Shared activity patterns arising at genetic susceptibility loci reveal underlying genomic and cellular architecture of human disease.

ONE SENTENCE SUMMARY: *We discover that variants associated with a specific disease share expression profiles across tissues and cell types, enabling fine mapping and identification of new disease-associated variants, illuminating key cell types involved in disease pathogenesis.*

J. Kenneth Baillie*, Andrew Bretherick, Christopher S. Haley, Sara Clohisey, Alan Gray, Jeffrey Barrett, Eli A. Stahl, Albert Tenesa, Robin Andersson, J. Ben Brown, Geoffrey J. Faulkner, Marina Lizio, Ulf Schaefer, Carsten Daub, Masayoshi Itoh, Naoto Kondo, Timo Lassmann, Jun Kawai, IIBDGC Consortium, FANTOM5 Consortium, Vladimir B. Bajic, Peter Heutink, Michael Rehli, Hideya Kawaji, Albin Sandelin, Harukazu Suzuki, Jack Satsangi, Christine A. Wells, Nir Hacohen, Thomas C Freeman, Yoshihide Hayashizaki, Piero Carninci, Alistair R.R. Forrest*, David A. Hume*

*to whom correspondence should be addressed (j.k.baillie@ed.ac.uk, alistair.forrest@perkins.uwa.edu.au, david.hume@roslin.ed.ac.uk)

Abstract

Genetic variants underlying complex traits, including disease susceptibility, are enriched within the transcriptional regulatory elements, promoters and enhancers. There is emerging evidence that regulatory elements associated with particular traits or diseases share patterns of transcriptional regulation. Accordingly, shared transcriptional regulation (coexpression) may help prioritise loci associated with a given trait, and help to identify the biological processes underlying it. Using cap analysis of gene expression (CAGE) profiles of promoter- and enhancer-derived RNAs across 1824 human samples, we have quantified coexpression of RNAs originating from trait-associated regulatory regions using a novel analytical method (network density analysis; NDA). For most traits studied, sequence variants in regulatory regions were linked to tightly coexpressed networks that are likely to share important functional characteristics. These networks implicate particular cell types and tissues in disease pathogenesis; for example, variants associated with ulcerative colitis are linked to expression in gut tissue, whereas Crohn's disease variants are restricted to immune cells. We show that this coexpression signal provides additional independent information for fine mapping likely causative variants. This approach identifies additional genetic variants associated with specific traits, including an association between the regulation of the OCT1 cation transporter and genetic variants underlying circulating cholesterol levels. This approach enables a deeper biological understanding of the causal basis of complex traits.

Introduction

Genome-wide association studies (GWAS) have considerable untapped potential to reveal new mechanisms of disease[1]. Variants associated with disease are strongly over-represented in regulatory, rather than protein-coding, sequence; this enrichment is particularly strong in promoters and enhancers[2–4]. There is emerging evidence that gene products associated with a specific disease participate in the same pathway or process[5], and therefore share transcriptional control[6].

We have recently shown that cell-type specific patterns of activity at multiple alternative promoters[7] and enhancers[3] can be identified using cap-analysis of gene expression (CAGE) to detect capped RNA transcripts, including mRNAs, lncRNAs and eRNAs[3,5]. In the FANTOM5 project, we used CAGE to locate transcription start sites at single-base resolution and quantified the activity of

267,225 regulatory regions in 1824 human samples (primary cells, tissues, and cells following various perturbations)[8].

Unlike analysis of chromatin modifications or accessibility, the CAGE sequencing used in FANTOM5 combines extremely high resolution in three relevant dimensions: maximal spatial resolution on the genome, quantification of activity (transcript expression) over a wide dynamic range, and high biological resolution – quantifying activity in a much wider range of cell types and conditions than any previous study of regulatory variation[2,4]. Since a majority of human protein-coding genes have multiple promoters[5] with distinct transcriptional regulation, CAGE also provides a more detailed survey of transcriptional regulation than microarray or RNAseq resources. Heritability of traits studied by GWAS is substantially enriched in these FANTOM5 promoters[9].

Genes that are coexpressed are more likely to share common biology[10,11]. Similarly, regulatory regions that share activity patterns are more likely to contribute to the same biological pathways[5]. Transcriptional activity of regulatory elements (both promoters and enhancers[3]) is associated with variable levels of expression arising at these elements in different cell types and tissues[5].

In order to determine whether coexpression can provide additional information to prioritise genome-wide associations that would otherwise fall below genome-wide significance, we developed network density analysis (NDA). The NDA method combines genetic signals (disease association in a GWAS) with functional signals (correlation in expression across numerous cell types and tissues, Figure 1), by mapping genetic signals onto a pairwise coexpression network of regulatory regions, and then quantifying the density of genetic signals within the network. Every regulatory region that contains a GWAS SNP is assigned a score quantifying its proximity in the network to every other regulatory region containing a GWAS SNP for that trait. We then identified specific cell types and tissues in which there is preferential activity of regulatory elements associated with selected disease-related phenotypes, thereby providing appropriate cell culture models for critical disease processes.

Results

Discovery and prioritisation of GWAS hits in regulatory sequence

We defined regulatory regions as the transcription start site (TSS) -300bp and +100bp for promoters[5], and the region between bidirectional TSS for enhancers[3] (See Online Methods). For each of 7 GWAS studies for which high-resolution complete datasets were publicly available, we identified a set of regulatory regions containing variants with GWAS p-values below a permissive threshold (5e-8; Table 1). We devised NDA to examine the similarity in activity patterns among the set of regulatory regions detected in each GWAS (that is, the similarity in expression profile of transcripts arising from these regulatory regions).

NDA detected significant coexpression (see below) among the sets of transcripts arising from regulatory regions containing variants associated with each of the following diseases and traits: ulcerative colitis, Crohn's disease, height, HDL cholesterol, LDL cholesterol, total cholesterol and triglyceride levels (Table 1). One lower-resolution study, of blood pressure, was also analysed: in this smaller study, no coexpression signal was detected among transcripts arising near variants associated with either systolic or diastolic blood pressure (Table 1).

Significant coexpression was only detected within loci containing variants with low p-values (Fig 2a). Similar expression profiles are often seen arising from regulatory regions that are close to each other on the same chromosome, which may also span linkage disequilibrium blocks. The effect of this on the coexpression signal was mitigated by grouping nearby (within 100,000bp) regulatory regions into a single unit, unless they have notably different expression patterns (Fig 2c; Online Methods). SNPs in nearby regulatory regions are also more likely to be in linkage disequilibrium, and these regulatory regions themselves are more likely to share cis- or short range trans- regulatory signals in common. We checked for significant linkage disequilibrium between regulatory regions assigned to independent groups (Supplementary files 1, 4-12). At a threshold of $r^2 > 0.8$, there is no linkage disequilibrium between significantly coexpressed groups; three examples of weaker linkage relationships were detected with $0.08 \leq r^2 \leq 0.6$ (Supplementary file 1).

Regulatory regions around individual TSS with higher coexpression scores contain variants with stronger GWAS p-values (Fig 2b), indicating that this independent signal provides additional information that may be used for fine-mapping causative loci (Fig 2c).

In order to enable the detection of new regulatory regions with strong coexpression relationships, we chose a permissive p-value threshold for trait association of $5x10^{-6}$ (see Online Methods). GWAS data for Crohn's disease[12] were used for initial optimisation of the NDA approach; among GWAS datasets for phenotypes that were not used in algorithm development (i.e. all apart from Crohn's disease), 0-24% of regulatory regions containing a GWAS SNP showed significant coexpression with other regulatory elements associated with the same phenotype (FDR < 0.05, compared with 100 permuted subsets of equal size; see Online Methods).

For a given disease, regulatory regions containing GWAS variants are coexpressed if they share similar activity patterns (i.e. similar expression patterns among transcripts arising from these regulatory regions) with other regulatory regions implicated in that disease. Figure 3 shows significant coexpression superimposed on a two-dimensional representation of the entire network of pairwise correlations. Since activity (transcript expression) was measured in numerous samples, the true proximity of regulatory regions to one another cannot be accurately represented in two dimensions – a perfect representation would require as many dimensions as there are unique samples. However, the NDA method is designed to quantify proximity in network space, so that significantly coexpressed elements are detected, even if they are not directly adjacent on a two-dimensional representation of the network (Figure 3). Among strong coexpression was seen between loci that were widely separated on the genome (Figure 4).

The coexpression signal essentially combines the signal for association in a GWAS with the location and activity pattern of regulatory regions on the genome. We deliberately chose a permissive GWAS p-value threshold in order to enable the detection of new signals that did not achieve genome-wide significance in the original studies. For example, we found that coexpressed transcripts for both LDL and total cholesterol (TC) arise from promoters for well-studied genes such as APOB[13] and ABCG5[14], but also from regulatory regions not previously associated with cholesterol levels. A promoter for SLC22A1, which encodes an organic cation transporter, OCT1[15], is strongly coexpressed among elements associated with both conditions (Supplementary File 1). OCT1 transcription is regulated by cholesterol[16] and the transporter regulates hepatic steatosis through its role in thiamine transport[17]. This action of OCT1 is inhibited by metformin[17], an oral hypoglycaemic agent whose cholesterol-lowering effect[18] is not well understood[19]. Full results of coexpression analyses are in Supplementary File 1, and online at www.coexpression.net.

Cell-type and tissue specificity

The significantly-coexpressed networks detected here could be regarded as revealing the signature expression profile, at least within the FANTOM5 dataset, for a given disease or trait. We next explored whether these signature expression patterns reveal cell types or biological processes that may contribute to the trait or disease susceptibility.

We therefore ranked cell types and tissues by transcriptional activity for each of the significantly-coexpressed loci for each trait, and combined the rankings using a robust rank aggregation[20] (Online Methods). By first detecting the characteristic expression signature associated with a given phenotype using only high-resolution GWAS data, and then detecting the cell type and tissue activity profiles that underlie this signature, we improve on the statistical power of previous methods that have attempted to detect cell-type specific signatures of disease[4,6,21]. Strong signals reported previously are highly significant in our analysis; for example genetic loci associated with cholesterol are transcriptionally active in hepatocytes and liver tissue[6](Supplementary File 8).

This analysis reveals robust cell-type associations that have important implications for understanding disease pathogenesis. For example, cell-type associations with Crohn's disease were restricted to immune cells, particularly monocytes exposed to inflammatory stimuli (Supplementary File 4). In contrast, cell type associations with ulcerative colitis were statistically significant in rectum, colon and intestine samples, and in a distinct group of immune cells: macrophages exposed to bacterial lipopolysaccharide (Supplementary File 5). This is consistent with the view that ulcerative colitis, in which disease processes are primarily restricted to the colon and rectum, is a consequence of dysregulation of processes that are intrinsic to the large bowel, including epithelial barrier function[22], whereas Crohn's disease is a multisystem autoimmune disorder with more diverse extra-intestinal manifestations[23], consistent with a primary immune aetiology.

Discussion

The development of high-throughput genotyping methods has led to an explosion of associations between genetic markers and human diseases[24]. The results presented here are a step towards overcoming the next challenge for this field: making sense of these associations to advance the practice of medicine. There has been increasing recognition of the potential to utilise prior knowledge to improve detection and interpretation of genome-wide signals[25]. The results of our analysis demonstrate that there is biological information in the coexpression of genetic variants associated with a particular disease that can provide the basis for prioritising variants that would not otherwise meet standard thresholds for genome-wide statistical significance.

We report relationships between numerous regulatory regions that are not associated with named genes – a restriction that has previously limited the transition from genetic discovery to biological understanding[26–30]. The analysis reveals the impact of specific enhancers and promoters that may be remote from the genes they regulate, or may contribute to tissue-specific regulation of a gene that may otherwise appear to be more widely-expressed.

Even for those disease-associated variants that can be reliably assigned to a named gene, previous attempts to draw functional inferences have, by necessity, relied on published data[26], annotated biological pathways[31], or gene sets[30,32]. Although many important insights have been gained from these approaches, they share a fundamental limitation: reliance on existing knowledge. This restricts the ability to exploit the potential of genomics to deliver insights into new, previously unseen, mechanisms of disease[33].

The data used for development and testing of the coexpression approach were from large meta-analyses that incorporate genotyping (or imputation) of genetic variants at extremely high resolution, increasing the probability that variants will be found within regulatory regions. In future, the availability of whole-genome sequencing can reasonably be expected to produce many additional high-quality datasets for coexpression analysis. In principle, the NDA approach can be generalised to any network in which it is desirable to quantify the proximity of a subset of nodes.

The scale, depth and breadth of the FANTOM5 expression atlas, together with the NDA approach, enable detection of subtle coexpression signals for regulatory regions that have previously been undetectable. As additional genetic studies become available at greater genotyping resolution, we anticipate that this method will detect new genetic associations with disease, coexpressed modules underlying pathogenesis, identify critical cell types implicated in mechanisms of disease.

DATA ACCESS

The FANTOM5 atlas is accessible from http://fantom.gsc.riken.jp/data/

An online service running the coexpression method is available at https://coexpression.roslin.ed.ac.uk

username: fantom5

password: review

ACKNOWLEDGEMENTS

We would like to express our gratitude for the diligence and professionalism of the entire FANTOM5 consortium and to the members of the IIBDGC group, GIANT consortium, and Global Lipids consortium for freely sharing their data. We are particularly grateful to the tens of thousands of patients and healthy volunteers who donated DNA and other material to these studies.

JKB gratefully acknowledges funding support from a Wellcome Trust Intermediate Clinical Fellowship (103258/Z/13/Z) and a Wellcome-Beit Prize (103258/Z/13/A), BBSRC Institute Strategic Programme Grant to the Roslin Institute (BBS/E/D/20241864), the UK Intensive Care Foundation, and the Edinburgh Clinical Academic Track (ECAT) scheme. Funds were provided to the Roslin Institute through a BBSRC Strategic Programme Grant (JKB, SC, CSH, GJF, TCF, DAH; BBS/E/D/20211551, BBS/E/D/20231760). We acknowledge the financial support provided by the MRC-HGU Core Fund (CSH, AT). FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to YH and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to YH. RIKEN Centre for Life Science Technologies, Division of Genomic Technologies members (RIKEN CLST (DGT)) are supported by institutional funds from the MEXT, Japan. ARRF is supported by a Senior Cancer Research Fellowship from the Cancer Research Trust and funds raised by the Ride to Conquer Cancer. JCB is supported by Wellcome Trust grant WT098051. GJF acknowledges the support of an NHMRC Career Development Fellowship (GNT1045237), NHMRC Project Grants (GNT1042449, GNT1045991, GNT1067983 and GNT1068789), and the EU FP7 under grant agreement No. 259743 underpinning the MODHEP consortium. MR was supported by grants from the Deutsch Forschungsgemeinschaft, the German Cancer Aid and the Rudolf Bartling Foundation. RA was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 638273). US and VBB are supported by the KAUST Base Research Fund to VBB and KAUST CBRC Base Fund. RMP is supported by grants from the US National Institutes of Health (R01-AR057108, R01-AR056768, U01-GM092691 and R01-AR059648) and holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. RA and AS were supported by funds from FP7/2007-2013/ERC grant agreement 204135, the Novo Nordisk foundation, and the Lundbeck Foundation and the Danish Cancer Society. CAW is supported by a Queensland Government Smart Futures Fellowship, and samples were collected under Australian National Health and Medical Research council project grants 455947 and 597452, under agreement from the Australian Red Cross 11-02QLD-10 and the University of QLD ethics committee.

Authors' contributions

JKB conceived the study, designed and led the analyses and wrote the manuscript. AB and AG contributed to computational optimisation and description of methods. AB and SC generated network and circos images, respectively. CH, JB, JBB, TF, and AT advised on statistical and network analysis methods. ML managed the data collection, including annotation, expression profiling, metadata association and archiving. CW, JS, NH and TF contributed biological expertise. CW, RA, JS, AS, MR, VB, and PH advised on methodology. ARRF, MI, CD, NK, TL, JK, HS, HK, YH, and PC organised the FANTOM5 project including sample collection, data production, mapping and tag clustering. JKB, ARRF, DAH, TF, GJF, PC and YH provided resources. DAH and ARRF advised on methodology, and contributed to the manuscript. All authors contributed to and approved the final version of the manuscript.

DISCLOSURE DECLARATION

All authors report that they have no conflicts of interests to declare in respect of this manuscript.

REFERENCES

1. Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14,** 139–149 (2013).

2. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).

3. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).

4. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).

5. Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J.K., et al. A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).

6. Marbach, D. *et al.* Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Meth* **13,** 366–370 (2016).

7. The FANTOM Consortium *et al.* The Transcriptional Landscape of the Mammalian Genome. *Science* **309,** 1559–1563 (2005).

8. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347,** 1010–1014 (2015).

9. Finucane, H. K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *bioRxiv* 14241 (2015). doi:10.1101/014241

10. Hume, D. A., Summers, K. M., Raza, S., Baillie, J. K. & Freeman, T. C. Functional clustering and lineage markers: insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations. *Genomics* **95,** 328–338 (2010).
11. Mabbott, N. A., Baillie, J. K., Hume, D. A. & Freeman, T. C. Meta-analysis of lineage-specific gene expression signatures in mouse leukocyte populations. *Immunobiology* **215,** 724–736 (2010).
12. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* **42,** 1118–1125 (2010).
13. Tybjærg-Hansen, A., Steffensen, R., Meinertz, H., Schnohr, P. & Nordestgaard, B. G. Association of Mutations in the Apolipoprotein B Gene with Hypercholesterolemia and the Risk of Ischemic Heart Disease. *New England Journal of Medicine* **338,** 1577–1584 (1998).
14. Lee, M.-H. *et al.* Identification of a gene, ABCG5, important in the regulation of dietary cholesterol absorption. *Nat Genet* **27,** 79–83 (2001).
15. Klaassen, C. D. & Aleksunes, L. M. Xenobiotic, Bile Acid, and Cholesterol Transporters: Function and Regulation. *Pharmacol Rev* **62,** 1–96 (2010).
16. Dias, V. & Ribeiro, V. The expression of the solute carriers NTCP and OCT-1 is regulated by cholesterol in HepG2 cells. *Fundam Clin Pharmacol* **21,** 445–450 (2007).
17. Chen, L. *et al.* OCT1 is a high-capacity thiamine transporter that regulates hepatic steatosis and is a target of metformin. *Proc Natl Acad Sci U S A* **111,** 9983–9988 (2014).
18. Bailey, C. J. & Turner, R. C. Metformin. *N. Engl. J. Med.* **334,** 574–579 (1996).
19. Shaw, R. J. *et al.* The kinase LKB1 mediates glucose homeostasis in liver and therapeutic effects of metformin. *Science* **310,** 1642–1646 (2005).
20. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28,** 573–580 (2012).
21. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
22. Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448,** 427–434 (2007).

23. Mekhjian, H. S., Switz, D. M., Melnyk, C. S., Rankin, G. B. & Brooks, R. K. Clinical features and natural history of Crohn's disease. *Gastroenterology* **77,** 898–906 (1979).
24. Li, M. J. *et al.* GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res* **40,** D1047–D1054 (2012).
25. MacLeod, I. M. *et al.* Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* **17,** 144 (2016).
26. Raychaudhuri, S. *et al.* Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. *PLoS Genetics* **5,** e1000534 (2009).
27. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6,** 5890 (2015).
28. Wojcik, G. L., Kao, W. L. & Duggal, P. Relative performance of gene- and pathway-level methods as secondary analyses for genome-wide association studies. *BMC Genet* **16,** (2015).
29. 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).
30. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 15545–15550 (2005).
31. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28,** 27–30 (2000).
32. Nam, D., Kim, J., Kim, S.-Y. & Kim, S. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucl. Acids Res.* **38,** W749–W754 (2010).
33. Baillie, J. K. Targeting the host immune response to fight infection. *Science* **344,** 807–808 (2014).

Trait	SNPs included, p$< 5 \times 10^{-6}$ (SNPs per million bases)	Regulatory regions containing a SNP (SNPs per million bases)	Fold enrichment for SNPs in regulatory regions	Distinct regulatory regions	Significantly co-expressed TSS ($FDR < 0.05$)(% of distinct regions)
Crohn's disease*	1924 (0.6)	133 (3.5)	5.7	70	23 (33%)
Ulcerative colitis	2162 (0.7)	146 (3.8)	5.5	83	20 (24%)
HDL	5410 (1.7)	260 (7.2)	4.2	101	17 (17%)
LDL	4644 (1.5)	205 (5.2)	3.5	92	19 (21%)
Total cholesterol	6421 (2.0)	316 (8.3)	4.1	128	29 (23%)
Triglycerides	4863 (1.5)	254 (7.0)	4.6	97	23 (24%)
Height	8882 (2.8)	358 (7.6)	2.7	166	29 (17%)
SBP	417 (0.1)	20 (0.4)	3.0	13	0 (0%)
DBP	711 (0.2)	20 (0.4)	1.9	14	0 (0%)

Table 1: Results of coexpression analysis for a range of human traits for which high-quality data are available: Crohn's disease, ulcerative colitis, high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol, triglycerides, height, systolic blood pressure (SBP) and diastolic blood pressure (DBP). *Initial optimisation and parameterisation of the algorithm was undertaken using a random subset of data from this study.

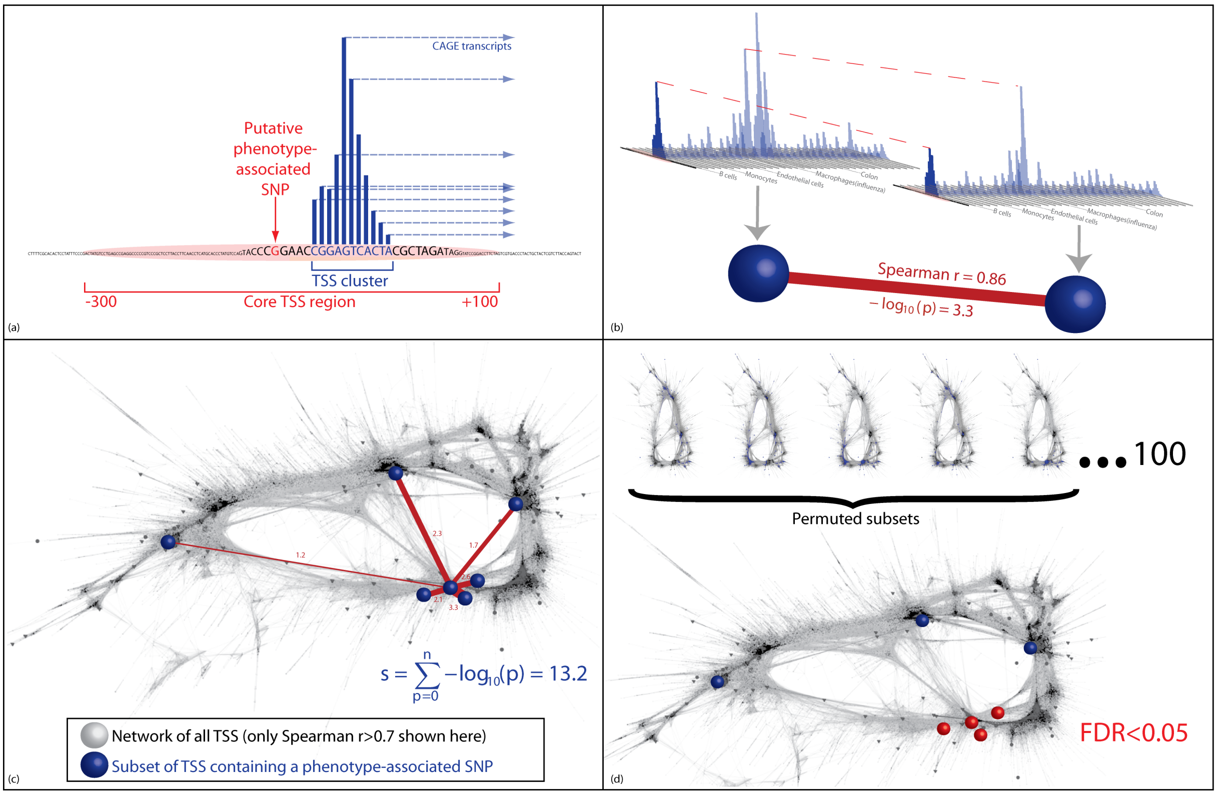


Figure 1: Use of NDA to detect coexpression. a) A subset of regulatory elements is identified containing disease-associated SNPs. b) The strength of the links between pairs of these regulatory regions is quantified, first as the Spearman correlation, then as the $-log_1 0$p-value quantifying the probability, specific to this regulatory region, of a Spearman correlation of at least this strength arising by chance. This is determined from the empirical distribution of correlations between this regulatory region and all other regulatory regions in the entire network of all regulatory regions in the genome. c) The subset of regulatory regions containing disease-associated SNPs form an unexpectedly dense grouping in the network, but this may not be visible in a two-dimensional representation (for illustration, this network shows all correlations between regulatory regions with Spearman $r > 0.7$, layout generated by the FMMM algorithm). The NDA score assigned to any one node is the sum of the links it shares with other nodes in the chosen subset (see Supplementary Methods for a full explanation). d) NDA scores from the input subset of regulatory elements are compared with NDA scores from permuted subsets of regulatory elements in order to quantify the false discovery rate (FDR).

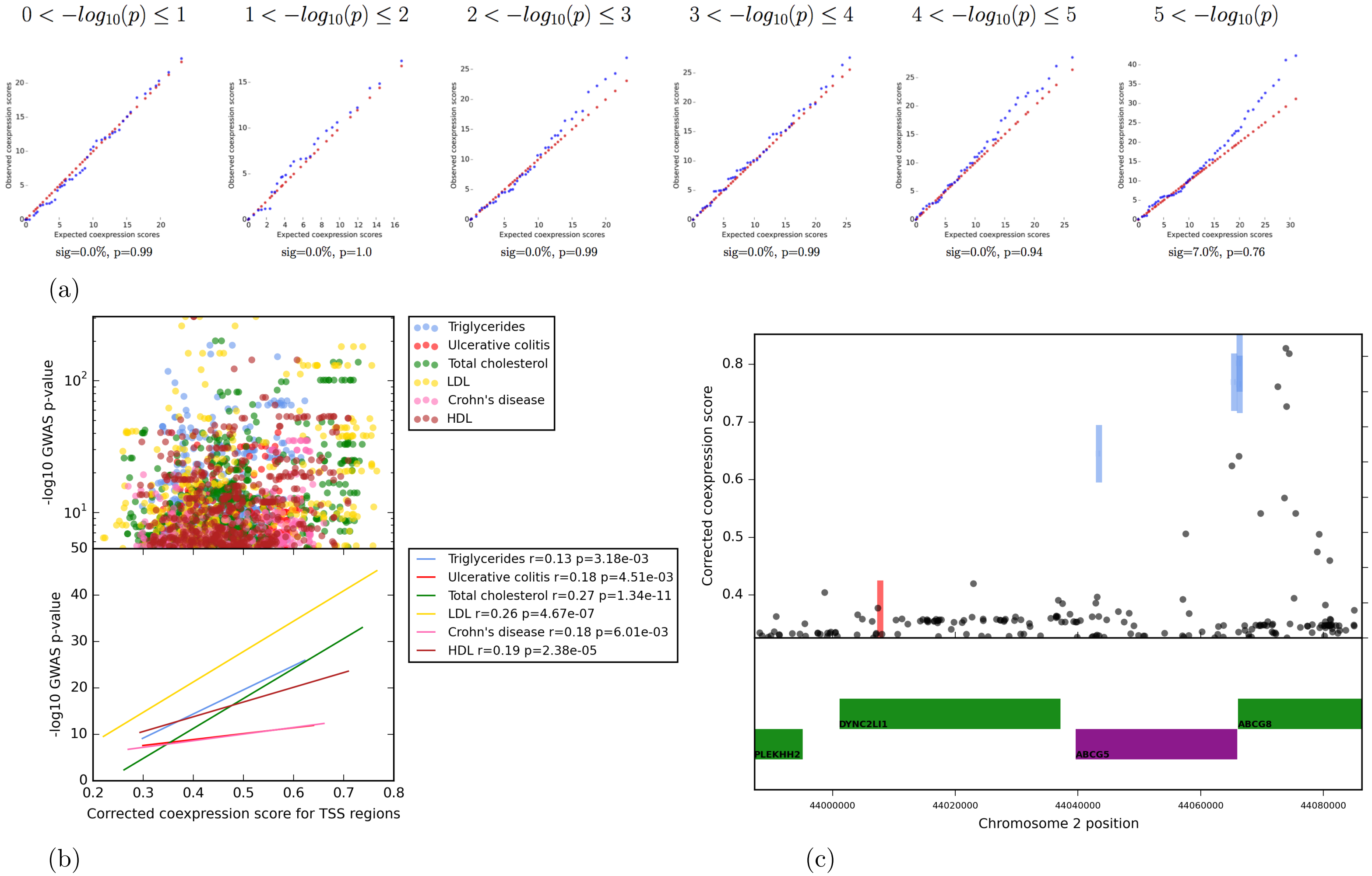
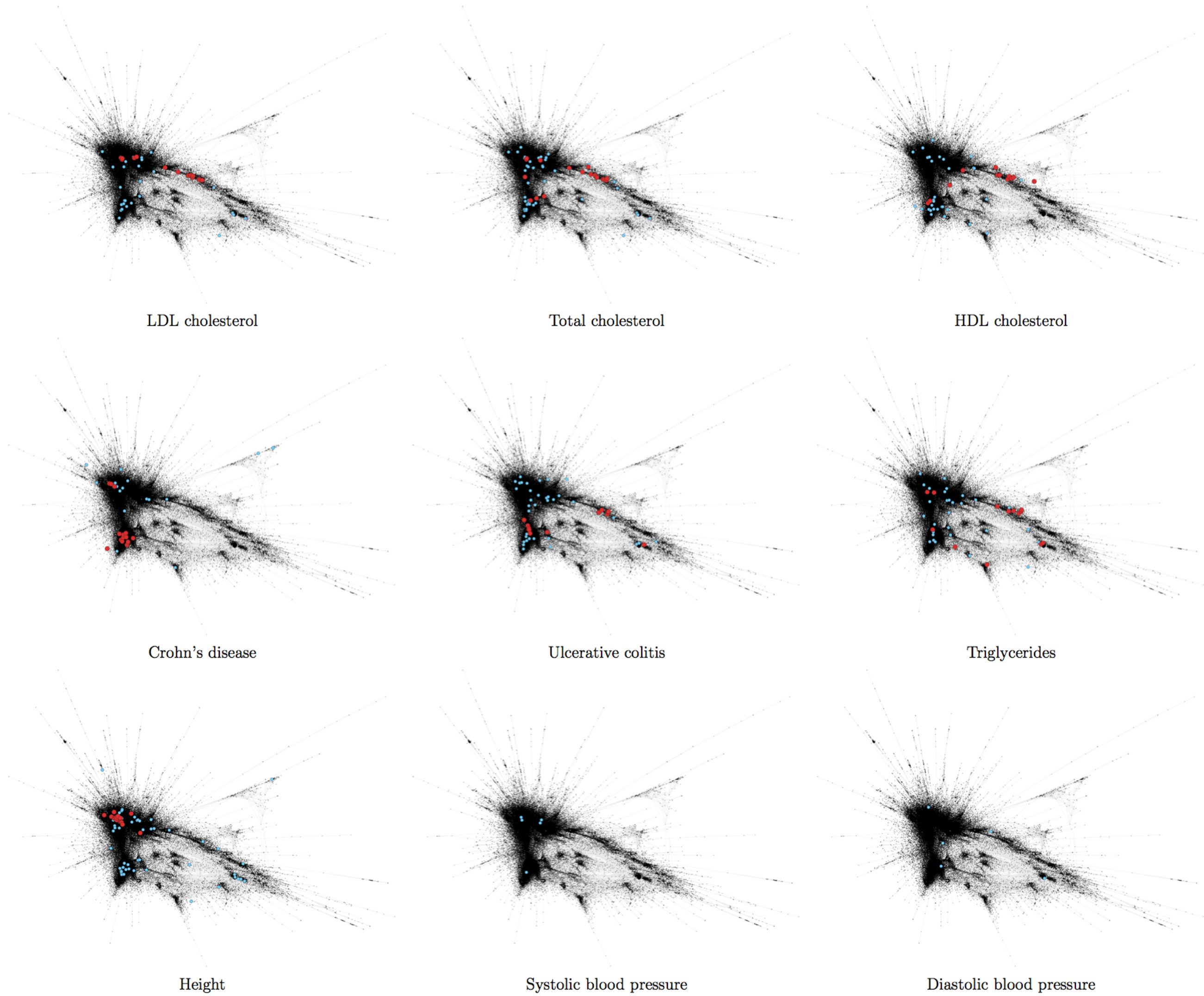


Figure 2: a. Change in coexpression signal in 800 SNPs selected at random from GWAS of Crohn's disease $-log_{10}(p)$ bins from 0 to 5. No signal for coexpression is detected at weak p-values. Percentage of significantly coexpressed entities (hits, $FDR < 0.05$) and p-value (Kolmogorov-Smirnov test) comparing observed and expected distributions are shown below each plot. b. Relationship between GWAS p-value for a SNP, and coexpression scores of individual promoters assigned to that SNP. Top panel: GWAS p-values (log scale) vs corrected coexpression scores. Bottom panel: linear regression lines for data in top panel; Spearman's r and associated p-values are shown for each trait. Only significantly coexpressed ($FDR < 0.05$) promoters are included. c. Detail of chromosomal region containing variants associated with LDL cholesterol. Top panel: Rectangles show corrected coexpression scores of individual regulatory regions; groups of regulatory regions considered as a single unit share the same colour. Black circles show GWAS p-values for individual SNPs. Bottom panel: known protein coding transcripts in sense (green) and antisense (purple).



LDL cholesterol

Total cholesterol

HDL cholesterol

Crohn's disease

Ulcerative colitis

Triglycerides

Height

Systolic blood pressure

Diastolic blood pressure

Figure 3: Network layouts (Spearman $r > 0.5$, FMMM layout algorithm, largest component only is shown) showing position of significant hits on a two-dimensional network representation of FANTOM5 regulatory regions. Red circles: significantly-coexpressed ($FDR < 0.05$) regulatory regions containing a putative GWAS hit ($p < 5 \times 10^{-6}$) for this trait. Blue circles: regulatory regions containing a putative GWAS hit ($p < 5 \times 10^{-6}$) for this trait that are not significantly coexpressed ($FDR > 0.05$).

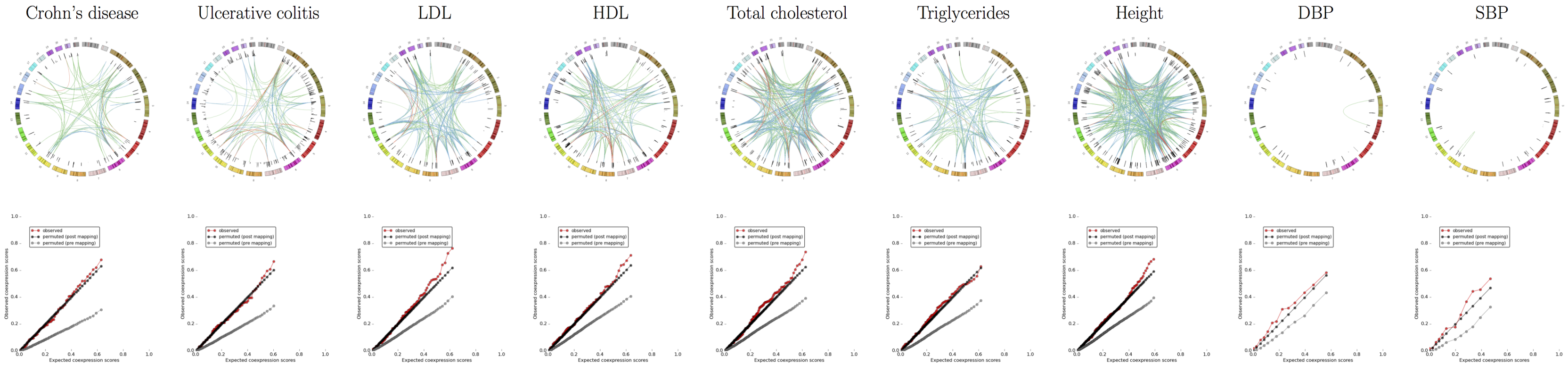


Figure 4: (Top panels) Circular plots of coexpression links between different locations on the genome, illustrating the spatial separation of highly-correlated regulatory regions. The coloured outer circle shows an end-to-end concatenated view of the human chromosomes. The black inner circle shows log_{10} GWAS p-values for included SNPs. Links depict an association between two regulatory regions containing these wSNPs and are coloured according to $-log_{10}(p)$(line colour indicates $log_{10}(p)$: red>3, blue> 2, green> 1.5). (Bottom panels) Quantile-quantile plots showing observed and expected coexpression scores. Expected coexpression scores are derived from circular permuted subsets of regulatory regions (post-mapping permutations; black circles) or SNPs chosen by circular permutations against the background of all SNPs genotyped in each study. Data are shown for Crohn's disease, ulcerative colitis, high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol, triglycerides, height, systolic blood pressure (SBP) and diastolic blood pressure (DBP)

SUPPLEMENTARY METHODS

1 Regulatory regions

For the purpose of this analysis, promoters identified in the FANTOM5 dataset were defined as the region from -300 bases to +100 bases from a transcription start site (Figure 1a, main paper). Previous analysis demonstrated that this covers the areas of maximal sequence conservation across species[4] and the core region of transcription factor binding[12]. Since eRNA TSS are considerably longer than promoter TSS (median length(IQR) 272(173-367) vs 15(9-26)), enhancers were defined by the range covered by eRNA transcription start sites[10].

2 Coexpression algorithm

For each GWAS study, SNPs were identified that lie within either a functional promoter or enhancer. Any promoter or enhancer that contained a variant putatively associated with a given phenotype was considered to be candidate phenotype-associated regulatory region. A pairwise coexpression matrix was then generated across the full FANTOM5 dataset of promoters and enhancers, in which each node is a regulatory region, and edges reflect the similarity in activity (expression) patterns arising at these regulatory regions, across different cell types and tissues.

To test the hypothesis that regulatory regions genetically associated with a given phenotype are more likely to be coexpressed, we devised a method to quantify coexpression among a pool of putative phenotype-associated regulatory regions (network density analysis; NDA). This approach avoids arbitrary cut-offs between clusters (or communities) of nodes, and yields a single value for each node, quantifying the closeness with all other nodes in a particular subset (network density). NDA was used to integrate the putative association between a regulatory sequence and the phenotype of interest (indicated by the presence of a phenotype-associated SNP), with the coexpression similarity between this node with other nodes that are also putatively associated with the same phenotype.

2.1 Principle of network density analysis (NDA)

NDA integrates information from two distinct and independent sources: the relationships between nodes in the network, and the choice of subset. In the present work, nodes are regulatory regions, the subset is those regulatory regions that contain variants associated with a particular phenotype, and the relationships are Spearman's rank correlations. However, the NDA approach is generalisable to any network of pairwise relationships.

Within a network of all possible pairwise relationships between nodes, a subset of nodes is selected that share a particular characteristic.

Within this subset of nodes, every pair of nodes is considered. Each relationship between two nodes is expressed as the $-\log_{10}$ of the empirical probability of a relationship at least as strong occurring between the chosen node and another, randomly-chosen, node from anywhere in the whole network. These probabilities are specific to each node and are directional. The NDA score is the sum of the $-\log_{10}(p)$ values for a node in the chosen subset and all other nodes within the subset. The NDA score therefore quantifies the density of this subset of nodes in network space. The purpose of using the empirical probability of a correlation, rather than the raw correlation metric, is to control for bias in favour of highly-connected nodes, as would occur if one expression profile were very common. Finally, the NDA score is assigned its own p-value by comparison to that obtained using randomly permuted subsets (see below). If the network contains no additional information about this subset of nodes, then the relationships between nodes in the chosen subset will be no stronger than the relationships seen in permuted subsets.

2.2 Application to coexpression of regulatory regions

From the set of all nodes in a network, a subset is selected because they share some characteristic. In the case of the genomic analyses reported here, the nodes are TSS, and the subset of interest is those TSS that contain a variant that has some evidence of association with a particular trait. Throughout this paper, we have defined the set of phenotype-associated transcription start sites, R, as follows: the set of regulatory elements associated with phenotype-associated single nucleotide polymorphism within 300bp (promoters) or 0bp (enhancers) upstream from a FANTOM5 transcription start site (TSS) and 100bp (promoters) or 0bp (enhancers) downstream. In order to enable the detection of new associations, we use a deliberately permissive threshold. We define as "putatively-significant" a SNP-phenotype association of $p < 5 \times 10^{-6}$. Let the integer variable i be used to index the base pairs (bp) of the genome. For a given trait, the set of input SNPs, K, are those that have a putatively-significant association with that trait at our chosen threshold. If we let TSS_{start} equal the base pair index 300bp (promoters) or 0bp (enhancers) upstream from a FANTOM5 transcription start site (TSS) and TSS_{end} 100bp (promoters) or 0bp (enhancers) downstream, the set, P, of putative trait-associated promoters is given by:

$$P = \{i : i \in K, TSS_{start} - 300 \leq i \leq TSS_{end} + 100\} \qquad (1)$$

and the set E of enhancers containing a putative trait-associated SNP is given by:

$$E = \{i : i \in K, TSS_{start} \leq i \leq TSS_{end}\} \qquad (2)$$

giving a total set of regulatory regions:

$$R = P \cup E \qquad (3)$$

2.3 Linkage disequilibrium (LD) - grouping nearby regulatory regions

Input SNPs from GWAS results tend to be in LD with nearby variants. There is therefore a risk of spurious coexpression, since nearby regulatory regions are also likely to share regulatory influences, such as chromatin accessibility, enhancers, and lncRNAs. One solution to this would be to filter input SNPs by LD. However this would require that LD relationships for all SNPs be known for all of the populations from which SNP association data were derived, which is not the case. It would also risk removing functionally important regulatory regions from the analysis, by choosing only one SNP per LD block.

In order to overcome these problems, we sought to identify those regulatory region-associated SNPs within a given region that are most likely to contribute to a given subnetwork of putative phenotype-associated regulatory regions. By the definitions described above, these will be those regulatory regions with the highest NDA score. Regulatory regions are considered for combination if they are separated by 100,000bp or less. If any regulatory regions within this range has a correlation p-value of less than 0.1 with any other regulatory regions in the range, they are combined. A single representative regulatory region is then chosen - the regulatory region with the largest NDA score in the group, derived from a network comprised of all other groups.

In order to confirm that spuriously significant coexpression signals are not being generated because of LD, we used the ENSEMBL Perl API for the 1000 genomes phase 3 data (CEU) to search for variants in LD with each SNP lying within the chosen regulatory region for each group. Variants in LD with a variant in any other chosen regulatory region are reported.

2.4 Coexpression matrix

Let A be the set of all nodes in the whole network. Each member of A is a node in an interaction network. For each $i \in R$, Spearman's rank correlation, x, is calculated with each other node in R. The probability, p, of a correlation as strong or stronger as the index correlation, x, arising by a chance pairing between the index node and any other node ($n_{(r>x)}$) is inferred from the empirical distribution of all correlations (r) of the index node in A.

$$p = \frac{n_{(r>x)}}{n_A} \tag{4}$$

2.5 Network density analysis

For every node in the set R, a score s is calculated to summarise the strength of interactions with all other nodes in R. Since the only thing that the elements of R have in common is that they are TSS identified by the set of input SNPs, unexpectedly strong inter-relationships between elements of R are taken as indirect evidence of a relationship between the input SNPs themselves. The NDA score, s, is defined as the sum of $-log_{10}(p)$ values for interaction strength within the matrix.

$$s = \sum_{p=0}^{n} -log_{10}(p) \tag{5}$$

Raw p-values are calculated from the empirical distribution of values of s for 10000 permuted networks. The Benjamini-Hochberg method is used to estimate false discovery rate (FDR). Significant network density scores are taken as those with $FDR < 0.05$. In order to enable coexpression scores to be compared loosely between different analyses, each raw coexpression scores (s) is corrected by dividing by the total number of independent groups of regulatory regions included in each analysis, n_res, yeilding a corrected coexpression score, ccs:

$$ccs = s/n_res \tag{6}$$

2.6 Iterative recalculation

The node in the network with the highest NDA score has, by definition, numerous strong correlations with other nodes in the subset R. The NDA scores assigned to these other nodes are therefore inflated by their association with the stongest node. This inflation may reflect biological reality, since both TSS have a putative genetic association with the phenotype of interest, and both share strong links. However, there is a risk that TSS sharing a chance association with a strongly coexpressed TSS will be spuriously inflated to significance. For this reason, we have applied a stringent correction in order to ensure that we have confidence in each significantly coexpressed TSS independently of all TSS with stronger coexpression in the network: the NDA score for each TSS is calculated after removing all TSS with stronger NDA scores from the network.

2.7 Input datasets

Of 267,225 robust promoters and enhancers identified by FANTOM5[6], 93,558 (50.6%) were promoters within 400 bases of the 5′ end of a known transcript model[6]. These were annotated with the name of the transcript. Alternative promoters were named in order of the highest transcriptional activity[6]. Where necessary, coordinates for GWAS SNPs (see 2.11 were translated to hg19 coordinates using LiftOver[9], or coordinates were obtained for SNP IDs from dbSNP[14] version 138.

2.8 Permutations

A circular permutation method was devised to prevent systematic bias by maintaining the underlying structure of GWAS SNP data. The NDA score for a given regulatory region was compared with NDA scores obtained from randomly permuted subsets of genes to give an empirical p-value for coexpression. If permuted networks consist of randomly-selected regulatory regions, then this p-value quantifies coexpression alone (see 2.8.1); if the permuted networks are generated by mapping randomly-selected SNPs to regulatory regions, then the final p-value is a composite of two measures: coexpression, and the enrichment for true GWAS hits in regulatory sequence (see 2.8.2)).

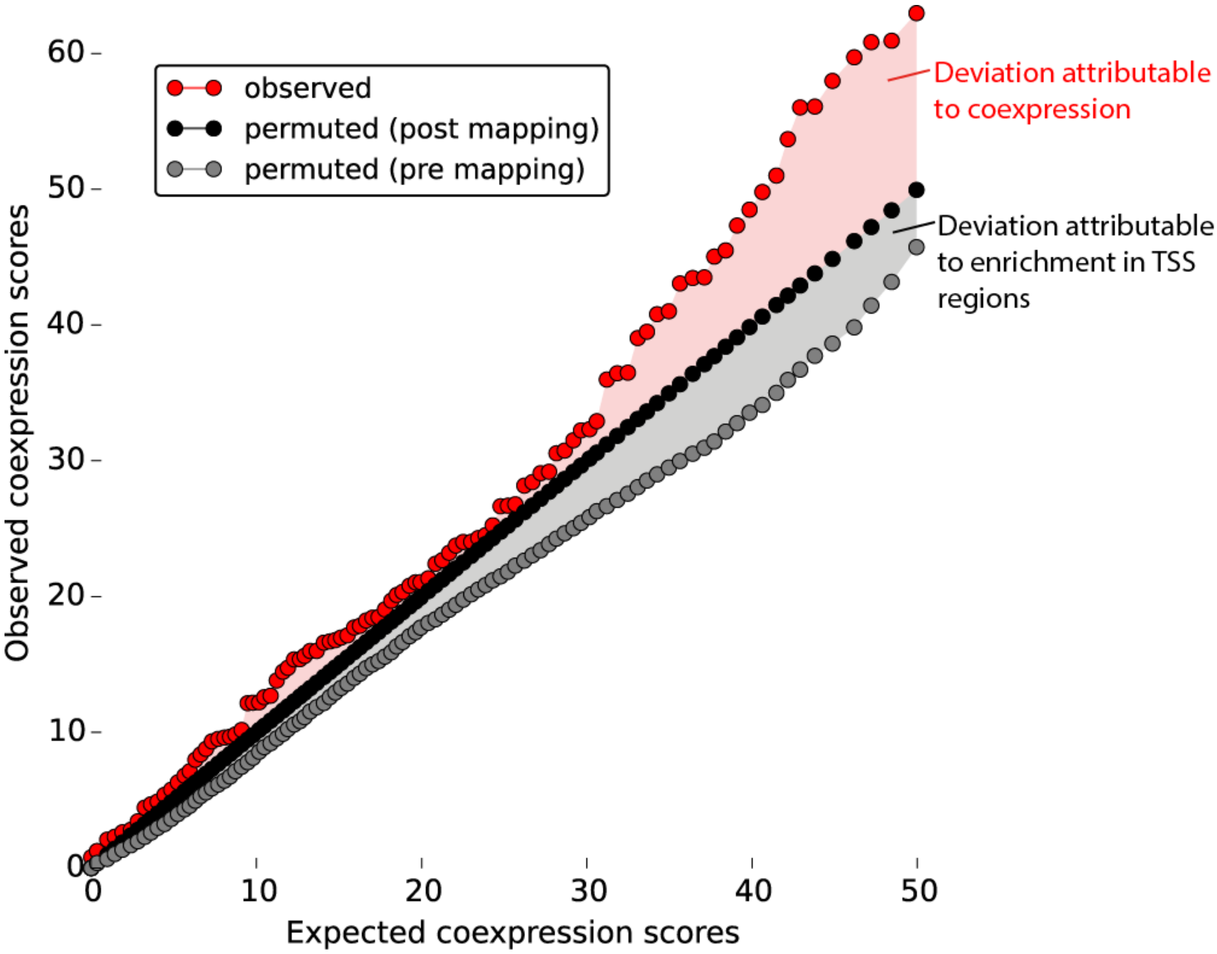
2.8.1 Pre-mapping permutations

Pre-mapping permutations use a random set of SNPs generated by rotation of the input set of SNPs, K, on a concatenated circular genome. The choice of background is critical - some more recent GWAS studies consider only a subset of variants with a high probability of association with a given trait, such as the immunochip[16] or the metabochip[17]. In the present analyses, background data were chosen to reflect as accurately as possible the pool of variants included in the original study. For this reason, results are presented only for phenotypes for which the the entire summary dataset was available, including a p-value for every SNP, so that the background used to generate permuted networks is exactly the same background from which the real dataset is drawn.

2.8.2 Post-mapping permutations

In order to quantify the effect of coexpression alone (i.e. eliminating the inflation of NDA scores that occurs due to enrichment of trait-associated SNPs in regulatory regions), permuted networks were generated after mapping to TSS regions. Let A be the whole set of FANTOM5 TSS. Post-mapping permutations select a random subset of A in a similar circular manner, by randomly displacing the members of the set R (Equation 3) by a random number of places on the list. Where the displacement pushes members of R off the end of the list, they are re-entered at the beginning.

This process generates a pool of variants that are likely to be grouped in a similar distribution on the genome as the input set. If the input set contains a large group of TSS regions in close proximity on the genome, it is likely that this group of TSS regions will be joined as a single unit (see above) for analysis. During generation of permutations, the same number of consecutive TSS regions elsewhere on the genome may not be in sufficient proximity (and expression correlation) to be grouped together. This would create extra network nodes, falsely inflating the NDA scores in the permuted sets. In order to mitigate against this, those TSS from each permutation that do



Supplementary Figure 1: Distribution of observed NDA scores for Crohn's disease, and expected NDA scores from pre- and post-mapping permutations

not conform to the input set distribution are re-entered into a further circular permutation until an identical distribution is found. If no matching grouping is found after 8 repeat permutations, additional regulatory regions are added from consecutive positions above and below whichever group is nearest in size to the relevant group in the original input dataset.

The difference between the distributions of NDA scores derived from pre- and post-mapping permutations reveals the different components of the measure. When compared to a random pool of SNPs (pre-mapping permutations), two factors inflate the NDA scores for real GWAS data: firstly, more regulatory regions are identified because true GWAS hits are enriched within regulatory regions; secondly, the coexpression signal itself is greater for real data. In contrast, post-mapping permutations have precisely the same number of regulatory regions included as the real dataset, so there is no component of inflation due to enrichment in regulatory regions. The effects of these different components are shown in Figure 1, which reveals the NDA score to be a composite measure of both signals.

False discovery rates (FDR) are calculated using the Benjamini-Hochberg method[2].

2.9 Choice of samples and regulatory regions

The enrichment for GWAS hits from a pooled resource comprising the NCBI GWAS catalog and the GWASdb database (observed $SNPs.Mb^{-1}$: expected $SNPs.Mb^{-1}$) was quantified at increasing search window sizes upstream and downstream from the transcription start site (TSS). A table of GWAS hits for a broad range of phenotypes was obtained from the NCBI GWAS catalog[8] and from a larger, less selective catalog of GWAS p-values meeting permissive criteria for genome-wide significance, GWASdb[13]. The GWASdb dataset is less fastidiously curated than the NCBI GWAS catalog, but contains a much greater range of SNPs since it does not restrict inclusion to the strongest associations, or to putative causative variants. Since both databases are limited by the variation in reporting, and quality, of the original GWAS studies from which data are drawn, this analysis was restricted to variants meeting genome-wide significance at a widely-accepted threshold ($p < 5 \times 10^{-8}$). These catalogues were combined and filtered to remove duplicate entries. Data were obtained from:

- NHGRI GWAS catalog, June 2014 `http://www.genome.gov/gwastudies`
- GWASdb2, June 2014 update `ftp://jjwanglab.org/GWASdb/20140629/gwasdb_20140629_snp_trait.gz`

Overlapping phenotypes, such as "urate" and "uric acid" were manually merged as shown in SF2_phenotype_matching.txt. Phenotypes that were considered to be too broad to be informative were excluded, as were those that were not related to human disease. A complete table of phenotypes in GWASdb and NCBI GWAS catalog, showing mergers and inclusion/exclusion in the present work, is provided in a supplementary file (SF2_phenotype_matching.txt).

The coexpression signal obtained for the test input set was evaluated using different subsets of FANTOM5 samples (cell lines, timecourses following a perturbation in primary cells or selected cell lines, tissue samples, primary cells, or various combinations of these), and different types of regulatory region (enhancers, promoters assigned to annotated genes, other promoters, or all regulatory regions combined)(Supplementary Figure 3). A weak signal for coexpression is seen in cell lines, but the addition of cell lines to the combined sample set of primary cells, timecourses and tissues did not improve the coexpression signal seen for any subset of regulatory regions. The strongest coexpression is seen in the combined sample set. A "minimal detail" sample set was also tested, comprising a single average value for each of the timecourses, primary cell types and tissue types, and excluding data from unstimulated cell lines. The complete dataset, including all cell types and tissues, provided the strongest signal, demonstrating that there is additional biologically-relevant information contained in the expression profiles from all sample subsets (Supplementary Figure 3).

2.10 Anti-correlation

Strong anti-correlation between pairs of TSS associated with the same phenotype may have biological importance, such as down-regulation at one TSS but expression at another, or negative regulation of a signalling pathway on which expression of a TSS is dependent. For this reason, anti-correlations may improve detection of true associations in this analysis. However, in order to confer an overall improvement on the performance of the algorithm, true inverse expression relationships between phenotype-associated TSS would need to be sufficiently common to overcome the noise added by incorporating all strong anti-correlations into the NDA score.

Anti-correlations do not contribute any net improvement to the NDA scores for a training set (Crohn's disease, 50% of all SNPs, chosen at random), and were therefore excluded.

2.11 GWAS data sources

Full GWAS or meta-analysis data, reporting every SNP genotyped or imputed in a given study, are required in order to permute subsets against the appropriate background for a given study (see 2.8). These were obtained from the following sources:

- Crohn's disease[7] summary *p*-values were obtained from the International Inflammatory Bowel Disease Genetics Consortium `ftp://ftp.sanger.ac.uk/pub4/ibdgenetics/cd-meta.txt.gz`
- Ulcerative colitis[1] summary *p*-values were obtained from the International Inflammatory Bowel Disease Genetics Consortium `ftp://ftp.sanger.ac.uk/pub4/ibdgenetics/ucmeta-sumstats.txt.gz`
- Summary *p*-values for human height[3] were obtained from the GIANT consortium `https://www.broadinstitute.org/collaboration/giant/images/4/47/GIANT_HEIGHT_LangoAllen2010_publicrelease_HapMapCeuFreq.txt`
- Summary *p*-values for total cholesterol, LDL cholesterol, HDL cholesterol and triglycerides[5] were obtained from the Global Lipids Consortium `http://csg.sph.umich.edu/abecasis/public/lipids2013/`
- Summary *p*-values for systolic and diastolic blood pressure. [15] were obtained from the International Consortium on Blood Pressure study `http://www.georgehretlab.org/icbp_088023401234-9812599.html`

A permissive threshold for trait association of $p < 5 \times 10^{-6}$ was used for whole GWAS / meta-analysis coexpression analyses.

3 Cell type specificity

In order to better understand the pathophysiological implications of disease variants in regulatory regions, we sought to identify whether these regions exhibit unexpectedly specific expression in any given cell types or tissue samples. In order to reduce noise, technical and biological replicates were averaged for this and subsequent analyses. The full table of samples in FANTOM5, showing which samples were averaged as technical replicates, and which were excluded, is in supplementary table (SF2_phenotype_matching.txt).

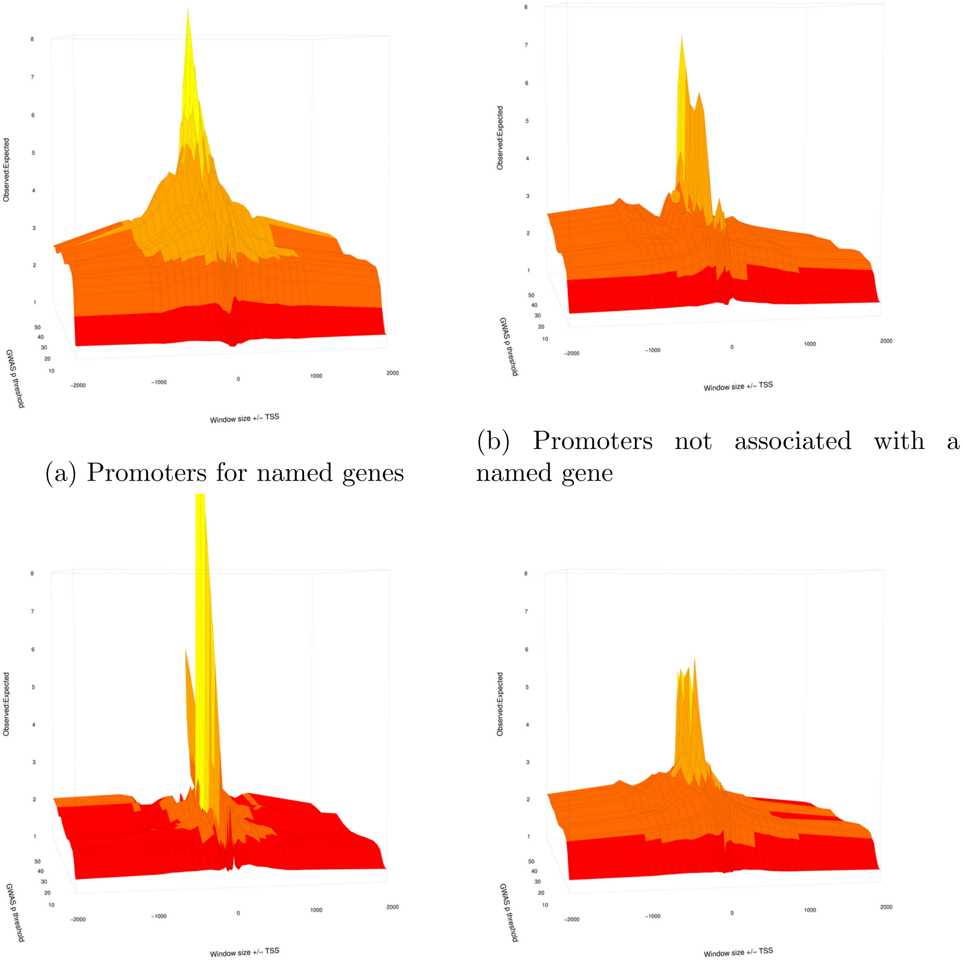
For a given trait, we took the subset of regulatory regions for which a significant coexpression pattern was detected for that trait (coexpression $FDR \leq 0.05$). For each regulatory region, we created a list of all cell types in which that region was active, ranked by expression level. We then combined the cell type lists for each regulatory region using a robust rank aggregation (RRA)[11].

There are several possible sources of bias in this raw measurement. For example, some cell types have more cell-type specific transcriptional activity, perhaps because these cell types fulfil a specialised role; other cell types are particularly well-represented in the FANTOM5 samples. We therefore controlled for the probability that a given cell type would be highly ranked in the initial

RRA analysis, by permuting RRA results for at least 100,000 random selections of n regulatory regions. We then calculated the empirical p-value for a each cell type, i.e. the probability that this cell type would be assigned a raw RRA p-value at least as strong by random chance. We then corrected for multiple comparisons using the Benjamini-Hochberg method to estimate false discovery rate (FDR).

4 Code availability

Computer code required to run the NDA method, specifically for the detection of coexpression in FANTOM5 regulatory regions, can be obtained from https://github.com/baillielab/coexpression/



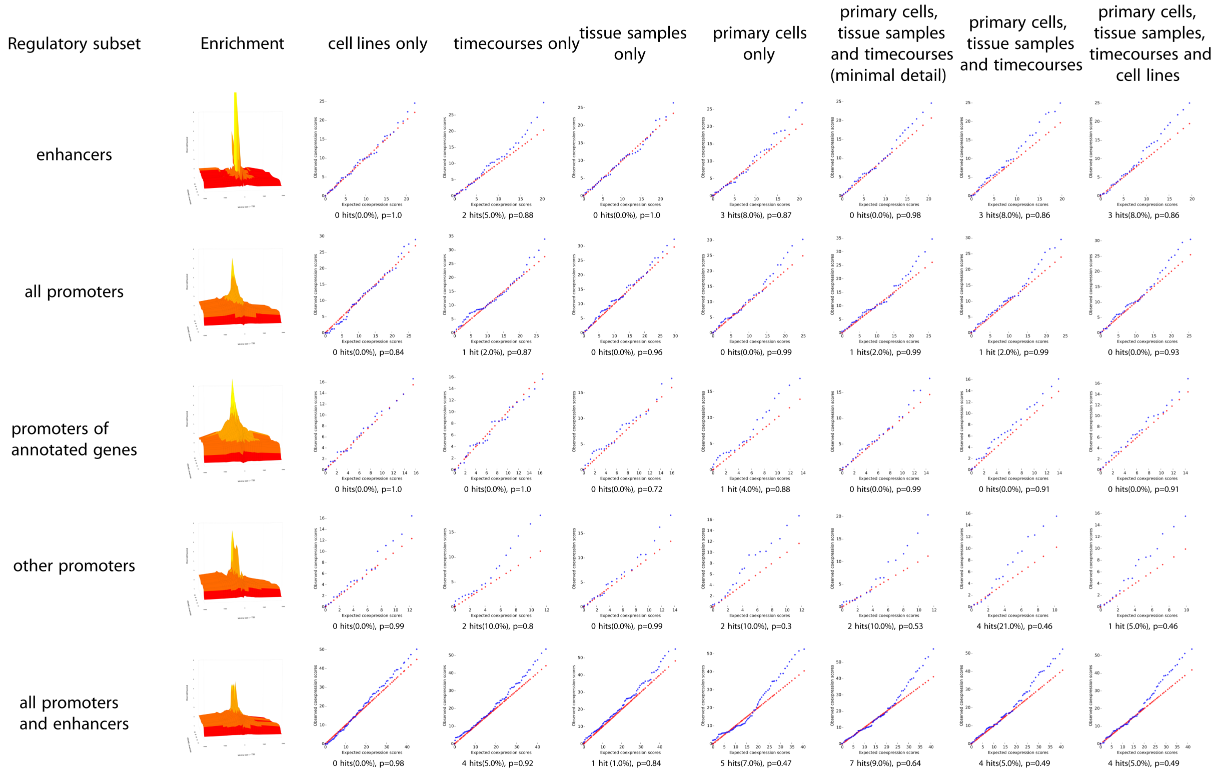
(a) Promoters for named genes

(b) Promoters not associated with a named gene

(c) Enhancers

(d) All promoters and enhancers

Supplementary Figure 2: Enrichment for GWAS hits at increasing distances above and below TSS



Supplementary Figure 3: Change in coexpression signal using different subsets of the FANTOM5 dataset, using the Crohn's disease GWAS as the input set. Enrichment column shows a miniature graph depicting the enrichment (observed $SNPs.Mb^{-1}$: expected $SNPs.Mb^{-1}$) at increasing search window sizes upstream and downstream from the transcription start site (TSS). Other columns show Q:Q plots of observed:expected NDA scores obtained using a given subset of samples (see SF3_sample_averaging.xlsx for full description of each subset). Rows indicate the subset of regulatory regions used in each analysis. Percentage of significantly coexpressed entities (hits, $FDR < 0.05$) and p-value (Kolmogorov-Smirnov test) comparing observed (blue) and expected (red) distributions are shown below each plot.

References

[1] Carl A Anderson, Gabrielle Boucher, Charlie W Lees, Andre Franke, Mauro D'Amato, Kent D Taylor, James C Lee, Philippe Goyette, Marcin Imielinski, Anna Latiano, Caroline Lagac, Regan Scott, Leila Amininejad, Suzannah Bumpstead, Leonard Baidoo, Robert N Baldassano, Murray Barclay, Theodore M Bayless, Stephan Brand, Carsten Bning, Jean-Frdric Colombel, Lee A Denson, Martine De Vos, Marla Dubinsky, Cathryn Edwards, David Ellinghaus, Rudolf S N Fehrmann, James A B Floyd, Timothy Florin, Denis Franchimont, Lude Franke, Michel Georges, Jrgen Glas, Nicole L Glazer, Stephen L Guthery, Talin Haritunians, Nicholas K Hayward, Jean-Pierre Hugot, Gilles Jobin, Debby Laukens, Ian Lawrance, Marc Lmann, Arie Levine, Cecile Libioulle, Edouard Louis, Dermot P McGovern, Monica Milla, Grant W Montgomery, Katherine I Morley, Craig Mowat, Aylwin Ng, William Newman, Roel A Ophoff, Laura Papi, Orazio Palmieri, Laurent Peyrin-Biroulet, Julin Pans, Anne Phillips, Natalie J Prescott, Deborah D Proctor, Rebecca Roberts, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Philip Schumm, Frank Seibold, Yashoda Sharma, Lisa A Simms, Mark Seielstad, A Hillary Steinhart, Stephan R Targan, Leonard H van den Berg, Morten Vatn, Hein Verspaget, Thomas Walters, Cisca Wijmenga, David C Wilson, Harm-Jan Westra, Ramnik J Xavier, Zhen Z Zhao, Cyriel Y Ponsioen, Vibeke Andersen, Leif Torkvist, Maria Gazouli, Nicholas P Anagnou, Tom H Karlsen, Limas Kupcinskas, Jurgita Sventoraityte, John C Mansfield, Subra Kugathasan, Mark S Silverberg, Jonas Halfvarson, Jerome I Rotter, Christopher G Mathew, Anne M Griffiths, Richard Gearry, Tariq Ahmad, Steven R Brant, Mathias Chamaillard, Jack Satsangi, Judy H Cho, Stefan Schreiber, Mark J Daly, Jeffrey C Barrett, Miles Parkes, Vito Annese, Hakon Hakonarson, Graham Radford-Smith, Richard H Duerr, Sverine Vermeire, Rinse K Weersma, and John D Rioux. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics*, 43(3):246–252, 2011.

[2] Yoav Benjamini. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[3] Sonja I. Berndt, Stefan Gustafsson, Reedik Mgi, Andrea Ganna, Eleanor Wheeler, Mary F. Feitosa, Anne E. Justice, Keri L. Monda, Damien C. Croteau-Chonka, Felix R. Day, Tnu Esko, Tove Fall, Teresa Ferreira, Davide Gentilini, Anne U. Jackson, Jian'an Luan, Joshua C. Randall, Sailaja Vedantam, Cristen J. Willer, Thomas W. Winkler, Andrew R. Wood, Tsegaselassie Workalemahu, Yi-Juan Hu, Sang Hong Lee, Liming Liang, Dan-Yu Lin, Josine L. Min, Benjamin M. Neale, Gudmar Thorleifsson, Jian Yang, Eva Albrecht, Najaf Amin, Jennifer L. Bragg-Gresham, Gemma Cadby, Martin den Heijer, Niina Eklund, Krista Fischer, Anuj Goel, Jouke-Jan Hottenga, Jennifer E. Huffman, Ivonne Jarick, sa Johansson, Toby Johnson, Stavroula Kanoni, Marcus E. Kleber, Inke R. Knig, Kati Kristiansson, Zoltn Kutalik, Claudia Lamina, Cecile Lecoeur, Guo Li, Massimo Mangino, Wendy L. McArdle, Carolina Medina-Gomez, Martina Mller-Nurasyid, Julius S. Ngwa, Ilja M. Nolte, Lavinia Paternoster, Sonali Pechlivanis, Markus Perola, Marjolein J. Peters, Michael Preuss, Lynda M. Rose, Jianxin Shi, Dmitry Shungin, Albert Vernon Smith, Rona J. Strawbridge, Ida Surakka, Alexander Teumer, Mieke D. Trip, Jonathan Tyrer, Jana V. Van Vliet-Ostaptchouk, Liesbeth Vandenput, Lindsay L. Waite, Jing Hua Zhao, Devin Absher, Folkert W. Asselbergs, Mustafa Atalay, Antony P. Attwood, Anthony J.

Balmforth, Hanneke Basart, John Beilby, Lori L. Bonnycastle, Paolo Brambilla, Marcel Bruinenberg, Harry Campbell, Daniel I. Chasman, Peter S. Chines, Francis S. Collins, John M. Connell, William O. Cookson, Ulf de Faire, Femmie de Vegt, Mariano Dei, Maria Dimitriou, Sarah Edkins, Karol Estrada, David M. Evans, Martin Farrall, Marco M. Ferrario, Jean Ferrires, Lude Franke, Francesca Frau, Pablo V. Gejman, Harald Grallert, Henrik Grnberg, Vilmundur Gudnason, Alistair S. Hall, Per Hall, Anna-Liisa Hartikainen, Caroline Hayward, Nancy L. Heard-Costa, Andrew C. Heath, Johannes Hebebrand, Georg Homuth, Frank B. Hu, Sarah E. Hunt, Elina Hyppnen, Carlos Iribarren, Kevin B. Jacobs, John-Olov Jansson, Antti Jula, Mika Khnen, Sekar Kathiresan, Frank Kee, Kay-Tee Khaw, Mika Kivimki, Wolfgang Koenig, Aldi T. Kraja, Meena Kumari, Kari Kuulasmaa, Johanna Kuusisto, Jaana H. Laitinen, Timo A. Lakka, Claudia Langenberg, Lenore J. Launer, Lars Lind, Jaana Lindstrm, Jianjun Liu, Antonio Liuzzi, Marja-Liisa Lokki, Mattias Lorentzon, Pamela A. Madden, Patrik K. Magnusson, Paolo Manunta, Diana Marek, Winfried Mrz, Irene Mateo Leach, Barbara McKnight, Sarah E. Medland, Evelin Mihailov, Lili Milani, Grant W. Montgomery, Vincent Mooser, Thomas W. Mhleisen, Patricia B. Munroe, Arthur W. Musk, Narisu Narisu, Gerjan Navis, George Nicholson, Ellen A. Nohr, Ken K. Ong, Ben A. Oostra, Colin N. A. Palmer, Aarno Palotie, John F. Peden, Nancy Pedersen, Annette Peters, Ozren Polasek, Anneli Pouta, Peter P. Pramstaller, Inga Prokopenko, Carolin Ptter, Aparna Radhakrishnan, Olli Raitakari, Augusto Rendon, Fernando Rivadeneira, Igor Rudan, Timo E. Saaristo, Jennifer G. Sambrook, Alan R. Sanders, Serena Sanna, Jouko Saramies, Sabine Schipf, Stefan Schreiber, Heribert Schunkert, So-Youn Shin, Stefano Signorini, Juha Sinisalo, Boris Skrobek, Nicole Soranzo, Alena Stankov, Klaus Stark, Jonathan C. Stephens, Kathleen Stirrups, Ronald P. Stolk, Michael Stumvoll, Amy J. Swift, Eirini V. Theodoraki, Barbara Thorand, David-Alexandre Tregouet, Elena Tremoli, Melanie M. Van der Klauw, Joyce B. J. van Meurs, Sita H. Vermeulen, Jorma Viikari, Jarmo Virtamo, Veronique Vitart, Grard Waeber, Zhaoming Wang, Elisabeth Widn, Sarah H. Wild, Gonneke Willemsen, Bernhard R. Winkelmann, Jacqueline C. M. Witteman, Bruce H. R. Wolffenbuttel, Andrew Wong, Alan F. Wright, M. Carola Zillikens, Philippe Amouyel, Bernhard O. Boehm, Eric Boerwinkle, Dorret I. Boomsma, Mark J. Caulfield, Stephen J. Chanock, L. Adrienne Cupples, Daniele Cusi, George V. Dedoussis, Jeanette Erdmann, Johan G. Eriksson, Paul W. Franks, Philippe Froguel, Christian Gieger, Ulf Gyllensten, Anders Hamsten, Tamara B. Harris, Christian Hengstenberg, Andrew A. Hicks, Aroon Hingorani, Anke Hinney, Albert Hofman, Kees G. Hovingh, Kristian Hveem, Thomas Illig, Marjo-Riitta Jarvelin, Karl-Heinz Jckel, Sirkka M. Keinanen-Kiukaanniemi, Lambertus A. Kiemeney, Diana Kuh, Markku Laakso, Terho Lehtimki, Douglas F. Levinson, Nicholas G. Martin, Andres Metspalu, Andrew D. Morris, Markku S. Nieminen, Inger Njlstad, Claes Ohlsson, Albertine J. Oldehinkel, Willem H. Ouwehand, Lyle J. Palmer, Brenda Penninx, Chris Power, Michael A. Province, Bruce M. Psaty, Lu Qi, Rainer Rauramaa, Paul M. Ridker, Samuli Ripatti, Veikko Salomaa, Nilesh J. Samani, Harold Snieder, Thorkild I. A. Srensen, Timothy D. Spector, Kari Stefansson, Anke Tnjes, Jaakko Tuomilehto, Andr G. Uitterlinden, Matti Uusitupa, Pim van der Harst, Peter Vollenweider, Henri Wallaschofski, Nicholas J. Wareham, Hugh Watkins, H.-Erich Wichmann, James F. Wilson, Goncalo R. Abecasis, Themistocles L. Assimes, Ins Barroso, Michael Boehnke, Ingrid B. Borecki, Panos Deloukas, Caroline S. Fox, Timothy Frayling, Leif C. Groop, Talin Haritunian, Iris M. Heid, David Hunter, Robert C. Kaplan, Fredrik Karpe, Miriam F. Moffatt, Karen L. Mohlke, Jeffrey R. O'Connell, Yudi Pawitan, Eric E. Schadt, David Schlessinger, Valgerdur Steinthors-

dottir, David P. Strachan, Unnur Thorsteinsdottir, Cornelia M. van Duijn, Peter M. Visscher, Anna Maria Di Blasio, Joel N. Hirschhorn, Cecilia M. Lindgren, Andrew P. Morris, David Meyre, Andr Scherag, Mark I. McCarthy, Elizabeth K. Speliotes, Kari E. North, Ruth J. F. Loos, and Erik Ingelsson. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics*, 45(5):501–512, 2013.

[4] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Par G Engstrom, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine A Wells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume, and Yoshihide Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635, 2006.

[5] Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, 45(11):1274–1283, 2013.

[6] Kawaji H. Rehli M.-Baillie J.K. et al Forrest, A. R. R. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, 2014.

[7] Andre Franke, Dermot P B McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, Carl A Anderson, Joshua C Bis, Suzanne Bumpstead, David Ellinghaus, Eleonora M Festen, Michel Georges, Todd Green, Talin Haritunians, Luke Jostins, Anna Latiano, Christopher G Mathew, Grant W Montgomery, Natalie J Prescott, Soumya Raychaudhuri, Jerome I Rotter, Philip Schumm, Yashoda Sharma, Lisa A Simms, Kent D Taylor, David Whiteman, Cisca Wijmenga, Robert N Baldassano, Murray Barclay, Theodore M Bayless, Stephan Brand, Carsten Bning, Albert Cohen, Jean-Frederick Colombel, Mario Cottone, Laura Stronati, Ted Denson, Martine De Vos, Renata D'Inca, Marla Dubinsky, Cathryn Edwards, Tim Florin, Denis Franchimont, Richard Gearry, Jrgen Glas, Andre Van Gossum, Stephen L Guthery, Jonas Halfvarson, Hein W Verspaget, Jean-Pierre Hugot, Amir Karban, Debby Laukens, Ian Lawrance, Marc Lemann, Arie Levine, Cecile Libioulle, Edouard Louis, Craig Mowat, William Newman, Julin Pans, Anne Phillips, Deborah D Proctor, Miguel Regueiro, Richard Russell, Paul Rutgeerts, Jeremy Sanderson, Miquel Sans, Frank Seibold, A Hillary Steinhart, Pieter C F Stokkers, Leif Torkvist, Gerd Kullak-Ublick, David Wilson, Thomas Walters, Stephan R Targan, Steven R Brant, John D Rioux, Mauro D'Amato, Rinse K Weersma, Subra Kugathasan, Anne M Griffiths, John C Mansfield, Severine Vermeire, Richard H Duerr, Mark S Silverberg, Jack Satsangi, Stefan Schreiber, Judy H Cho, Vito Annese, Hakon Hakonarson, Mark J Daly, and Miles Parkes. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature Genetics*, 42(12):1118–1125, 2010.

[8] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional impli-

cations of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–9367, 2009.

[9] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The ucsc genome browser database: update 2006. *Nucleic Acids Research*, 34(suppl 1):D590–D598, 2006.

[10] Tae-Kyung Kim, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F. Worley, Gabriel Kreiman, and Michael E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, 2010.

[11] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

[12] Boris Lenhard, Albin Sandelin, and Piero Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature reviews. Genetics*, 13(4):233–245, 2012.

[13] Mulin Jun Li, Panwen Wang, Xiaorong Liu, Ee Lyn Lim, Zhangyong Wang, Meredith Yeager, Maria P. Wong, Pak Chung Sham, Stephen J. Chanock, and Junwen Wang. Gwasdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Research*, 40(D1):D1047–D1054, 2012.

[14] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[15] The International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–109, 2011.

[16] Gosia Trynka, Karen A. Hunt, Nicholas A. Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F. Bakker, Maria Teresa Bardella, Leena Bhaw-Rosun, Gemma Castillejo, Emilio G. de la Concha, Rodrigo Coutinho de Almeida, Kerith-Rae M. Dias, Cleo C. van Diemen, Patrick C. A. Dubois, Richard H. Duerr, Sarah Edkins, Lude Franke, Karin Fransen, Javier Gutierrez, Graham A. R. Heap, Barbara Hrdlickova, Sarah Hunt, Leticia Plaza Izurieta, Valentina Izzo, Leo A. B. Joosten, Cordelia Langford, Maria Cristina Mazzilli, Charles A. Mein, Vandana Midah, Mitja Mitrovic, Barbara Mora, Marinita Morelli, Sarah Nutland, Concepcin Nez, Suna Onengut-Gumuscu, Kerra Pearce, Mathieu Platteel, Isabel Polanco, Simon Potter, Carmen Ribes-Koninckx, Isis Ricao-Ponce, Stephen S. Rich, Anna Rybak, Jos Luis Santiago, Sabyasachi Senapati, Ajit Sood, Hania Szajewska, Riccardo Troncone, Jezabel Varad, Chris Wallace, Victorien M. Wolters, Alexandra Zhernakova, Spanish Consortium on the Genetics of Coeliac Disease (cegec), PreventCD Study Group, Wellcome Trust Case Control Consortium (wtccc), B. K. Thelma,

Bozena Cukrowska, Elena Urcelay, Jose Ramon Bilbao, M. Luisa Mearin, Donatella Barisani, Jeffrey C. Barrett, Vincent Plagnol, Panos Deloukas, Cisca Wijmenga, and David A. van Heel. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics*, 43(12):1193–1201, 2011.

[17] Benjamin F. Voight, Hyun Min Kang, Jun Ding, Cameron D. Palmer, Carlo Sidore, Peter S. Chines, Nol P. Burtt, Christian Fuchsberger, Yanming Li, Jeanette Erdmann, Timothy M. Frayling, Iris M. Heid, Anne U. Jackson, Toby Johnson, Tuomas O. Kilpelinen, Cecilia M. Lindgren, Andrew P. Morris, Inga Prokopenko, Joshua C. Randall, Richa Saxena, Nicole Soranzo, Elizabeth K. Speliotes, Tanya M. Teslovich, Eleanor Wheeler, Jared Maguire, Melissa Parkin, Simon Potter, N. William Rayner, Neil Robertson, Kathleen Stirrups, Wendy Winckler, Serena Sanna, Antonella Mulas, Ramaiah Nagaraja, Francesco Cucca, Ins Barroso, Panos Deloukas, Ruth J. F. Loos, Sekar Kathiresan, Patricia B. Munroe, Christopher Newton-Cheh, Arne Pfeufer, Nilesh J. Samani, Heribert Schunkert, Joel N. Hirschhorn, David Altshuler, Mark I. McCarthy, Gonalo R. Abecasis, and Michael Boehnke. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*, 8(8):e1002793, 2012.