archive ouverte UNIGE
http://archive-ouverte.unige.ch

Article

Estimating the causal tissues for complex traits and diseases

ONGEN, Halit, *et al*.

Reference

ONGEN, Halit, *et al*. Estimating the causal tissues for complex traits and diseases. *Nature Genetics*, 2017, vol. 49, no. 12, p. 1676-1683

PMID : 29058715
DOI : 10.1038/ng.3981

Available at:
http://archive-ouverte.unige.ch/unige:112643

Disclaimer: layout of this document may differ from the published version.

UNIVERSITÉ DE GENÈVE

nature genetics

Estimating the causal tissues for complex traits and diseases

Halit Ongen[1–3], Andrew A Brown[1–3], Olivier Delaneau[1–3], Nikolaos I Panousis[1–3], Alexandra C Nica[1], GTEx Consortium[4] & Emmanouil T Dermitzakis[1–3]

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

How to interpret the biological causes underlying the predisposing markers identified through genome-wide association studies (GWAS) remains an open question. One direct and powerful way to assess the genetic causality behind GWAS is through analysis of expression quantitative trait loci (eQTLs). Here we describe a new approach to estimate the tissues behind the genetic causality of a variety of GWAS traits, using the *cis*-eQTLs in 44 tissues from the Genotype-Tissue Expression (GTEx) Consortium. We have adapted the regulatory trait concordance (RTC) score to measure the probability of eQTLs being active in multiple tissues and to calculate the probability that a GWAS-associated variant and an eQTL tag the same functional effect. By normalizing the GWAS–eQTL probabilities by the tissue-sharing estimates for eQTLs, we generate relative tissue-causality profiles for GWAS traits. Our approach not only implicates the gene likely mediating individual GWAS signals, but also highlights tissues where the genetic causality for an individual trait is likely manifested.

Over the last decade, GWAS have become the norm in describing genetic variants associated with common complex human diseases and traits[1,2]. Although an impressive number of GWAS findings have been accumulated, the vast majority of the variants identified lie in the noncoding genome[3], rendering their biological interpretation difficult. Furthermore, GWAS identify genetic markers associated with organismal traits and fail to pinpoint the specific tissues underlying these associations[4]. Regulatory variants, such as eQTLs, identified in multiple tissues could aid greatly in the interpretation of GWAS results, not only by linking the noncoding genome to genes but also by identifying the causal tissues behind the genetic associations[5–7]. The GTEx project was founded with the intention of characterizing eQTLs across multiple tissues[8] and currently comprises 44 tissues from 449 individuals (70–361 samples per tissue) for a total of 7,051 transcriptomes (**Supplementary Fig. 1**). This makes GTEx an ideal data set in which to determine the identity of the tissues from which the genetic causality of a GWAS trait arises. Here we aimed to take advantage of this opportunity by first assessing the sharing of eQTLs across tissues (the probability of an eQTL identified in one tissue being active in other tissues) on an individual variant basis and then using these estimates of tissue sharing to infer in which tissues, among the 44 GTEx tissues, GWAS variants likely exert their functions.

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [2]Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland. [3]Swiss Institute of Bioinformatics, Geneva, Switzerland. [4]A list of members and affiliations appears in the **Supplementary Note**. Correspondence should be addressed to H.O. (halit.ongen@unige.ch) or E.T.D. (emmanouil.dermitzakis@unige.ch).

Received 8 February 2016; accepted 29 September 2017; published online 23 October 2017; doi:10.1038/ng.3981

RESULTS

Tissue specificity of eQTLs in the 44 GTEx tissues

For a given eQTL discovered in one tissue, we wanted to derive the probability that this eQTL was active in each of the other 43 tissues. Previously, methods have been described for joint eQTL discovery across multiple tissues[9], assessment of the tissue specificity of eQTLs by integrating orthogonal data from biochemically active regions of the genome in different cell types[10], and eQTL discovery using gene networks[11]; however, in this study, we aimed to quantify the probability of two eQTLs, discovered separately in different tissues and that colocalize, tagging the same underlying functional effect. We have previously described the RTC score, which quantifies the extent to which a colocalizing GWAS variant and eQTL (two variants located in the same genomic region delimited by recombination hotspots) tag the same functional variant[12] (Online Methods and **Supplementary Fig. 2**). This method can easily be extended to assess whether eQTLs identified in two separate tissues represent a functional variant shared by the two tissues (Online Methods). However, the RTC score is not a probability in itself and is affected by the number of variants and the linkage disequilibrium (LD) in a given region. Therefore, we derived a probability from the RTC score by simulating two scenarios for each region: (i) a scenario in which two variants tag different functional effects (H0) and (ii) a scenario in which two variants tag the same functional effect (H1). Subsequently, we generated a distribution centered on the real RTC score for the region and quantified the overlap between this distribution and the distributions of simulated RTC scores under H0 and H1. We then applied Bayes' theorem, in conjunction with the overall tissue-sharing estimates quantified by the π_1 statistic[13], to compute a probability of shared functional effect, which we call *P*(shared), for a given RTC score in a given region (Online Methods and **Supplementary Figs. 3–5**). By converting the RTC score into a probability, we create a metric that accounts for differences in power when calling shared functional effects in different regions and that can be used to discover the tissue specificity of eQTLs.

Being able to calculate the probability of two variants sharing a functional effect allowed us to estimate tissue sharing of eQTLs among the

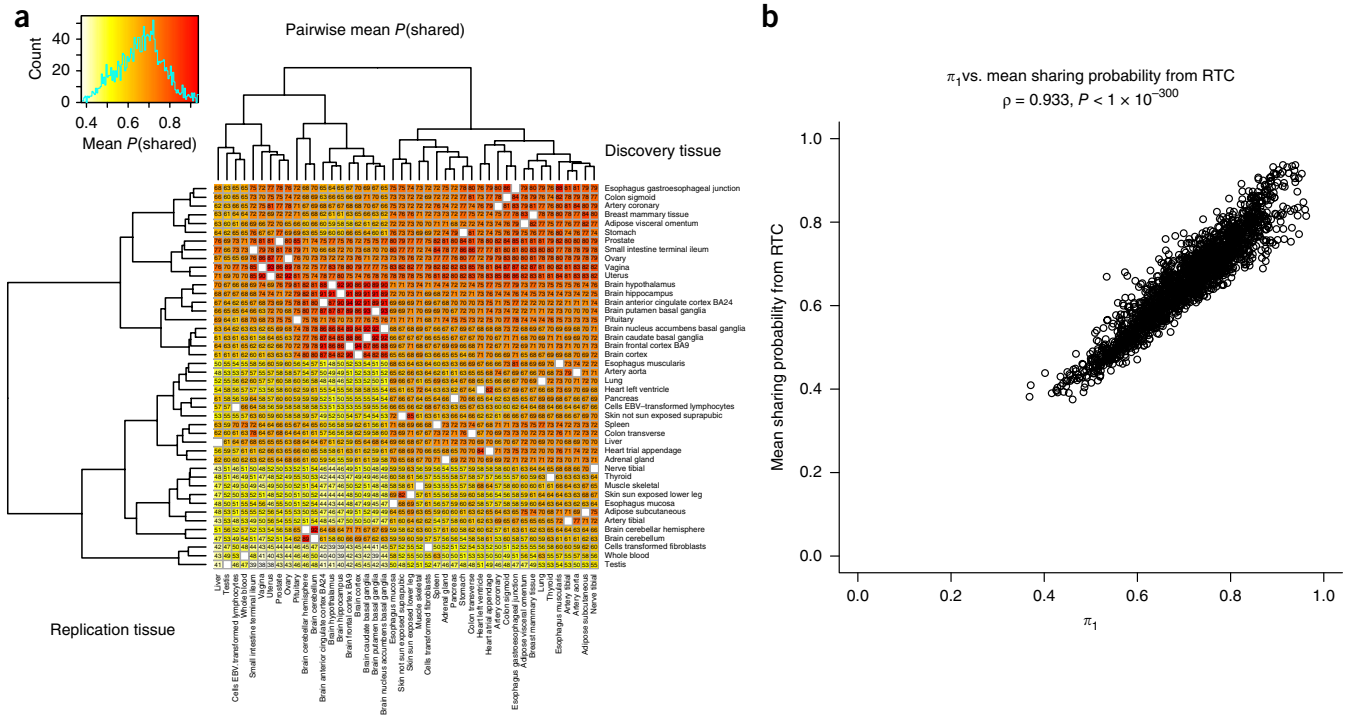

Figure 1 Estimates of tissue sharing for eQTLs among the 44 GTEx tissues. (**a**) Tissue-sharing matrix based on sharing probabilities calculated through RTC. Rows correspond to discovery tissues and columns to replication tissues; hierarchical clustering was performed with the complete linkage method using the Euclidian distances between the mean probabilities of sharing for each tissue pair. Each cell contains the mean probability of sharing multiplied by 100. (**b**) Significant positive correlation between the mean probability of tissue sharing obtained by RTC and the π_1 statistic.

44 GTEx tissues. Gold-standard methods used to quantify tissue sharing of eQTLs, such as the π_1 method, estimate overall sharing between tissues; in contrast, we aimed to estimate the probabilities of each eQTL being shared across tissues using π_1 as the baseline. To ascertain a near-complete list of *cis*-eQTLs, we conducted conditional *cis*-eQTL discovery and identified 858–13,259 independent *cis*-eQTLs at a false discovery rate (FDR) threshold of FDR = 5% (Online Methods and **Supplementary Fig. 6**). Subsequently, we took the union of the eQTLs identified in all of the tissues (Online Methods) and calculated sharing probabilities using the methodology described in the preceding paragraph. We found a high degree of eQTL sharing among biologically related tissues. For example, brain tissues formed a cluster indicating a high level of sharing among these tissues, coronary artery showed the highest degree of sharing with aorta, and uterus and ovary had the most eQTLs in common among all pairs of tissues (**Fig. 1a** and **Supplementary Table 1**). We compared these tissue-sharing estimates to the more commonly used π_1 estimates[13] and found that the two metrics were significantly positively correlated ($r = 0.933$, $P < 1 \times 10^{-300}$; **Fig. 1b**), confirming the validity of our approach. The advantage of RTC over π_1 is that RTC can assess the tissue-sharing probabilities for an individual variant, whereas π_1 estimates the overall sharing and cannot directly make a statement about individual eQTLs.

Unlike the π_1 estimate, our RTC-based probability of eQTL sharing across tissues can be used to find the most likely set of tissues where an eQTL effect is active. We accomplished this by calculating the sharing probabilities for each eQTL in all combinations of the 44 GTEx tissues (Online Methods). Moreover, we recorded the frequency of other tissues identified in the set of most likely tissues for an eQTL. When we considered the distribution of the number of tissues in which each eQTL was likely to be active, the majority of eQTLs (94%) were active in at least one additional tissue, in agreement with previous findings[8,14,15] (**Fig. 2a**). Furthermore, the number of tissues with shared effects decreased sharply as the number of tissues increased, but there was a slight enrichment for eQTLs active across most or all of the 44 tissues (**Fig. 2a**). When we assessed the eQTL sharing estimates among the tissues in which significant eQTLs were found, we discovered that the majority of the tissues exhibited higher degrees of tissue sharing; however, eQTLs in some outlier tissues, like testis and whole blood, showed a higher degree of tissue specificity (**Fig. 2b**,**c**,**e** and **Supplementary Table 2**). As each eQTL identified in a given tissue was predicted to be active in a set of other tissues, we next identified the most frequent other tissues included across all these sets. This was done to measure the global impact of the individual estimates, unlike the tissue-sharing comparison in the previous section where we only quantified the global sharing between tissues. The results indicated that shared eQTL effects were also more frequently observed for tissues with biologically meaningful similarity. For example, brain tissues were most similar to other brain tissues, ovary was most similar to uterus and vagina, and left ventricle in heart was most similar to arterial appendage in heart (**Fig. 2d**,**f**, **Supplementary Fig. 7**, and **Supplementary Table 3**). In summary, our methodology uncovered outlier tissues with eQTLs showing high degrees of tissue specificity and others in which eQTLs showed high levels of sharing among tissues. Tissue-sharing estimates for individual eQTLs identified biologically relevant tissues as shared, indicating that the RTC method is capable of assessing tissue specificity on a variant-by-variant basis.

Colocalization analysis of GWAS variants with GTEx eQTLs

Given that GTEx comprises a wide range of tissues and that our novel methodology can assess tissue sharing for each eQTL variant identified in these tissues, we were in an unprecedented position to infer candidate causal regulatory effects and their target genes that might

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

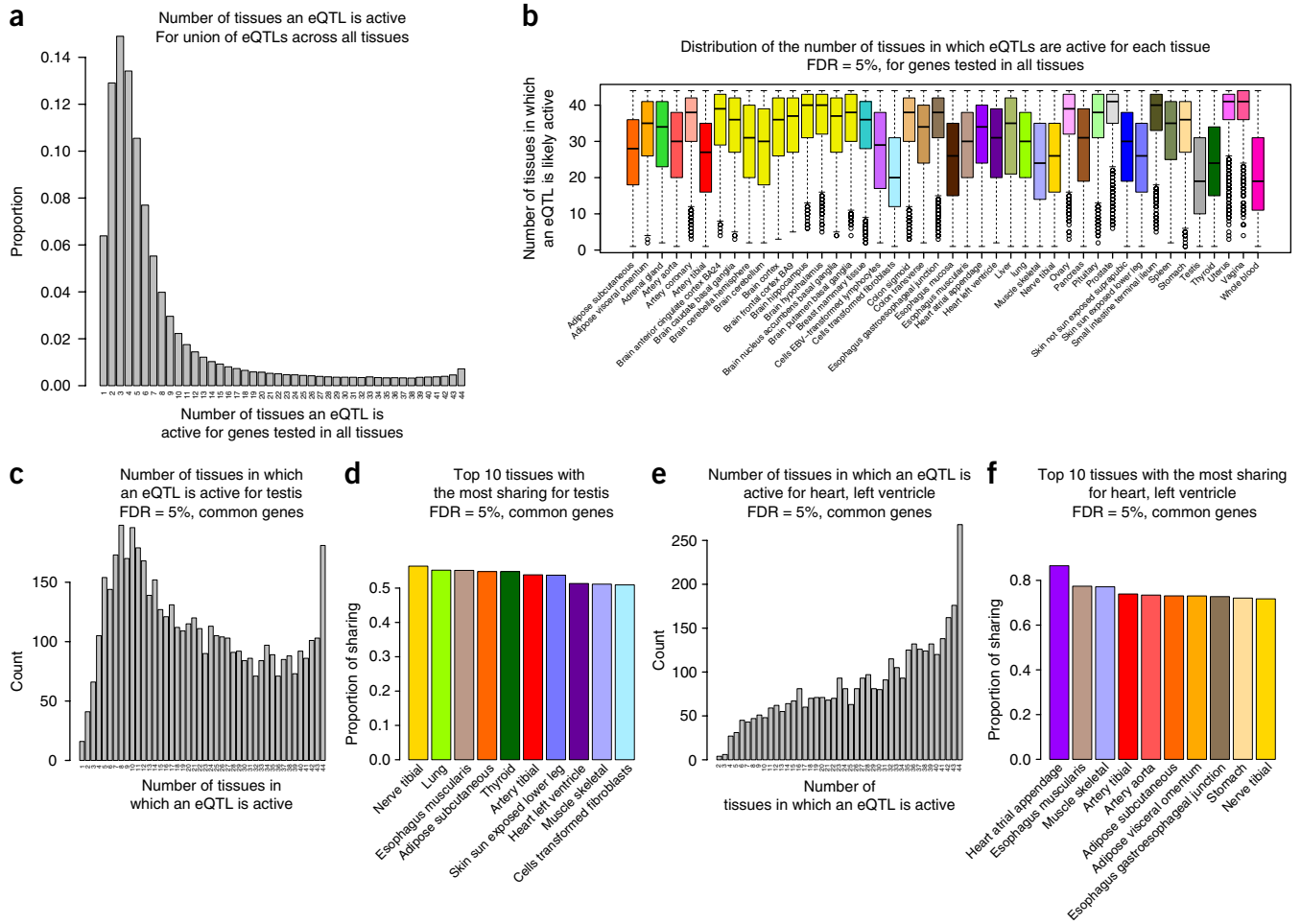

Figure 2 Finding the most likely set of tissues where an eQTL effect is active. (**a**) Frequency distribution of the number of tissues in which an eQTL is active (plotted for the union of eQTLs across all tissues), showing that most eQTLs are shared by at least one or a few other tissues in addition to the original tissue whereas eQTL sharing among high numbers of tissues is rare. (**b**) Distribution of the number of other tissues in which an eQTL is active for the significant (FDR = 5%) eQTLs in each of the tissues. The majority of eQTLs for each tissue are shared with other tissues, like those in brain tissues; however, there are outlier tissues in which eQTLs have higher rates of tissue specificity, like testis and whole blood. In the box plots, the black horizontal line represents the median, the boxes are delimited by the first and third quartiles, and whiskers extend to 1.5 times the box length; outlier data points are represented as circles. (**c**) Testis as an example of a tissue with a higher degree of tissue specificity for eQTLs. (**d**) The top ten tissues sharing eQTLs with testis. (**e**) Heart left ventricle as an example of a tissue sharing most of its eQTLs with other tissues. (**f**) The top ten tissues sharing eQTLs with heart left ventricle.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

mediate the associations of GWAS variants. Because RTC uses only discovered GWAS variants, we were able to test GWAS variant–eQTL overlap for all known GWAS variants and were not limited to GWAS signals with available summary statistics or raw data, which thus far are very sparse. To this end, we downloaded the NHGRI-EBI GWAS catalog[3] and filtered the complete list of 15,929 GWAS variants to include 5,751 variants with genome-wide significant associations ($P < 5 \times 10^{-8}$) that overlapped with GTEx variants. We ran the RTC analysis with the independent significant eQTLs (FDR = 5%) from each of the tissues, which corresponded to 4,664 GWAS variants that colocalized with eQTLs. Next, we created a null set of 5,751 variants that were matched to the list of real GWAS variants on the basis of minor allele frequency (MAF) and distance to the closest transcription start site (TSS) (Online Methods and **Supplementary Fig. 8**). We not only observed a large enrichment of high RTC scores across the GWAS variant–eQTL colocalizations, but also found that significantly fewer null GWAS variants colocalized with eQTLs (3,982 colocalizations; Fisher's exact test, $P = 3 \times 10^{-8}$), confirming, as previously described[5,12,16], that GWAS variants frequently colocalize and likely share functional effects with eQTLs. Thus, at least some of these variants influence traits through regulatory effects (**Fig. 3a**), although colocalization between eQTLs and GWAS variants should not be interpreted as a causal relationship. We also observed a bimodal distribution for probabilities of GWAS and eQTL variants tagging the same functional effect, where the majority of the probabilities were close to 0, but there was also an enrichment for high probabilities (**Fig. 3b**). We have previously shown that RTC score is a better estimate of shared causality for two variants than pairwise LD metrics (r^2 and D')[12]. When we compared the RTC score between two variants to the corresponding r^2 value, we observed that a high r^2 value generally meant a high RTC score; however, many causal links found by RTC may be missed when using r^2 as a metric, extending our previous finding that RTC is preferable to r^2 when predicting causality (**Fig. 3c**, **Supplementary Fig. 9**, and **Supplementary Table 4**). Cases where r^2 was low (<0.1) and RTC was high (>0.9) were due to the level of LD in a given region; more specifically, these regions had significantly lower

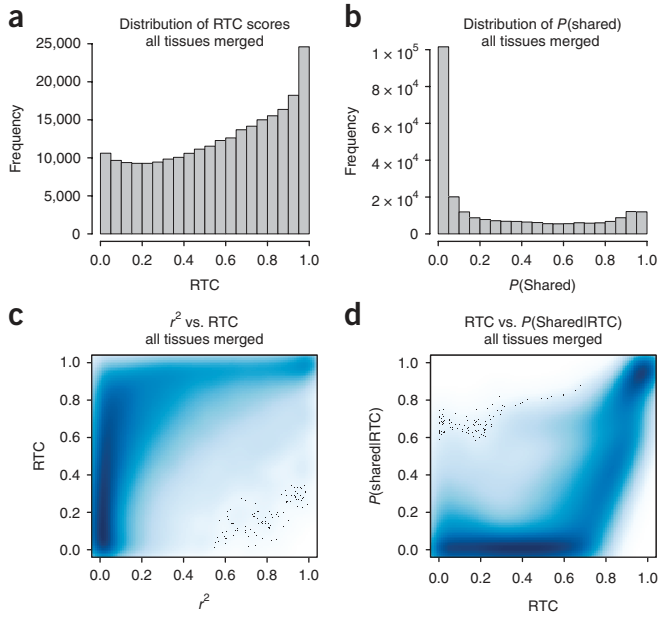

Figure 3 RTC score compared to other pairwise variant metrics. (**a**) Frequency distribution of RTC values for eQTLs from the 44 GTEx tissues showing an enrichment in high RTC scores. (**b**) Distribution of the probabilities of GWAS and eQTL variants tagging the same functional effect. (**c**) RTC score compared to r^2. High RTC score tends to mean high LD between two variants; however, low LD does not necessarily result in a low RTC score, indicating that the RTC score is independent of the LD between two variants. (**d**) Sharing probability calculated from simulations as compared to raw RTC score. Low RTC scores are much less likely to correspond to sharing than high RTC scores; however, there is substantial variation between regions. The density plots in **c** and **d** were generated using the smoothScatter function in R statistical computing software, which produces a smoothed color-density representation of a scatterplot, obtained through a kernel-density estimate. Darker shades of blue indicate a higher density of points in the scatterplot. Black points correspond to individual data points.

LD (Mann–Whitney *U* test, $P < 1 \times 10^{-16}$) when compared to other regions in the genome, causing even weak linkage between the two variants to have a high RTC score (**Supplementary Fig. 10**). We tested how the probability of shared functional effect, as calculated with our new methodology, varied with the raw RTC score and show that this probability behaved as expected, with high RTC scores indicating a high probability of a shared functional effect for the GWAS and eQTL variants. However, the probability was highly variable across regions with the same RTC score, indicating the necessity of calculating this probability on a region-by-region basis (**Fig. 3d**).

Comparison of RTC to another colocalization method, coloc

We compared the probabilities of sharing generated by RTC to the scores obtained with another colocalization method, coloc[17]. As coloc requires summary statistics from GWAS results, we downloaded results from a meta-analysis of total cholesterol levels[18] and calculated the probabilities of shared effect for genome-wide significant GWAS hits and liver eQTLs using both RTC and coloc. We found a strong significant positive correlation ($r = 0.73$; Mann–Whitney *U* test, $P = 3.9 \times 10^{-10}$; Online Methods and **Supplementary Fig. 11**) between the sharing probabilities calculated by the two methods, confirming the validity of our approach. To run coloc, we needed to intersect the lists of variants from the GWAS and eQTL discovery, hence possibly losing some of the most significant variants in a given region, a drawback our methodology does not have. Moreover, RTC assumes that there are two true signals in each region and specifically tests whether they are shared or independent, whereas coloc makes no such assumption. Thus, cases where there was a high RTC sharing probability (≥0.9) and a relatively low coloc sharing probability (≤0.8) are due to coloc attributing a high probability to no GWAS effect in the region, one of the five probabilities that coloc calculates (**Supplementary Fig. 12**). Furthermore, we assessed the performance of each method using a simulation analysis (Online Methods and **Supplementary Fig. 13**). We found that the two methods performed comparably ($\rho = 0.801$, $P = 1.5 \times 10^{-115}$). At a probability threshold of 0.9 in calling shared functional effect, RTC had a sensitivity of 0.91 and a specificity of 0.95, whereas for coloc these measures were 0.66 and 1, respectively (**Supplementary Fig. 14**). On the other hand, if the two methods were matched on sensitivity, coloc had a higher specificity than RTC. As the simulation study was designed such that we calculated GWAS and eQTL *P* values for every variant in the regions, this result indicates that, in scenarios where we know the GWAS and eQTL *P* values for every single variant in a given window, coloc would be the better choice, as it uses all the information in the locus.

Estimating the tissue-causality profiles of GWAS results

Although GWAS provide a list of markers that predispose to a certain disease or trait, they fail to identify the tissues where genetic causality arises. Given that we can test all filtered GWAS signals for eQTL overlap, we can attempt to address this gap in knowledge. However, we are limited to the tissues GTEx has sampled; thus, in some cases, the real causal tissue will be missing. This means that the exact property we are estimating is the relative contributions of the 44 tissues to the genetic causality of a given trait. To do this, we need to know not only whether colocalizing GWAS and eQTL variants are tagging the same functional effect, as inferred by RTC, but also the tissue-wide activity of the eQTL in question. We expected that weighting the probability of GWAS and eQTL variants being due to the same functional effect by the extent of tissue sharing for the eQTL would increase our power in detecting the causal tissue behind the genetic associations of a GWAS trait. To do this, for each eQTL in a given tissue that colocalized with a GWAS variant, we divided the probability of the GWAS variant and eQTL tagging the same functional variant by the sum of the tissue-sharing probabilities for the eQTL in that tissue. This enabled us to weight the GWAS variant–eQTL probabilities such that tissue-specific eQTLs would contribute to a tissue's GWAS enrichment more than eQTLs that were shared with many other tissues. Next, for each disease in each tissue, we divided the sum of the normalized GWAS variant–eQTL probabilities from the previous step by the number of independent eQTLs in the tissue, thereby controlling for different eQTL discovery power across the 44 tissues; we call this our normalized tissue causality score (NTCS). Lastly, we exactly reproduced this analysis with the set of 5,741 null GWAS variants (Online Methods) and compared our NTCS in a tissue for the disease-associated variants to the score we observed under the null for that tissue. The ratio of the real GWAS NTCS to the score under the null was defined as the enrichment metric (**Supplementary Table 5**). Moreover, by comparing the distributions of real NTCSs to the NTCSs under the null, we calculated a *P* value for the observed enrichment (**Supplementary Table 6**). We show that, by using our normalization technique, we could significantly reduce (Mann–Whitney *U* test, $P = 1.9 \times 10^{-18}$; **Supplementary Fig. 15**) the correlation between the number of eQTLs in a tissues and the GWAS enrichment metric, thus allowing us to estimate the relative contribution of tissues to the genetic causality of a trait.

We investigated the overall pattern of tissue causality for GWAS traits and looked at specific examples. For each GWAS trait, we ranked enrichment

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

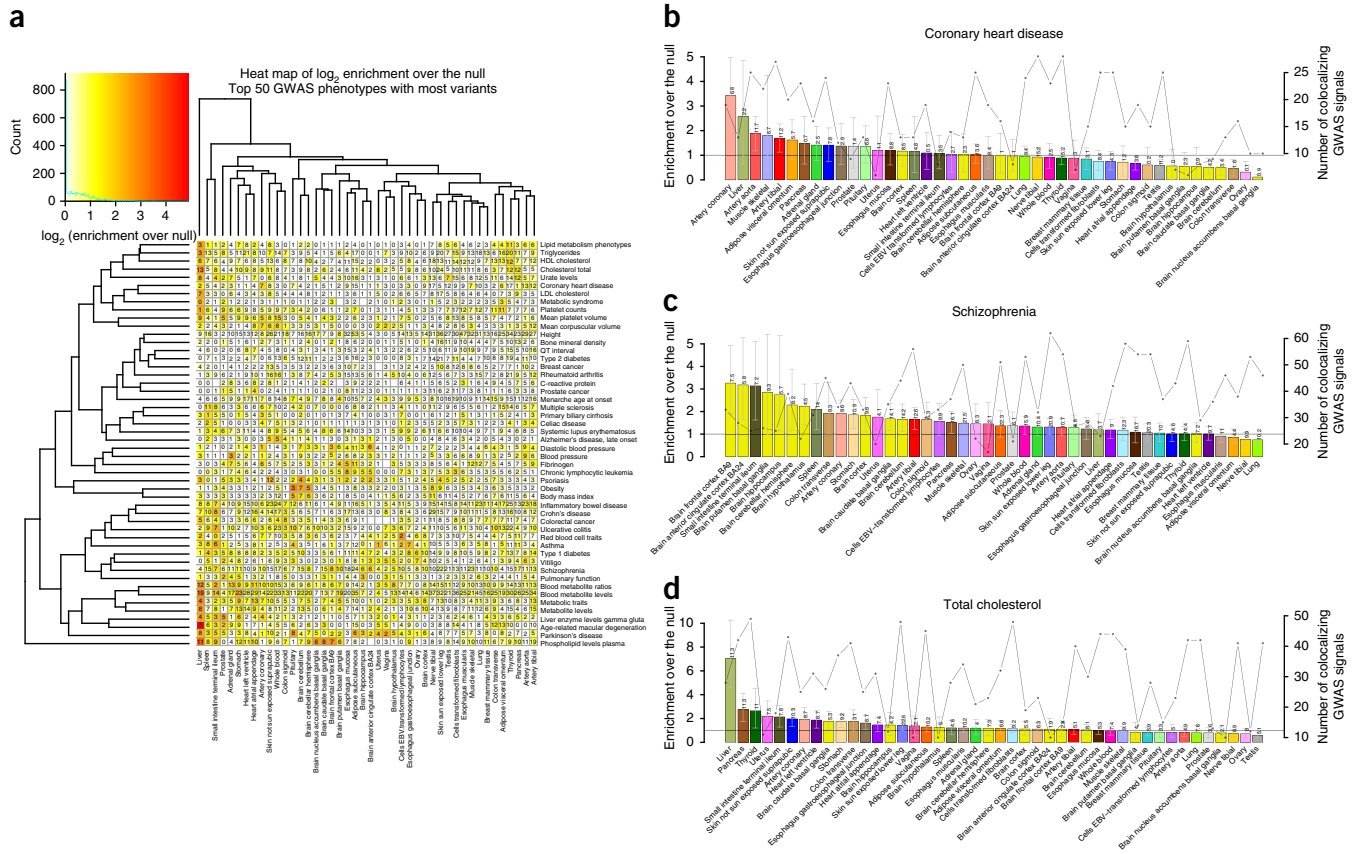

Figure 4 Patterns of tissue causality of GWAS traits. (**a**) Heat map of the tissue-causality profiles for tissues as given by $\log_2$-transformed enrichment over the null to any of the top 50 traits with the highest number of GWAS variants in the NHGRI-EBI GWAS catalog. Tissues that are depleted over the null are presented as zeros in this heat map, as biologically either a tissue is involved in a phenotype, which may be quantified, or it is not involved; thus, there should not be a magnitude of non-involvement. Darker shades of red indicate that higher likelihood of GWAS genetic causality is acting through the corresponding tissue. Rows correspond to the GWAS traits and columns to the tissues, and these are clustered with hierarchical clustering using the complete linkage method on the Euclidian distances calculated from the $\log_2$-transformed enrichment over the null. (**b**–**d**) Examples of traits with a prior on a biologically causal tissue: coronary heart disease (**b**), schizophrenia (**c**), and total cholesterol (**d**). On the primary *y* axis, the enrichments or depletions over the null per tissue are plotted as bars; on the secondary *y* axis, the number of GWAS variants that colocalized with eQTLs per tissue are plotted as a line. The horizontal black line indicates the null. On top of each of the bars is the $-\log_{10}$ Benjamini–Hochberg corrected *P* value for the enrichment or depletion. The 95% confidence interval of the enrichment or depletion as determined by bootstrapping the statistic 1,000 times is shown as a gray line. During each bootstrap iteration, we randomly sample, with replacement, both the observed probabilities of a given disease in a given tissue and the null probabilities in the same tissue, and we recalculate our statistic to assign confidence intervals to the enrichments or depletions.

over the null for each of the tissues. Tissues that were ranked higher were estimated to contribute more to the genetic causality of a GWAS trait. Tissues that showed a depletion when compared to the null were considered to have no enrichment and we ignored the magnitude of the depletion, as biologically we expect a tissue to have a quantitative contribution to the development of a trait; depletion is not meaningful as a quantity. We discovered that liver was the tissue most likely to be causal in most of the GWAS traits (11%), including, as expected, a variety of lipid measurements[19,20] and uric acid levels[21] (**Fig. 4a**, **Supplementary Fig. 16**, and **Supplementary Table 7**). Brain tissues were the top tissues relating to traits like height[22], schizophrenia[23,24], and age of onset of puberty[25]. Furthermore, for traits where we had a biological prior of a causal tissue and the tissue was assayed in GTEx, this tissue tended to be the tissue identified as most likely to be causal by our methodology. For example, the top causal tissue for coronary heart disease was coronary artery followed by liver; for schizophrenia, the top tissues were brain tissues; and for lipid metabolism traits, like total cholesterol levels, the top tissue tended to be liver (**Fig. 4b–d**). In the case of coronary heart disease, coronary artery is usually thought of as a 'passenger' tissue, where the effects of the disease are manifested rather than the tissue contributing to pathophysiology; however, our analysis identifies it as a likely causal tissue, indicating that there are potentially novel risk factors to be discovered. We also observed that there was overlap between the confidence intervals of tissues. While in some cases, like total cholesterol measurements, we had the statistical power to dissociate the top tissue from others, this was not the case in all diseases, indicating that we are still underpowered. However, larger sample sizes will likely make the tissues statically distinguishable, without affecting the ranking of the top tissues. Thus, we show that, by having access to eQTLs from multiple tissues and controlling for the tissue specificity of eQTLs using our new methodology, we can estimate the ranking of relevant tissues from which the genetic causality of GWAS traits arise.

Causal tissues correctly identify the causal gene for a GWAS result

As we estimated the tissue causality profiles for GWAS traits, we can compare the causal genes for the GWAS associations between tissues

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

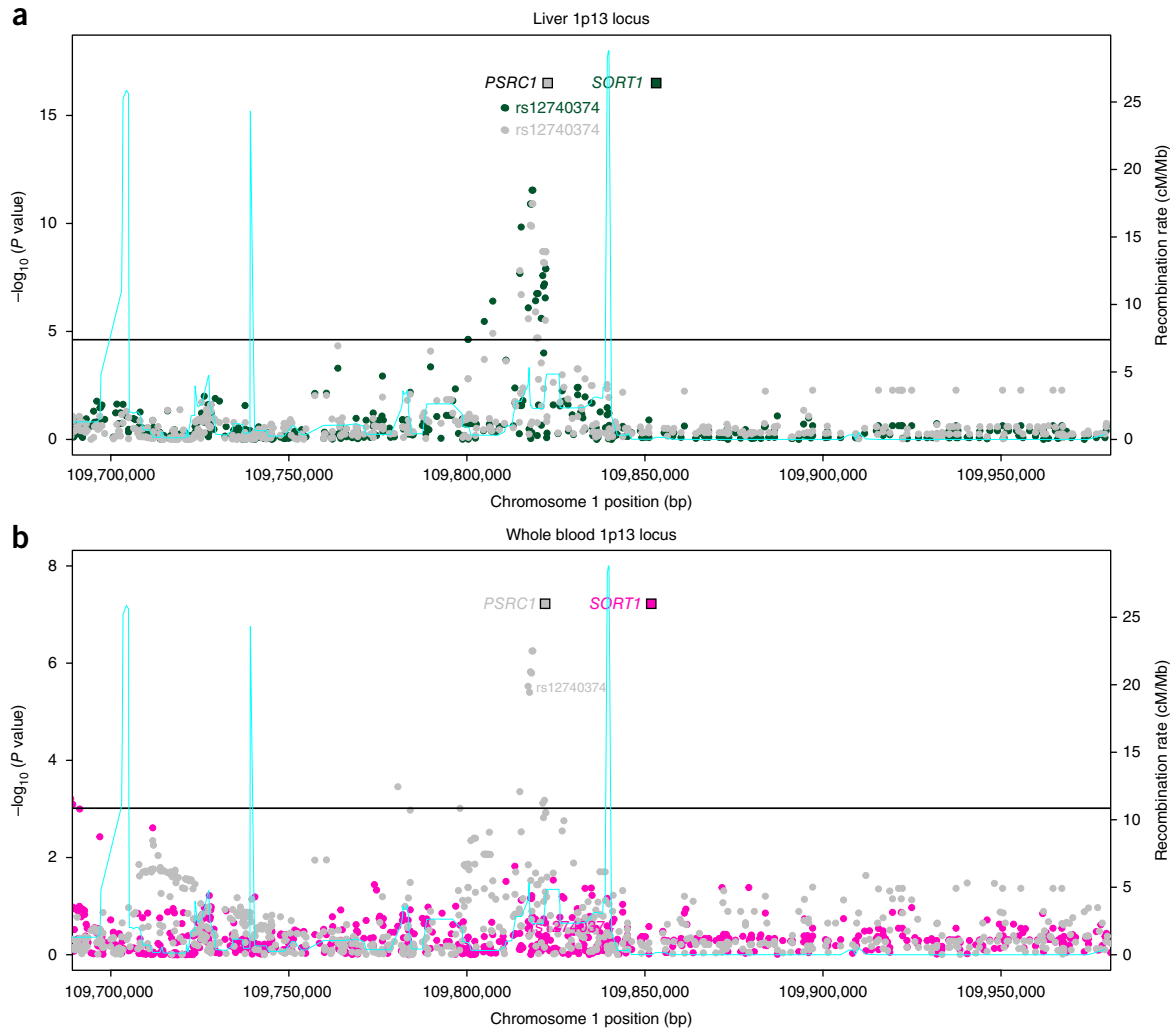

Figure 5 eQTL effects at the coronary artery disease (CAD)- and lipid levels–associated 1p13 locus. (**a**) Liver. (**b**) Whole blood. Points are the $-\log_{10}$ (*P* value). eQTL associations for *SORT1* are shown in green in liver and magenta in whole blood, and *PSRC1* is shown in gray. The cyan line is the recombination rate, given in the secondary *y* axis, and the boxes highlight the positions of the two genes. The genome-wide significance threshold for eQTL associations is represented as the horizontal black line. In both tissues, the best eQTL association is genome-wide significant (FDR = 5%); however, the eQTL gene, for which the eQTL and the causal rs12740374 variant are tagging the same functional effect as identified by our method, is different. Liver, which we estimate to have a key role in both the development of CAD and the regulation of lipid levels, correctly identifies *SORT1* as the causal gene for this GWAS association, as *SORT1* was the strongest eQTL effect of rs12740374 in liver for all genes tested in *cis*, and the eQTL effect of this variant on *SORT1* is ~2× more tissue specific than its eQTL effect on *PSRC1*. However, the more easily collectable whole blood, which is estimated not to contribute to these traits, fails to do so. If we had just whole-blood eQTLs and did not know the tissue-causality profile for these traits, we would have incorrectly identified *PSRC1* as a putative causal gene.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

likely contributing to the genetic causality of GWAS traits and those that are not. We examined the rs12740374 variant in the 1p13 locus, which is not only associated with coronary artery disease[26,27] and lipid measurements[28], but is also one of the few GWAS noncoding loci where the mechanistic causes are well established[29]. Liver is a key tissue in both heart disease and lipid levels (**Fig. 4b**,**d**), and in liver the causal gene for the rs12740374 association is correctly identified as *SORT1* (*P*(Shared) = 1)[29]. In tissues that do not contribute highly to the genetic causality of these traits, like testis and whole blood, we incorrectly identified another nearby gene, *PSRC1*, as the putative causal gene (*P*(Shared) = 0.96 and 0.97, respectively; **Fig. 5** and **Supplementary Table 8**). Notably, the tissues where *SORT1* was correctly identified contribute significantly (Mann–Whitney $P = 0.0004$) more to the genetic causality of heart disease and lipid levels than tissues where the causal gene was different (**Supplementary Fig. 17**). Finally, in liver, the eQTL effect of rs12740374 on *SORT1* was ~2× more tissue specific when compared to its effect on *PSRC1* (likely active in 12 versus 21 tissues, respectively), which downgrades the colocalization with *PSRC1* in our analysis. This result shows the importance of identifying the causal tissues for GWAS traits before stating which genes may be responsible for these associations.

Clustering of diseases with common pathophysiology based on tissue-causality profiles

Finally, we asked how different diseases with shared pathophysiology differ with respect to which tissues contribute to their genetic causality. To this end, we investigated autoimmune and cardiometabolic diseases and used hierarchical clustering to group the individual diseases per their relative tissue-causality profiles. Among the autoimmune diseases, we found that Crohn's disease and ulcerative colitis formed a cluster, whereas celiac disease had a different tissue-causality profile

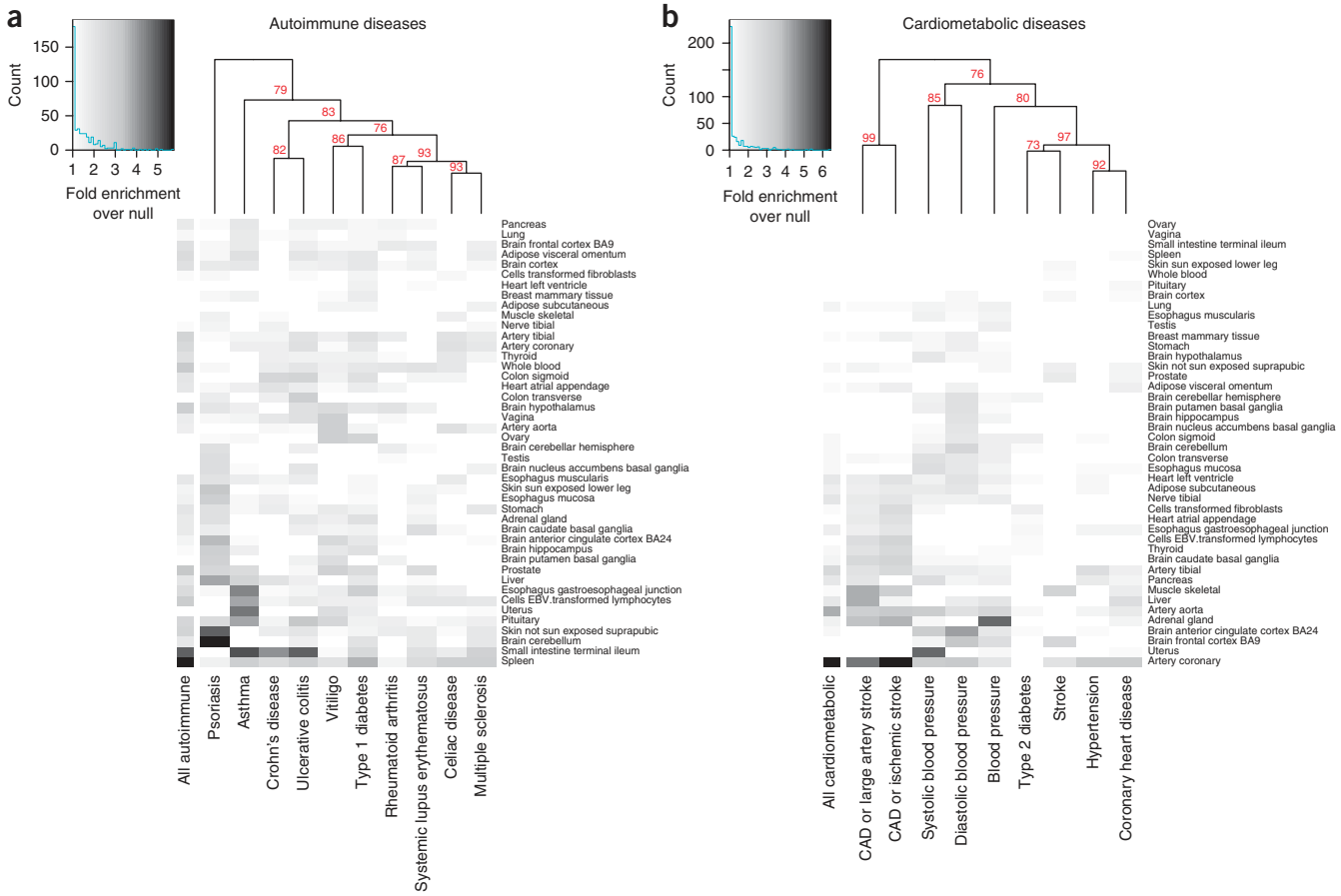

Figure 6 Enrichment over the null of tissues signifying their contribution to the genetic causality of complex diseases. (**a**) Autoimmune diseases. (**b**) Cardiometabolic disorders. Rows list tissues, and columns list diseases. Darker shades correspond to higher contribution per tissue. The leftmost column shows the relative tissue contributions across all diseases combined. The hierarchical clustering of the diseases is shown as a dendrogram. Clustering was conducted with hierarchical clustering using the complete linkage method on the Euclidian distances calculated from enrichment over the null. The red number on each node of the dendrogram is the approximately unbiased bootstrap probability for each node as calculated by the pvclust[30] R package using 1,000 bootstrap iterations.

and was most similar to multiple sclerosis. Type 1 diabetes and vitiligo seemed most similar to each other, and rheumatoid arthritis and lupus were clustered together. Asthma and psoriasis appeared markedly different when compared to other autoimmune disorders (**Fig. 6a**). For cardiometabolic diseases, blood pressure–related traits, coronary heart disease and hypertension, CAD or stroke phenotypes, and type 2 diabetes and stroke clustered together (**Fig. 6b**). We demonstrate that, by comparing the tissue-causality profiles of GWAS diseases, we can begin to disentangle the common as well as diverging biology underlying their development.

DISCUSSION

Here we describe a new approach that is designed to estimate the likely causal tissues underlying the genetic causality of GWAS traits. In this study, we use the eQTLs identified by the GTEx Consortium to find the relative contribution of the 44 tissues to the genetic causality of a given GWAS trait. The 44 tissues assayed in this study do not constitute a complete representation of all human tissues and thus will not be applicable to all GWAS and may fail to identify the real causal tissue as a result of it not being sampled; however, GTEx represents the most comprehensive eQTL data set of human tissues. Furthermore, in some cases, the tissues are not statistically distinguishable from each other, which may be owing to lack of power, lack of tissue specificity of GTEx *cis*-eQTLs, or the fact that GWAS traits truly operate through many diverse tissues. Given the tissue and sample size limitations, there is room for improvement in determining true tissue causality profiles for GWAS traits. However, our analysis represents an unbiased and fairly complete profiling of the relative tissue contributions to GWAS genetic causality at an unprecedented scale. As the sample sizes and the number of tissues assessed for eQTLs and our resolution of the genetic etiology of complex disorders increase, we expect our methodology to yield even more powerful conclusions. We believe that this type of approach will be paramount in the interpretation of new GWAS results using a publicly available data set, like GTEx, and will aid in the design of downstream functional experiments to identify the mechanistic causes of complex disorders and traits, as well as new avenues of treatment and prevention.

URLs. GTEx Portal, http://gtexportal.org/home/; QTLtools, https://qtltools.github.io/qtltools/; Vital-IT, http://www.vital-it.ch/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

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

ACKNOWLEDGMENTS

This research was supported by grants from the US National Institutes of Health (NIH-R01MH101814), European Commission Framework Programme 7 (UE7-SYSCOL-258236), the European Research Council (UE7-POPRNASEQ-260927), the Swiss National Science Foundation (31003A-149984 and 31003A-170096), and the Louis Jeantet Foundation. Computations were performed at the Vital-IT Centre for High-Performance Computing of the SIB Swiss Institute of Bioinformatics.

AUTHOR CONTRIBUTIONS

H.O. and E.T.D. designed the study. H.O., A.A.B., and O.D. conducted the analysis and developed software. A.C.N. designed the original RTC method. N.I.P. tested the software. H.O. wrote and E.T.D. edited the manuscript. The GTEx Consortium generated the data.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Bush, W.S. & Moore, J.H. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).
2. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
3. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
4. Dermitzakis, E.T. From gene expression to disease risk. *Nat. Genet.* **40**, 492–493 (2008).
5. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
6. Nica, A.C. & Dermitzakis, E.T. Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.* **17** R2, R129–R134 (2008).
7. Montgomery, S.B. & Dermitzakis, E.T. From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet.* **12**, 277–282 (2011).
8. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
9. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486 (2013).
10. Brown, C.D., Mangravite, L.M. & Engelhardt, B.E. Integrative modeling of eQTLs and *cis*-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLoS Genet.* **9**, e1003649 (2013).
11. Hore, V. *et al.* Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48**, 1094–1100 (2016).
12. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
13. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
14. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**, 1246–1250 (2009).
15. Gutierrez-Arcelus, M. *et al.* Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet.* **11**, e1004958 (2015).
16. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
17. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
18. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
19. Nguyen, P. *et al.* Liver lipid metabolism. *J. Anim. Physiol. Anim. Nutr. (Berl.)* **92**, 272–283 (2008).
20. Saltiel, A.R. & Kahn, C.R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001).
21. Maiuolo, J., Oppedisano, F., Gratteri, S., Muscoli, C. & Mollace, V. Regulation of uric acid metabolism and excretion. *Int. J. Cardiol.* **213**, 8–14 (2016).
22. Taki, Y. *et al.* Correlation among body height, intelligence, and brain gray matter volume in healthy children. *Neuroimage* **59**, 1023–1027 (2012).
23. Buckner, R.L., Andrews-Hanna, J.R. & Schacter, D.L. The brain's default network: anatomy, function, and relevance to disease. *Ann. NY Acad. Sci.* **1124**, 1–38 (2008).
24. Harrison, P.J. The neuropathology of schizophrenia. A critical review of the data and their interpretation. *Brain* **122**, 593–624 (1999).
25. Han, S.K. *et al.* Activation of gonadotropin-releasing hormone neurons by kisspeptin as a neuroendocrine switch for the onset of puberty. *J. Neurosci.* **25**, 11349–11356 (2005).
26. Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* **41**, 334–341 (2009).
27. Samani, N.J. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443–453 (2007).
28. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
29. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
30. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

ONLINE METHODS

GTEx project. GTEx comprises 44 tissues, each having more than 60 samples, collected from 7,051 post-mortem biopsies from 449 individuals, where each tissue has a different number of samples (**Supplementary Fig. 1**). For details of data production and quality control, see ref. 31.

Conditional eQTL discovery. Multiple independent signals for a given expression phenotype were mapped using a greedy forward stepwise regression algorithm followed by a backward selection step. First, the set of GTEx eGenes for a given tissue is taken and the maximum beta-adjusted *P* value (correcting for multiple testing across the SNPs) over these genes is taken as the gene-level threshold. Then, for each gene, FastQTL[32] is run iteratively. At each iteration, it performs a *cis* scan of the window, correcting for all previously discovered SNPs and all standard GTEx covariates. If the beta-adjusted *P* value for the peak SNP is not significant at the gene-level threshold, the forward stage is complete and the procedure moves on to the backward step. If the *P* value is significant, the peak SNP is added to the list of discovered eQTLs as an independent signal and the forward step moves on to the next iteration.

Once the forward stage is complete for a given gene, we have a list of associated SNPs; we refer to these as forward signals. The backward stage consists of testing each forward signal separately, controlling for all other discovered signals. To do this, for each forward signal, we run a *cis* scan over all variants in the window in FastQTL[32] using all standard covariates and all other discovered signals as covariates. If no SNP is significant at the gene-level threshold, the signal being tested is dropped; otherwise, the peak SNP from the scan is chosen as the variant that represents the signal best in the full model (**Supplementary Fig. 6**).

Regulatory trait concordance score. The methodology described in this section is implemented in QTLtools[33]. When assessing tissue specificity of eQTLs, we use the same method; however, in that case, the GWAS variant becomes an eQTL in a different tissue.

RTC method. The RTC algorithm assesses the likelihood of a shared functional effect between a GWAS SNP and an eQTL by quantifying the change in the statistical significance of the eQTL after correcting the eQTL phenotype (here, gene expression) for the genetic effect of the GWAS SNP and comparing its correction impact to that of all other SNPs in the interval[12]. We mapped all common autosomal variants in each of the tissues to the recombination hotspot intervals as defined by McVean *et al.*[34]. These coordinates were transformed into GRCh37 coordinates using the liftOver tool[35]. The RTC method is as follows: for a GWAS variant falling into the same region flanked by recombination hotspots (a coldspot) as an eQTL, with *N* variants in a given coldspot:

1. Correct the phenotype for each of the variants in the region separately by linear regression, yielding *N* pseudo-phenotypes (residuals);
2. Redo the eQTL variant association with all of these pseudo-phenotypes;
3. Sort (in decreasing order) the resulting *P* values and find the rank of the eQTL to GWAS SNP–pseudo-phenotype among all eQTL to pseudo-phenotype associations;
4. RTC = $(N - \text{rank}_{\text{GWAS SNP}})/N$.

This results in an RTC score that ranges from 0 to 1, where higher values indicate a more likely shared functional effect for the GWAS and eQTL variants (**Supplementary Fig. 2**). If there are multiple independent eQTLs for a given phenotype, the RTC for each independent eQTL is assessed after correcting the phenotype for all the other eQTL variants for that phenotype. This correction is done using linear regression and taking the residuals after regressing the phenotype with the other eQTLs. For example, for $\text{eQTL}_{\text{independent 1}}$, the phenotype tested is equal to the residuals resulting from the following linear model: phenotype ~$\text{eQTL}_{\text{independent 2}}$ + … + $\text{eQTL}_{\text{independent } N}$.

Simulating RTC scores under the null hypothesis (H0). The H0 scenario arises when two variants in a coldspot tag different functional effects. To simulate this, we do the following:

1. For a coldspot that harbors colocalized GWAS and eQTL variants ($\text{eQTL}_{\text{real}}$), we randomly pick two hidden causal variants ($\text{GWAS}_{\text{causal}}$ and $\text{eQTL}_{\text{causal}}$);
2. We find two variants (GWAS and eQTL) that are linked ($r^2 \geq 0.5$) to the hidden causal variants ($\text{GWAS}_{\text{causal}}$ and $\text{eQTL}_{\text{causal}}$, respectively);
3. We generate a pseudo-phenotype for $\text{eQTL}_{\text{causal}}$ based on the β and intercept of $\text{eQTL}_{\text{real}}$ and randomly distributed residuals of $\text{eQTL}_{\text{real}}$;
4. We rerun the RTC analysis with this new pseudo-phenotype and using the GWAS and eQTL variants.

We repeat these steps 200 times for each coldspot with an eQTL and a GWAS variant, in all of the 44 tissues separately, and record the H0 RTC distributions for each region (**Supplementary Figs. 3** and **18**). Multiple independent eQTLs are handled as described in the previous section.

Simulating RTC scores under the alternate hypothesis (H1). The H1 scenario arises when two variants tag the same functional variant. The scheme here is exactly the same as for H0, except that there is only one hidden causal variant and both the GWAS and eQTL variants are randomly selected from variants that are linked to the same hidden causal variant. This is implemented as follows:

1. For a coldspot that harbors colocalized GWAS and eQTL variants ($\text{eQTL}_{\text{real}}$), we randomly pick one hidden causal variant ($\text{eQTL}_{\text{causal}}$);
2. We find two variants (GWAS and eQTL) that are linked ($r^2 \geq 0.5$) to the hidden causal variant;
3. We generate a pseudo-phenotype for $\text{eQTL}_{\text{causal}}$ based on the β and intercept of $\text{eQTL}_{\text{real}}$ and randomly distributed residuals of $\text{eQTL}_{\text{real}}$;
4. We rerun the RTC analysis with this new pseudo-phenotype and using the GWAS and eQTL variants.

We repeat these steps 200 times for each coldspot with an eQTL and a GWAS variant, in all 44 tissues separately, and record the H1 RTC distributions for each region (**Supplementary Figs. 4** and **18**).

Conversion of RTC score into a probability of sharing. Each region is characterized by an RTC score for a GWAS–eQTL localization and a distribution of RTC scores under the null and alternate hypotheses. We can use these and Bayes' theorem to estimate the probability of the two variants having the same functional effect for a given RTC score, expressed as $P(\text{shared}|\text{RTC} = rtc)$. To estimate the probability of overall sharing by GWAS and eQTL variants in a given tissue, we first calculate the eQTL *P* values for the GWAS variants from which we calculate the π_1 statistic, which estimates the proportion of true positives[13], and this becomes our overall probability of sharing, $P(\text{shared})$, and by extension the overall probability of not sharing, $P(\text{not shared})$, is defined as $1 - \pi_1(\pi_0)$. To estimate $P(\text{RTC} = rtc|\text{shared})$ and $P(\text{RTC} = rtc|\text{not shared})$, we do the following (**Supplementary Fig. 5**). First, we merge and sort the RTC values ascertained from simulations under the null and alternative hypotheses. We then take 10% of the values flanking our real RTC score to produce a range from which we can estimate a point probability. For example, 200 simulations under H0 and H1 would result in 400 values, which we sort and use to find the position of the real RTC value in this distribution; if this happens to be the 100th value, we take the 60th and 140th simulated RTC values to define our range. Subsequently, we calculate the proportion of values within the range identified in the previous step under H0, which equates to $P(\text{RTC} = rtc|\text{not shared})$, and the proportion of overlap with this range under H1 becomes $P(\text{RTC} = rtc|\text{shared})$. Finally, we apply Bayes' theorem to estimate $P(\text{shared}|\text{RTC} = rtc)$:

$$P(\text{shared}|\text{RTC}=rtc)=\frac{p(\text{RTC}=rtc\,|\,\text{shared})*\pi_1}{p(\text{RTC}=rtc\,|\,\text{not shared})*\pi_0+p(\text{RTC}=rtc\,|\,\text{shared})*\pi_1}$$

This was done for each region and each tissue separately.

Tissue sharing of eQTLs using the RTC score. The methodology described above can also be used to assess tissue specificity of eQTLs; when doing so, the GWAS variant in the previous section becomes an eQTL in a different tissue.

Taking the union of eQTLs across the 44 tissues. To quantify tissue sharing in every region that harbors an eQTL in any of the 44 tissues, we took the union of significant eQTLs. First, we mapped all the significant eQTLs in all the tissues to recombination coldspots. Subsequently, if certain tissues did not

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

doi:10.1038/ng.3981

have a significant eQTL in a given coldspot for a given gene, we took the most significant variant associated with the expression of that gene in that coldspot for all the missing tissues.

Tissue-sharing calculations. Tissue-sharing calculations were conducted for pairs of tissues in both directions: that is, we tested the union of eQTLs found in the previous section for tissue A (discovery tissue) in tissue B (replication tissue) and also, reciprocally, for tissue B in tissue A. This resulted in 1,892 separate runs for the 44 tissues. The RTC score calculations and simulations were conducted as described above. The sharing probabilities are calculated using the same method as above, but with the following exception: we use different *P*(shared) estimates for eQTLs that are significant in either of the two tissues and eQTLs that are not significant in either tissue. The method of estimating *P*(shared) is the same; that is, we calculate the *P* values of eQTLs for tissue A in tissue B and determine the π_1 statistic (**Supplementary Fig. 19**).

Most likely set of tissues in which an eQTL is active. As this method yields a probability of sharing for each eQTL variant in tissue A with all the other 43 tissues, we can calculate the set of tissues in which each eQTL is most likely to be active. This is accomplished in the following manner:

1. Because RTC is a metric that is designed to assess a whole region, we condense the values for each separate coldspot. To estimate probabilities of sharing by tissue A and tissue B, we take the mean of the pairwise estimates in both directions, that is, the mean of tissue A in tissue B and the mean of tissue B in tissue A. If there are multiple independent eQTLs in a coldspot, then we take the combination of values with the highest sharing probabilities. This results in a vector of 44 probabilities for each coldspot;
2. We iterate over the number of tissues, that is, 1 to 44, and call this *N*;
3. At each iteration, we identify the most likely set of *N* tissues in which an eQTL is active. This is done by sorting (in decreasing order) the 44 sharing probabilities and multiplying the product for the top *N* sharing probabilities by the product of 1 minus the sharing probabilities for the remaining 44 – *N* tissues. This yields the maximum probability of an eQTL being active in only *N* tissues;
4. Once we have all the probabilities for 1 to 44 tissues, we take the maximum of these, which corresponds to the most likely number of tissues, called *n*. The set of tissues in which an eQTL is most likely to be active is defined as the top *n* tissues in the sharing probability vector sorted in decreasing order.

Validation of tissue-sharing estimates from RTC with π_1. Considering all significant independent eQTLs in tissue A, we took the mean of the probabilities of their being shared with tissue B as the replication probability of eQTLs from tissue A in tissue B. We then compared this mean probability to the π_1 statistic for replication of eQTLs from tissue A in tissue B. Briefly, the π_1 statistic, which is obtained by evaluating a *P*-value distribution, assesses the proportion of the *P* values that do not originate from the null distribution of *P* values, thus quantifying the proportion of true signal in the data.

Response operator curve for using r^2 to call shared effects. We took RTC scores ≥0.9 to indicate a real shared functional effect for GWAS and eQTL variants[12]. Subsequently, using different r^2 thresholds (from 0 to 1, separated by steps of 0.01), we asked what percentage of the shared signals based on r^2 had RTC scores ≥0.9 (true-positive rate of r^2) and what percentage of the shared signals based on r^2 had RTC scores <0.9 (false-positive rate of r^2) if we called the effects for variant pairs with r^2 values equal to or greater than the given threshold as shared and those for variants with r^2 values below this threshold as not shared.

Tissue enrichments for GWAS traits. *Generation of a matched set of null variants for real GWAS variants.* For each real GWAS variant, we calculated the MAF and the distance to the closest TSS. We then selected one matched null variant for each real GWAS variant such that the null variant's MAF was ±2.5% with respect to the real GWAS variant's MAF and the relative distance to the closest TSS was ±5 kb with respect to the corresponding distance for the real GWAS variant. This resulted in 5,741 matched null GWAS variants.

Calculation of tissue enrichments over the null. We applied our methodology as described above to variants from the NHGRI-EBI GWAS catalog[3] downloaded on 15 June 2015 and the significant independent eQTLs identified in all 44 tissues. We filtered the catalog for GWAS variants that had reported $P < 5 \times 10^{-8}$. This yielded in 5,751 unique GWAS variants from 742 diseases or traits that overlapped with GTEx variants. To normalize the GWAS variant–eQTL probabilities of tissue specificity for the eQTL in a given GWAS variant–eQTL colocalization for a given disease, we divided the GWAS variant–eQTL sharing probability in a given tissue by the sum of the tissue-sharing probabilities of that eQTL in that tissue. This enabled us to increase the impact of tissue-specific eQTLs on disease tissue enrichment as compared to tissue-shared eQTLs. Subsequently, we took the sum of all normalized GWAS variant–eQTL probabilities for a disease in each tissue and divided by the number of independent eQTLs in each of the tissues, thereby accounting for the different power of discovery among the 44 tissues, and this became our NTCS. We then redid the analysis exactly as described here with the matched null GWAS variants and recalculated an NTCS for all of the 5,741 null variants in each of the tissues. Subsequently, we compared the distribution of real GWAS NTCS values for a given disease in a tissue to the null distribution in the same tissue. When making such a comparison for a particular disease, we scaled the NTCS for the null distribution such that the null NTCS was of the same order as the real NTCS, as the denominator in this calculation is a sum and the numbers of variants for a disease and null variants are different. This was accomplished by multiplying the null NTCS value by the ratio of the number of real GWAS variants assessed to the total number of null GWAS variants. Finally, to calculate a *P* value for enrichment, we compared the distribution of real NTCS values to that for the null variants using the Mann–Whitney test. We tested how our enrichment metric correlated with the number of independent eQTLs for each disease before and after our normalization scheme. Before normalization, a clear majority of the diseases exhibited a significant correlation between the number of eQTLs and GWAS tissue enrichment values, whereas after normalization only 11 diseases were still significantly associated (**Supplementary Fig. 15**). In the following equations, calculation of NTCS is described, where *T* is a given tissue; *G* is a GWAS trait; G_n is the total number of GWAS-associated variants for the trait; GWAS *P*(shared) G_i eQTL_i is the probability that a GWAS variant G_i and colocalizing eQTL_i tag the same functional variant; *T P*(shared) eQTL_i is the tissue-sharing probability of eQTL_i in tissue *T* with other tissues; T_n is the total number of eQTLs in tissue *T*; N_n is the total number of null variants; and N_i is a null variant.

$$\text{NTCS} = \frac{1}{T_n} \times \sum_{i=1}^{G_n} \frac{\text{GWAS}\,P(\text{shared})G_i\text{eQTL}_i}{T\,P(\text{shared})\text{eQTL}_i}$$

$$\text{Null NTCS} = \frac{G_n}{N_n T_n} \times \sum_{i=1}^{N_n} \frac{\text{GWAS}\,P(\text{shared})N_i\text{eQTL}_i}{T\,P(\text{shared})\text{eQTL}_i}$$

The enrichment score for GWAS trait *G* in tissue *T* is defined as the NTCS over the null NTCS, and the *P* value is calculated using a Mann–Whitney test comparing the distributions containing each of the *i*th elements in the formulas for the real GWAS and under the null.

Comparison of RTC P(shared) with coloc. Because the coloc method requires summary statistics from GWAS, we cannot directly compare the methodologies for all GWAS variants assessed with RTC in this manuscript. However, we downloaded summary statistics from a GWAS meta-analysis of total cholesterol levels originating from 188,577 individuals[18]. We then mapped all variants with GWAS $P < 5 \times 10^{-8}$ to the same recombination regions used in the RTC analysis, keeping only the most significant GWAS variant in a given region. Subsequently, we ran the RTC analysis with this set of GWAS variants associated with total cholesterol in the same way as described previously. We then calculated the eQTL *P* values for all the variants in a given region for the genes that colocalized with the total cholesterol GWAS list and also their MAFs. These data were then merged with the overlapping total cholesterol GWAS *P* values for each gene and region separately. These values were inputted into the coloc R package using the coloc.abf() function, to calculate probabilities of the two variants tagging the same functional effect. Finally, we compared the *P*(shared) value obtained from RTC for each gene and region with the corresponding coloc H4 probability (where the traits are associated and share a single causal variant; the same probability assessed with RTC).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Simulations to compare RTC and coloc. We randomly chose 256 regions delimited by recombination hotspots across the genome and subsequently extracted the genotypes in these regions from the 1000 Genomes Project Phase 3 release only for European samples[36]. In each region, we randomly selected a causal variant using HAPGEN2 software[37], and we simulated a GWAS with 10,000 cases and controls where each alternate allele conferred risk of 1.1 to the disease phenotype. We then randomly selected 500 controls from genotypes generated by HAPGEN2, which made up our eQTL cohort. In each region, we simulated two scenarios: the null where the eQTL and GWAS tagged independent causal effects and the alternate where the variants are due to the same underlying effect. Thus, we randomly selected one eQTL variant that had $r^2 < 0.2$ with the causal GWAS variant for the null and another eQTL variant with $r^2 > 0.8$ with the causal GWAS as the alternate case. Then, we created a phenotype for each of the two eQTL variants chosen on the basis of randomly selected β values from a distribution of β values for real eQTLs and random error (normally distributed) for each genotype. The GWAS *P* values necessary for coloc were generated using logistic regression, and eQTL *P* values were generated using linear regression. Finally, we ran both coloc and RTC for each of the 256 regions for both the null and alternate hypotheses.

Accession codes. The GTEx data used in this paper are available through controlled access at the database of Genotypes and Phenotypes (dbGaP) under accession phs000424.v6.p1.

Data availability. Data from the study are available from the authors upon reasonable request. A **Life Sciences Reporting Summary** is available.

31. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* http://dx.doi.org/10.1038/nature24277 (2017).
32. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
33. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
34. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
35. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
36. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

nature research

Corresponding Author: Emmanouil T. Dermitzakis

Date: Sep 27, 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined. N/A

2. Data exclusions

   Describe any data exclusions. N/A

3. Replication

   Describe whether the experimental findings were reliably reproduced. N/A

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups. N/A

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. N/A

   Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

   Describe the software used to analyze the data in this study. QTLtools was used in the analysis, which is referenced in the manuscript

Nature Genetics: doi:10.1038/ng.3981

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	N/A
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	N/A

Nature Genetics: doi:10.1038/ng.3981