a score of 1, and the score monotonically decreases (toward 0) as the genes are farther in physical distance from the representative SNP, providing robustness to r^2 variability (seen here as 'noise') in local genomic regions. These r^2 scores are available for all candidate genes and all traits in **Supplementary Table 2**.

Prix fixe subnetwork enrichment. For each collection of disjoint gene sets, we searched through the CFN to find prix fixe subnetworks (i.e., where each locus was represented by a single gene). Because enumerating all possible such subnetworks is often computationally intractable, we used a genetic algorithm to enrich for dense prix fixe subnetworks, where density is defined as the number of edges within the subnetwork. An initial population of 5,000 random prix fixe subnetworks was chosen (where the gene representing each locus was chosen uniformly at random). Each 'generation' then consisted of a mutation step and a mating step. In the mutation step, genes representing each locus in the prix fixe subnetworks were swapped with other genes from the same locus. Each locus was mutated with a 5% probability, and the replacement gene was chosen from the remaining available genes in that locus uniformly at random. The mating procedure incorporates the notion of fitness by preferentially selecting denser prix fixe subnetworks for mating (and thus propagation to the next generation). The density d_i (edge-count) of each subnetwork i was computed and (cubically) transformed to a selection score, $s_i = d_i^3$, which was then normalized to

$$s_i^* = \frac{s_i}{\sum_{j=1}^{5{,}000} s_j}$$

Pairs of subnetworks were sampled (with replacement), where the probability of selecting a parent subnetwork i was equal to s^*_i. Each mating resulted in a new subnetwork, where the gene chosen for each associated locus was randomly selected from either parent (in a 50/50 coin-flip procedure). After 5,000 such matings, each new population of subnetworks replaced the parental population and the procedure was repeated, starting again with the mutation step. The optimization cycle terminated when the newly generated population's average density failed to improve upon the previous generation's average density by more than 0.5%.

To measure the statistical significance of the final population of subnetworks, we used a randomization strategy intended to simulate the null case where non-informative collections of loci were provided in lieu of the true trait-associated loci. For a set of L loci with G_i genes in locus i, we generated L matched random and disjoint sets of genes, again with G_i genes per set i. To account for possible node-degree effects within the CFN, each random gene was selected such that its degree approximately matches the true candidate gene's degree in the CFN. We chose approximate degree matching over precise degree matching to prevent frequent selection of the actual true genes in the random trials, due to possible uniqueness in the true genes' degree distribution. All genes in the CFN were distributed amongst 128 equal-sized bins based on the genes' degrees (i.e., we used quantile-based binning of the degrees and associated nodes). Each original candidate gene was replaced with a random gene selected from the same bin, thus preserving approximate degree.

Each matched collection of random gene sets was then subjected to the genetic algorithm optimization method, and the average density of the final population in the random trial was used as a test statistic. The observed test statistic for the original loci was compared to test statistics for 1,000 random trials (as described above), resulting in an empirical P value representing the fraction of random trials producing final populations of subnetworks with higher average density than the average density seen with the true loci inputs (**Supplementary Fig. 4**).

Prix fixe gene-scoring. To score each candidate gene, we began with a single prix fixe subnetwork from the final population and modified this subnetwork one locus at time, while keeping the subnetwork constant for all other loci. Consider a single prix fixe subnetwork and a locus i containing G_i genes (g_1, g_2, …, g_{Gi}), where $g^\star$ represents the gene 'chosen' for that locus within the subnetwork. During the scoring procedure, $g^\star$ is 'forgotten' and all G_i genes are considered, whereas the chosen genes for all other loci remain fixed. First, each gene g_i is iteratively used in place of $g^\star$ and the density (edge count) for the subnetwork is recomputed. Then, the density of the subnetwork is recomputed in the absence of any gene for locus i (i.e., as if locus i were to be completely removed from the association study results): the 'empty' locus case. The difference in densities for each gene g_i and this empty locus case indicate the contribution made by gene g_i to the cohesiveness of the rest of the subnetwork. Thus, if two genes are in locus i and have identical connectivity patterns to all other subnetwork loci (for example, if the two genes are paralogs resulting from a localized duplication event), they will acquire the same score for this subnetwork, even if only one gene was chosen for this prix fixe subnetwork during the enrichment procedure described above. Genes with high connectivity to the other loci in the subnetwork will be assigned high scores, whereas genes with low connectivity are similar to the empty locus case and are given low scores. Each locus is similarly considered in turn, and thus all candidate genes are given a single score for each prix fixe subnetwork in the final population. These scores are then averaged over the full population of subnetworks, leading to the aggregate score for each gene. Scores for all genes across all traits are available in **Supplementary Table 2**. Genes at a locus but not found in our CFN were given "NA" scores, indicating the absence of information.

Rank-based prioritization evaluation. To evaluate gene-scoring methods within a locus, we used a rank-based system seeking to identify the rank of a known causal gene (for those loci containing such a known gene). Genes were first ranked according to score, and to compare ranks between loci containing different numbers of genes, we used a relative rank that was rescaled to lie between 0% and 100%. For example, in a locus containing five genes with

doi:10.1038/nmeth.3215

an SCGC gene ranked second, the normalized SCGC gene's rank is 3/4 (with the bottom- and top-ranked genes having relative ranks of 0% and 100%, respectively).

In the GWA studies we analyzed, many of the candidate genes described in the source publications were selected solely on the basis of their distance (either physical or genetic) from the associated tagSNP. Thus, we compared the prix fixe rankings to an alternative closest-gene strategy of ranking genes by tagSNP proximity (as defined by LD value). Genes were assigned scores based on the modeled r^2 decay between the tagSNP and SNPs proximal to the genes (as described above).

Although the LD-based approach alone fared poorly (see Results), we wondered if enhanced SCGC rankings could be achieved using a combined strategy incorporating both LD and prix fixe scores. For these 34 loci, we found that linear combinations of these two scores showed almost no improvement over the prix fixe strategy alone (results not shown), suggesting that within haplotype blocks, local LD structure may be of little additional use in prioritizing candidate disease genes. We also note that although some GWA authors may use existing literature to identify top candidates for a locus, this risks falling into (and contributing to) the cycle of confirmation bias, thus limiting the ability to identify truly novel disease genes via GWA studies.

Replication and parameter variation. As the prix fixe subnetwork enrichment is a stochastic process, we repeated the prix fixe method for all traits to assess score reproducibility. For each trait, prix fixe scores were recomputed, and we assessed correlation between the scores resulting from our primary analysis (presented above) and the replicate scores (using Kendall's τ rank correlation coefficient). All such correlations were found to be very high (ranging between 0.97 and 1.0; **Supplementary Fig. 6a**), verifying that the stochastic search process robustly avoids finding only local optima.

To assess how parameter settings may affect results, we next repeated our analysis for all traits using two different r^2 thresholds: $r^2 \geq 0.25$ and $r^2 \geq 0.75$ (corresponding to 'relaxed' and 'tightened' genomic regions, respectively). Again we computed Kendall's τ across gene scores for each trait with respect to the prix fixe scores for that trait's initial analysis (where the LD threshold was $r^2 \geq 0.50$). As varying the genomic regions often forces inclusion or exclusion of candidate genes, correlations were computed across only those candidate genes shared by both analyses. For both the relaxed and tightened genomic regions, these correlations were generally high (**Supplementary Fig. 6b**,**c**), indicating robustness of results to reasonable settings of LD. Despite resulting in varying numbers of candidate genes, alternative r^2 parameter settings also led to continued enriched prioritization of causal cancer genes (**Supplementary Figs. 7** and **8**), although in both cases with slightly weaker significance levels.

We further extended our repeat analyses to include alternative CFNs. For these trials, we kept the LD threshold fixed at $r^2 \geq 0.50$ (as it led to the best causal cancer gene prioritization; see above). Analyses were then repeated for all traits using three different CFNs: HumanFunc (HF) only, GeneMANIA (GM) only, and the union of HumanFunc, GeneMANIA and the high-confidence subset of STRING[57]. Each analysis was then compared to the primary (i.e., as presented in the main text) analysis for a given trait, again using score correlations. When using either the HF or GM CFN alone, score correlations with the initial combined HM ∪ GM CFN remained high (**Supplementary Fig. 9a**,**b**), though generally lower than those seen while adjusting the LD threshold parameter. This suggests that the prix fixe method exhibits greater sensitivity to the underlying network than to genomic region boundaries. For the addition of STRING data, we first recomputed STRING v9 scores (as described in Franceschini *et al.*[57]) to remove the text-mining contribution to the final STRING score, in an attempt to prevent literature-born confirmation bias. Prix fixe score correlations between the primary analyses and those scores obtained with this augmented CFN remained very high (**Supplementary Fig. 9c**), but we found no improvement in the ability to prioritize causal cancer genes with this larger CFN.

Functional enrichment. Functional enrichment analyses were performed using the FuncAssociate tool[29]. For each trait, we first ranked all candidate genes by their prix fixe score, independently of their genetic location. We then ran an ordered GO-term enrichment analysis, selecting for over-represented GO terms with a multiple testing–corrected *P*-value threshold of 0.05, and terms themselves were ordered by decreasing effect size (odds ratio). All over-represented GO terms for each trait are available in **Supplementary Table 3**.

We note that our CFNs were constructed using shared GO terms as examples of 'gold-standard positive' cofunctional links. For this reason, results should be interpreted primarily as answers to the question: "What types of cofunction examples were useful in this classification process?" and the interpretation of significance levels should account for this potential for circularity.

Independent replication of prix fixe results using T2D GWA. To examine the reproducibility of pathway identification across distinct GWA studies, we performed two type 2 diabetes mellitus (T2D) analyses: one (as part of our primary set of analyses) with loci identified in a study from 2010 (ref. 58) and one (for replication purposes) with loci found in two recent independent T2D GWA studies[59,60]. Our primary analysis of T2D was based on 26 loci, and a functional enrichment analysis revealed diabetes-related pathways such as "glucose homeostasis," "pancreas development" and "insulin secretion" (**Supplementary Table 3**). We then performed a new prix fixe analysis using loci from the 'new' T2D GWA studies that identified 17 loci, 8 of which were unique to these newer recent studies.

Despite sharing only 9 loci (among 26 and 17 total in the two analyses, respectively), the separate analyses both identified genes involved in diabetes-related biological functions, including "glucose homeostasis," "pancreas development" and "insulin secretion" (**Supplementary Tables 3** and **5**). Three of the top eleven scoring genes in our independent replication analysis have verified causal links to T2D, as annotated in the OMIM[41]. These include genes encoding transcription factors *TCF7L2* (*TCF4*), which has extensive evidence of being causal in T2D[61,62], and *HNF1B*, which is a known cause of maturity onset diabetes of the young[63]. Other high-ranking candidate genes have been identified as therapeutic targets in T2D (for example, *CTBP1* (ref. 64) and *LEP*[65]), and the high-scoring gene *HHEX* has recently been shown to play a key role in islet function[66].

Cancer differential expression analysis. We used data from TCGA to estimate differential expression characteristics of genes within cancer-associated loci. TCGA projects using the RNASeqv2 pipeline were chosen, and we downloaded paired tumor-versus-normal samples. Only the breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD) projects had sufficient numbers of matched RNA samples processed by the RNASeqv2 pipeline, and so we downloaded "Level 3" data for both of these projects. All samples were paired using the TCGA participant and sample-type barcodes (identifying patients and tissue types). Unpaired samples (i.e., normal tissue without tumor or vice versa) were not considered for this analysis.

The TCGA RNASeqv2 pipeline reports expected counts as produced by the RSEM[67] program. We rounded raw counts to the nearest integer and estimated differential expression with the edgeR R package[68] using the GLM (general linear model) functions to force treatment of tumor and normal samples in paired fashion. Genes were declared to be significantly differentially expressed if their mean estimated fold change in tumor-versus-normal was greater than 2 (in either direction) and the associated FDR (false discovery rate) was less than 5% (using Benjamini-Hochberg FDR estimation[69]).

Publication rate analysis. To measure the rates of publications referencing genes in the human genome, we used the gene2pubmed data available from the NCBI Gene database[70]. For each gene x, the earliest associated publication was identified and the corresponding year $t_{0,x}$ was used as the first publication year. Then the total number of publications n_x associated with each gene x was found. The subsequent publication rate for gene x was then computed as

$$r(x) = \frac{n_x}{2013 - t_{0,x}}$$

Each year from 1990 to 2012 (inclusive) was then used as a first publication year threshold $t^\star$. Rates for all genes x with $t_{0,x} \leq t^\star$ were averaged, giving the average rate of publications per year for all genes first described during or before year $t^\star$ (**Supplementary Fig. 2**).

Software availability. The methods described here are implemented and available as an R package as **Supplementary Software** and as a web application at http://llama.mshri.on.ca/prixfixe). With the recommended (default) parameter settings described here, most analyses require only a few minutes on standard commodity computers, with minimal memory requirements.

41. Amberger, J., Bocchini, C.A., Scott, A.F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37**, D793–D796 (2009).
42. Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–D312 (2012).
43. Gunsalus, K.C., Yueh, W.-C., MacMenamin, P. & Piano, F. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res.* **32**, D406–D410 (2004).
44. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
45. Östlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38**, D196–D203 (2010).
46. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
47. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
48. Taşan, M. *et al.* An *en masse* phenotype and function prediction system for *Mus musculus*. *Genome Biol.* **9** (suppl. 1), S8 (2008).
49. 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).
50. Musso, G. *et al.* Novel cardiovascular gene functions revealed via systematic phenotype prediction in zebrafish. *Development* **141**, 224–235 (2014).
51. Tian, W. *et al.* Combining guilt-by-association and guilt-by-profiling to predict *Saccharomyces cerevisiae* gene function. *Genome Biol.* **9** (suppl. 1), S7 (2008).
52. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
53. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
54. Ferrari, S. & Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* **31**, 799–815 (2004).
55. Hill, W.G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).
56. Sved, J.A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor. Popul. Biol.* **2**, 125–141 (1971).
57. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
58. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
59. THE SIGMA Type 2 Diabetes Consortium. Sequence variants in *SLC16A11* are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).
60. Hara, K. *et al.* Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum. Mol. Genet.* **23**, 239–246 (2014).
61. Boj, S.F. *et al.* Diabetes risk gene and Wnt effector Tcf7l2/TCF4 controls hepatic response to perinatal and adult metabolic demand. *Cell* **151**, 1595–1607 (2012).
62. Savic, D. *et al.* Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Res.* **21**, 1417–1425 (2011).
63. Bingham, C. & Hattersley, A.T. Renal cysts and diabetes syndrome resulting from mutations in hepatocyte nuclear factor-1β. *Nephrol. Dial. Transplant.* **19**, 2703–2708 (2004).
64. Farmer, S.R. Molecular determinants of brown adipocyte formation and function. *Genes Dev.* **22**, 1269–1275 (2008).
65. Coppari, R. & Bjørbæk, C. Leptin revisited: its mechanism of action and potential for treating diabetes. *Nat. Rev. Drug Discov.* **11**, 692–708 (2012).
66. Zhang, J., McKenna, L.B., Bogue, C.W. & Kaestner, K.H. The diabetes gene *Hhex* maintains δ-cell differentiation and islet function. *Genes Dev.* **28**, 829–834 (2014).
67. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. & Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).
68. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
69. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
70. Maglott, D., Ostell, J., Pruitt, K.D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39**, D52–D57 (2011).

doi:10.1038/nmeth.3215