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

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

Published in:
PLoS Computational Biology

DOI:
10.1371/journal.pcbi.1005934

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):

Download date: 25. feb., 2020
Shared activity patterns arising at genetic susceptibility loci reveal underlying genomic and cellular architecture of human disease

J. Kenneth Baillie¹,²,³ *, Andrew Bretherrick¹, Christopher S. Haley¹,⁴, Sara Clohisey¹, Alan Gray⁵, Lucile P. A. Neyton⁵, Jeffrey Barrett⁶, Eli A. Stahl⁷, Albert Tenesa¹, Robin Andersson⁸, J. Ben Brown⁹, Geoffrey J. Faulkner¹⁰, Marina Lizlo¹¹, Ulf Schaefer¹², Carsten Daub¹³, Masayoshi Itoh¹¹,¹³, Naoto Kondo¹¹, Timo Lassmann¹¹, Jun Kawai¹¹, IIBDGC Consortium¹¹, Damjen Mole¹², Vladimir B. Bajic¹⁴, Peter Heutink¹⁵, Michael Rehli¹⁶, Hideya Kawaji¹¹,¹³, Albin Sandelin⁶, Harukazu Suzuki¹¹, Jack Satsangi¹⁶, Christine A. Wells¹⁷, Nir Hacohen¹⁸, Thomas C. Freeman¹, Yoshihade Hayashizaki¹¹, Piero Carninci¹¹, Alistair R. R. Forrest¹⁹ *, David A. Hume¹¹,¹⁰ *

¹ Division of Genetics and Genomics, The Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom, ² Centre for Inflammation Research, University of Edinburgh, Edinburgh, United Kingdom, ³ Intensive Care Unit, Royal Infirmary Edinburgh, Edinburgh, United Kingdom, ⁴ Institute for Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom, ⁵ Edinburgh Parallel Computing Centre, The University of Edinburgh, Edinburgh, United Kingdom, ⁶ Statistical Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, ⁷ Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, United States of America, ⁸ The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen, Denmark, ⁹ Department of Statistics, University of California, Berkeley, United States of America, ¹⁰ Mater Research Institute, University of Queensland, University of Queensland, Brisbane, Australia, ¹¹ RIKEN Omics Science Center, Yokohama, Japan, Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan, ¹² Department for Infectious Disease Informatics, Public Health England, Colindale, United Kingdom, ¹³ RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Japan, ¹⁴ King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center, Thuwal, Kingdom of Saudi Arabia, ¹⁵ German Center for Neurodegenerative Diseases, Tübingen, Germany, ¹⁶ Dept. Hematology, University Hospital Regensburg, Regensburg, Germany, ¹⁷ Australian Institute for Bioengineering and Nanotechnology, University of Queensland, St Lucia, Brisbane Australia, ¹⁸ Broad Institute of Harvard and MIT, Cambridge, United States of America, ¹⁹ Harry Perkins Institute of Medical Research, and the Centre for Medical Research, University of Western Australia, QEII Medical Centre, Nedlands, Perth, Western Australia, Australia

* Members of the IIBDGC Consortium are listed in the Acknowledgments.

© 2018 Baillie et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The full datasets used for this work are available from the FANTOMS data interface at fantom.gsc.riken.jp/zenbu/.

Funding: 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 Grants to the Roslin Institute (BBS/E/D/20211551, BBS/E/D/20211552, BBS/E/D/20211553, BBS/E/D/20231760), the 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 similar patterns of transcriptional activity. Accordingly, shared transcriptional activity (coexpression) may help prioritise loci associated with a given trait, and help to identify underlying biological processes. Using cap analysis of gene expression (CAGE) profiles of promoter- and enhancer-derived RNAs across 1824 human samples, we have analysed coexpression of RNAs originating from trait-associated regulatory regions using a novel quantitative method (network density analysis; NDA). For most traits studied, phenotype-associated variants in regulatory regions were linked to tightly-coexpressed networks that are likely to share important functional characteristics. Coexpression provides a new signal, independent of
Author summary

We discover that genetic variants associated with specific diseases have more in common with each other than we have previously seen. Specifically, variants associated with the same disease tend to be in parts of the genome that are turned on or off in similar complex patterns across many different cell types. We discover that genetic variants associated with specific diseases are found within regulatory elements that share patterns of expression. Specifically, variants associated with the same disease tend to be in parts of the genome that are turned on or off together in similar complex patterns across many different cell types. Knowing this helps us to find new variants associated with some diseases, and to better understand the genetic causes of other diseases. Furthermore, we discover that the genetic causes of inflammatory bowel disease fall into two distinct patterns, indicating that two aetiologically-distinct endotypes of this condition exist. Unlike other methods to learn about disease mechanisms from genetic information, our approach does not require any knowledge or assumptions about the genes themselves—it depends only on the patterns in which parts of the genome are activated in different cell types.

Introduction

Genome-wide association studies (GWAS) have considerable untapped potential to reveal new mechanisms of disease[1]. Variants associated with disease are 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, IncRNAs 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 some GWAS is substantially
enriched in these FANTOM5 promoters[9][10].

Genes that are coexpressed are more likely to share common biology[11,12]. Similarly, reg-
ulatory regions that share activity patterns are more likely to contribute to the same biological
pathways[5]. We have previously shown 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]. Informative regulatory networks can be derived
from predicted transcription factor interactions with FANTOM5 regulatory regions[6]. We
therefore use transcript expression here as a surrogate for transcriptional regulatory activity.

In contrast to previous studies[6,13,14], we sought to explore the similarities in activity at
disease-associated sets of regulatory regions, rather than genes, and independent of transcrip-
tion factor binding predictions.

In order to determine whether coexpression of regulatory elements 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 pro-
moter and enhancer-associated transcript levels measured by CAGE across numerous cell
types and tissues, Fig 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 expressed regulatory region that contains a GWAS SNP associated with a given trait is
assigned a score quantifying its proximity in the network to every other regulatory region con-
taining 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 phe-
notypes, thereby providing appropriate cell culture models for critical disease processes.

Methods

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[15]. Previous analy-
sis demonstrated that this covers the areas of maximal sequence conservation across species
and the core region of transcription factor binding. Enhancers are widely transcribed across
the human genome (eRNAs). 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.

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 matrix was then generated from 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 share activity patterns, we devised the NDA method, which quantifies the strength of coexpression among a chosen pool of putative phenotype-associated regulatory regions. 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.

**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. Spearman’s rank correlation was chosen to quantify pairwise relationships, in view of the robustness of this measure in a variety of different distributions. 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 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.

**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 $TSS_{\text{start}}$ equal the base pair index 300bp (promoters) or 0bp (enhancers) upstream from a FANTOM5 transcription start site (TSS) and $TSS_{\text{end}}$ 100bp (promoters) or 0bp (enhancers) downstream, the set, $P$, of putative trait-associated promoters is given by:

$$P = \{ i : i \in K, TSS_{\text{start}} - 300 \leq i \leq TSS_{\text{end}} + 100 \}$$

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

$$E = \{ i : i \in K, TSS_{\text{start}} \leq i \leq TSS_{\text{end}} \}$$

giving a total set of regulatory regions:

$$R = P \cup E$$

**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 region 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 spurious coexpression signals are not being generated solely 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.

### Coexpression matrix

$A$ is defined as 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 as, or stronger than, 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 $A$.

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

### 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.

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

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 comparison of coexpression scores between different analyses, the raw coexpression score ($s$) is corrected by dividing by the total number of independent groups of regulatory regions included in each analysis, $n_{res}$ yielding a corrected coexpression score, $ccs$:

$$ccs = \frac{s}{n_{res}}$$

### 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 strongest 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.

**Input datasets**

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

**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; 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.

**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. 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.

**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. This is analogous to randomly reassigning the labels in the network, but aims to preserve the local relationships between regulatory regions, since we cannot assume that regulatory regions are randomly distributed on the genome, and since regional regulatory events, such as chromatin reorganisation, are expected to lead to coexpression between nearby regulatory regions.

Where $A$ is defined as a list of regulatory regions comprising the whole set of FANTOM5 TSS, post-mapping permutations select a subset of $A$ in a similar circular manner, by displacing the members of the set $R$ 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 to the input set. If the input set contains a large group of TSS regions in close proximity to each other 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, potentially inflating the NDA scores in the permuted sets. To mitigate against this, those TSS from each permutation that do 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.

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

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 per Mb: expected SNPs per Mb) 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 and from a larger, less selective catalog of GWAS p-values meeting permissive criteria for genome-wide significance, GWASdb. The GWASdb dataset is less 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. Because 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](http://www.genome.gov/gwastudies)

Overlapping phenotypes, such as "urate" and "uric acid" were manually merged. 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 (S2 Table).

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.
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. These were obtained from the following sources:

- Crohn’s disease 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 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 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 were obtained from the Global Lipids Consortium http://csg.sph.umich.edu/abecasis/public/lipids2013/
- Summary $p$-values for systolic and diastolic blood pressure were obtained from the International Consortium on Blood Pressure study http://www.georgehretlab.org/icbp_088023401234-9812599.html

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 S2 Table (S2 Table).

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).

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 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 ($FDR$).

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/
Results

Evaluation of the NDA method and FANTOM5 input dataset

Our initial evaluation demonstrated that coexpression is stronger among regulatory regions containing variants with low GWAS p-values (Fig 2). The coexpression signal obtained for the test input set was evaluated using different subsets of FANTOM5 samples (cell lines, time-courses 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) (Fig 3). 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 (Fig 3).

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

![Optimisation of GWAS p-value threshold](https://doi.org/10.1371/journal.pcbi.1005934.g002)
Fig 3. (a) Enrichment (y axis, observed SNPs per Mb: expected SNPs per Mb) at increasing search window sizes (x axis) upstream and downstream from the transcription start site (TSS) for increasingly strong GWAS signals (z axis, $-\log_{10} p$).

(b) Change in coexpression signal using different subsets of the FANTOM5 dataset, using the Crohn’s disease GWAS as the input set. Q-Q plots of observed:expected NDA scores obtained using a given subset of samples (see methods 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.

https://doi.org/10.1371/journal.pcbi.1005934.g003
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 Fig 4, which reveals the NDA score to be a composite measure of both signals.

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. 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. 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 results).

Fig 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 SNPs 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 high-density lipoprotein (HDL), low-density lipoprotein (LDL), and total cholesterol. See supplementary results for full results of all analyses.

https://doi.org/10.1371/journal.pcbi.1005934.g004
Fine mapping

Regulatory regions around individual TSS with higher coexpression scores contain variants with stronger GWAS p-values (Fig 5A), indicating that this independent signal provides additional information that may be used for fine-mapping causative loci (Fig 6; Supplementary results). Where data are available, we have compared our results to the recent fine mapping study by Huang et al, who use high-resolution genotyping in 67,852 subjects with inflammatory bowel disease to quantify the probability that a given variant is causal. A total of 9 variants with a causal probability > 0.1 lie within 150,000bp of a significantly coexpressed region; of these, 7 lie immediately adjacent to the most significantly coexpressed promoter/enhancer in the region.

Discovery and prioritisation of GWAS hits in regulatory sequence

In order to enable the detection of new regulatory regions with strong coexpression relationships, we chose a permissive threshold at GWAS $p < 5 \times 10^{-6}$. GWAS data for Crohn’s disease [16] were used for initial optimisation of the NDA approach. Of the 8 GWAS datasets for phenotypes that were not used in algorithm development (i.e. all apart from Crohn’s disease), 6 showed evidence of significant coexpression (Table 1). Among these, between 17 and 24% of regulatory regions identified as containing a GWAS SNP were found to be significantly coexpressed with other regulatory elements associated with the same phenotype (FDR < 0.05, compared with 100 permuted subsets of equal size; see Methods).

Although many coexpressed regulatory regions are not promoters for annotated genes (supplementary results; Fig 3), we compared the named genes in our results with gene-level burden of significance in original GWAS studies (PASCAL sum genescore). Log rank is shown on each axis (Rank 1 = highest scoring gene) for the subset of coexpression scores obtained for promoters of named genes. Open squares indicate significant coexpression (FDR<0.05).

Fig 5. (a) Relationship between GWAS p-value for a SNP, and coexpression scores of individual promoters assigned to that SNP for all phenotypes for which significant coexpression was detected. 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. (b) Rank comparison of named genes compared with gene-level burden of significance in original GWAS studies (PASCAL sum genescore). Log rank is shown on each axis (Rank 1 = highest scoring gene) for the subset of coexpression scores obtained for promoters of named genes. Open squares indicate significant coexpression (FDR<0.05).

https://doi.org/10.1371/journal.pcbi.1005934.g005

Coexpression of GWAS hits
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. Fig 7A shows significant coexpression superimposed on a two-dimensional representation of the entire network of pairwise correlations. Since activity (transcript expression) was measured in many samples, the true proximity of regulatory regions to one another cannot be accurately
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).

<table>
<thead>
<tr>
<th>Trait</th>
<th>SNPs included, p&lt;5e-6 (SNPs per million bases)</th>
<th>Regulatory regions containing a SNP (SNPs per million bases)</th>
<th>Fold enrichment for SNPs in regulatory regions</th>
<th>Distinct regulatory regions</th>
<th>Significantly coexpressed TSS (FDR&lt;0.05)(% of distinct regions)</th>
<th>p (KS test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crohn’s disease</td>
<td>1924 (0.6)</td>
<td>133 (3.5)</td>
<td>5.7</td>
<td>70</td>
<td>23 (33%)</td>
<td>1.61e-05</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>2162 (0.7)</td>
<td>146 (3.8)</td>
<td>5.5</td>
<td>83</td>
<td>20 (24%)</td>
<td>2.28e-06</td>
</tr>
<tr>
<td>LDL</td>
<td>4644 (1.5)</td>
<td>205 (5.2)</td>
<td>3.5</td>
<td>92</td>
<td>19 (21%)</td>
<td>1.48e-04</td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>6421 (2.0)</td>
<td>316 (8.3)</td>
<td>4.1</td>
<td>128</td>
<td>29 (23%)</td>
<td>6.55e-07</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>4863 (1.5)</td>
<td>254 (7.0)</td>
<td>4.6</td>
<td>97</td>
<td>23 (24%)</td>
<td>8.35e-06</td>
</tr>
<tr>
<td>Height</td>
<td>8882 (2.8)</td>
<td>358 (7.6)</td>
<td>2.7</td>
<td>166</td>
<td>29 (17%)</td>
<td>1.25e-06</td>
</tr>
<tr>
<td>HDL</td>
<td>5410 (1.7)</td>
<td>260 (7.2)</td>
<td>4.2</td>
<td>101</td>
<td>17 (17%)</td>
<td>3.51e-04</td>
</tr>
<tr>
<td>SBP</td>
<td>417 (0.1)</td>
<td>20 (0.4)</td>
<td>3.0</td>
<td>13</td>
<td>0 (0%)</td>
<td>4.89e-01</td>
</tr>
<tr>
<td>DBP</td>
<td>711 (0.2)</td>
<td>20 (0.4)</td>
<td>1.9</td>
<td>14</td>
<td>0 (0%)</td>
<td>5.41e-01</td>
</tr>
</tbody>
</table>

KS test: Kolmogorov-Smirnov test comparing distribution of coexpression scores for this study with permuted values. *Initial optimisation and parameterisation of the algorithm was undertaken using a random subset of data from this study.

https://doi.org/10.1371/journal.pcbi.1005934.t001

represented in two dimensions—a perfect representation would require as many dimensions as there are unique samples. In contrast, the NDA method quantifies proximity of regulatory regions in true network space without artificial dimensionality reduction. Thus significantly

Fig 7. Examples of detail of chromosomal regions surrounding regulatory regions significantly coexpressed in ulcerative colitis (STSS+/−150MbS). (a) Region surrounding IL10 (b) Region surrounding C1orf106. Top panel: Coloured rectangles show genomic location of individual regulatory regions (promoters or enhancers). Height of regulatory regions on y-axis depicts the coexpression score assigned to this regulatory region; groups of regulatory regions considered as a single unit (see Methods) share the same colour. Black circles show GWAS Sp±-values for individual SNPs. Red circles show causative probabilities estimated by Huang et al. (2017) for specific variants, where available. Bottom panel: genomic locations of known protein coding transcripts in sense (green) and antisense (purple).

https://doi.org/10.1371/journal.pcbi.1005934.g007
coexpressed elements are detected even if they are not directly adjacent on a two-dimensional representation of the network (Fig 7).

We saw no evidence of spurious coexpression due to genomic proximity with shared regulatory influences (see Methods). In each of the GWAS analyses for which significant coexpression was detected, strong coexpression links were seen between loci that were widely separated on the genome (Fig 4; supplementary results).

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[18] and ABCG5[19], but also from regulatory regions not previously associated with cholesterol levels. A promoter for SLC22A1, which encodes an organic cation transporter, OCT1[20], is strongly coexpressed among elements associated with LDL and TC (Supplementary results). OCT1 transcription is regulated by cholesterol[21] and the transporter regulates hepatic steatosis through its role in thiamine transport[22]. This action of OCT1 is inhibited by metformin[22], an oral hypoglycaemic agent whose cholesterol-lowering effect[23] is not well understood[24]. Full results of coexpression analyses are in the supplementary results, and online at http://baillielab.net/coexpression.

**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[25]. 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[4,6,26]. Signals that are strong enough to be detected in previous, less powerful studies are highly significant in our analysis; for example genetic loci associated with cholesterol are transcriptionally active in hepatocytes and liver tissue[6](Supplementary results).

**Discussion**

The development of high-throughput genotyping methods has led to an explosion of associations between genetic markers and human diseases[27]. 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[28]. 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[14,29–32]. Our 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 [29], annotated biological pathways[33], or gene sets[32,34]. 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[35].

Results for Crohn’s disease and ulcerative colitis were compared to the report by Huang et al [15], who used high-resolution genotyping in a large cohort, together with publicly-available functional genomics data, to identify immune cell signatures implicated in Crohn’s disease, and gut-specific cell types in ulcerative colitis. Our analysis, conducted in parallel and without knowledge of these findings, discovered the same associations, but goes further. Firstly, we demonstrate with a higher level of statistical confidence that these cell type associations are real (supplementary results). This is important in itself, because it 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 [36], whereas Crohn’s disease is a multisystem autoimmune disorder with more diverse extraintestinal manifestations[37], consistent with a primary innate immune aetiology affecting monocyte-macrophage differentiation and response to micro-organisms[38].

Secondly, our analysis extends current knowledge by revealing two distinct groups of significantly-coexpressed regulatory regions for each of these diseases, with differing expression profiles. For Crohn’s disease, one group is restricted to immune cells, particularly monocytes exposed to inflammatory stimuli, while another group of regulatory regions is active in epithelial cells. 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 (Fig 7; S5 Table 1.2). Based on the fundamental assumption of coexpression—that expression profile relates to function—we interpret this as evidence that two distinct biological processes are implicated in each of these diseases. This may be because a "two-hit" mechanism is required for disease pathogenesis. Alternatively these distinct processes may indicate genetically- (and hence aetiologically-) distinct sub-syndromes, or disease endotypes[39], within both Crohn’s disease and ulcerative colitis.

In either case the predominance of each process in an individual patient is likely to have therapeutic relevance. For example, the highly variable clinical response to immunomodulatory therapies, such as methotrexate[40] or anti-TNF monoclonal antibodies[41], may be influenced by the burden of disease-associated variants in Crohn’s disease Group 1 (Fig 7). This represents a conceptually new application of network theory to the detection of disease endotypes, and is likely to have more direct clinical consequences than other methods[42].

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 enable detection of subtle coexpression signals for regulatory regions that have previously been undetectable. The NDA approach developed here enables the identification of cell types and regulatory regions implicated in disease pathogenesis, and contributes a new independent signal to fine mapping of
causative loci. As additional genetic studies become available at greater genotyping resolution, we anticipate that this method will detect new genetic associations with disease and coexpressed modules underlying pathogenesis. The NDA method will enable the identification of critical cell types and processes implicated in mechanisms of disease, and enable further genetic stratification of disease endotypes by underlying mechanism.

Data access

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

An online service running the coexpression method is available at http://baillielab.net/coexpression

Code delivering the NDA coexpression method is available at https://github.com/baillielab/coexpression

Supporting information

S1 Table. Network density analysis results for each of the GWAS studies included here.
(XLSX)

S2 Table. Table of phenotypes in GWAS catalog and GWASdb showing combinations of similar/identical phenotypes for use in regional enrichment calculations.
(TXT)

S3 Table. Table of FANTOM5 samples showing combinations of similar samples that were averaged for use in coexpression network.
(XLSX)

S4 Table. Table showing fine mapping for every region included in the analyses described in this paper (see Fig 6 in main manuscript for legend).
(PDF)

S5 Table. Tables of network density analysis results for each GWAS studies included here.
(PDF)

Acknowledgments

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. We would like to thank GeNAS for data production.


The members of the International IBD Genetics Consortium are listed here: Murray Barclay, Laurent Peyrin-Biroulet, Mathias Chamaillard, Jean-Frederick Colombel, Mario Cottone, Anthony Croft, Renata D’Inca, Jonas Halfvarson, Katherine Hanigal, Paul Henderson, Jean-Pierre Hugot, Amir Karban, Nicholas A Kennedy, Mohammed Azam Khan, Marc Léman, Arie Levine, Dunecan Massey, Monica Milla, Grant W Montgomery, Sok Meng Evelyn Ng, Ioannis Oikonomou, Harald Peeters, Deborah D. Proctor, Jean-Francois Rahier, Rebecca Rob-erts, Paul Rutgeerts, Frank Seibold, Laura Stronati, Kirstin MTaylor, Leif Törkvist, Kullak Ublick, Johan Van Limbergen, Andre VanGossum, Morten H. Vatn, Hu Zhang, Wei Zhang, Australia and New Zealand IBDGCJane M. Andrews, Peter A. Bamption, Murray Barclay,
Coexpression of GWAS hits

Author Contributions

Conceptualization: J. Kenneth Baillie.


Validation: J. Kenneth Baillie, Eli A. Stahl, David A. Hume.

Visualization: J. Kenneth Baillie, Andrew Bretherick, Sara Clohisey, Lucile P. A. Neyton, David A. Hume.

Writing – original draft: J. Kenneth Baillie.

Writing – review & editing: J. Kenneth Baillie, Christopher S. Haley, Sara Clohisey, Alan Gray, Geoffrey J. Faulkner, Marina Lizio, Damian Mole, Michael Rehli, Albin Sandelin,
References


