The molecular basis, genetic control and pleiotropic effects of local gene co-expression

Authors

Diogo M. Ribeiro[1,2], Simone Rubinacci[1,2], Anna Ramisch[2,3,4], Robin J. Hofmeister[1,2], Emmanouil T. Dermitzakis[2,3,4], Olivier Delaneau[1,2]

[1]Department of Computational Biology, University of Lausanne, Switzerland,
[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[3]Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland
[4]Institute of Genetics and Genomics in Geneva, University of Geneva, Geneva, Switzerland
Corresponding author: olivier.delaneau@unil.ch

Abstract

Nearby genes are often expressed as a group. Yet, the prevalence, molecular mechanisms and genetic control of local gene co-expression are far from being understood. Here, by leveraging gene expression measurements across 49 human tissues and hundreds of individuals, we found that local gene co-expression occurs in 13% to 53% genes per tissue. By integrating various molecular assays (e.g. ChIP-seq and Hi-C), we estimated the ability of several mechanisms, such as enhancer-gene targeting, in distinguishing gene pairs that are co-expressed from those that are not. Notably, we identified 32,636 expression quantitative trait loci (eQTLs) which associate to co-expressed gene pairs and often overlap enhancer regions. Due to affecting several genes, these eQTLs are more often associated with multiple human traits than other eQTLs. Our extensive search for local gene co-expression opens the way to comprehend trait pleiotropy and comorbidity and provide functional interpretation of QTL and GWAS findings.

Introduction

Gene regulation is an essential component of cellular function, and its dysregulation often leads to disease[1,2]. The genetic makeup intrinsic to each person shapes their disease susceptibilities and response to treatment, making understanding the functional impact of genetic mutations one of the most pursued challenges in genetics research. Genome-wide association studies (GWAS) have associated tens of thousands of genetic variants with complex traits, but the functional link between them is unknown. In this context, expression quantitative trait loci (eQTL) analyses, which rely on studying the transcriptomic profiles of many individuals to reveal variants that modulate gene

expression, are pivotal in bridging the functional gap between genotypes and organismal phenotypes[2–4]. Indeed, in combination with GWAS, eQTL studies have been crucial in identifying causal genes and tissues associated with complex traits and disease[5]. Yet, understanding a variant's molecular link to complex traits and disease is still a major challenge, given that most are found in the genome's non-coding regions, act only in specific tissues and may affect several genes[6,7].

A pertinent clue in linking genome to phenome is the correlation of activity among regulatory elements and genes, which is key in understanding the shared genetic architecture between multiple complex traits and diseases as well as between all associated genes. In this respect, previous studies have found that gene order in eukaryotes is not random, and genes with similar expression profiles tend to be genomically linked[8]. Indeed, genes located near to each other (e.g. <1Mb) often display concerted co-expression[9–11]. Local gene co-expression is more pronounced in the immediate vicinity of a gene (e.g. <100 kb) but it has also been shown that it extends further and occurs regardless of strand, transcription orientation, shared functionality or tandem duplications[11,12]. The existence of structural and regulatory domains orchestrating the organised expression of nearby genes and mediating the genetic effects of regulatory variants on genes has been demonstrated[9], thus suggesting that regulatory element sharing and regulatory variant sharing may be key in controlling the co-expression of nearby genes. However, the molecular mechanisms, genetic control and tissue specificity of local gene co-expression are far from being fully understood. For instance, while genetic variants controlling expression have been found for almost all genes[2], the extent to which several genes may be controlled by the same genetic variants is yet unknown. Moreover, the full extent of local gene co-expression remains to be deeply assessed, particularly across tissues.

Here, we thoroughly investigate *(i)* the full genome-wide prevalence of local gene co-expression across human tissues, *(ii)* the molecular mechanisms that play a role in local gene co-expression, *(iii)* the regulation of local gene co-expression by genetic variants and *(iv)* their relevance on human disease and trait pleiotropy. For this, we developed a novel framework to detect local gene co-expression and associated regulatory variants using transcriptomic profiles across hundreds of genotyped individuals and 49 human tissues. To understand which molecular mechanisms play a role in local gene co-expression, we integrated various molecular features (e.g. chromatin contacts, enhancer targeting) and estimated their ability to distinguish gene pairs that are co-expressed from those that are not. Furthermore, our study identifies thousands of eQTLs that are predicted to regulate multiple co-expressed genes through shared regulatory elements as well as associated with multiple societal-relevant diseases and traits. Our extensive search for local gene co-expression reveals novel links between genetic variation and gene (co-)expression, opening the way to comprehend the

pleiotropy and disease comorbidity observed in humans and provide functional interpretation of QTL and GWAS findings.

Results

The prevalence of local gene co-expression across human tissues

Previous studies have found extensive local gene co-expression throughout the genome[10,11], but no consensus approach to identify local gene co-expression exists. Moreover, to date, studies of local gene co-expression were only performed on a limited number of individuals and tissues[11] and although gene expression measurements are often confounded by both known and unknown confounders[13], these did not account for either.

Here, we developed a robust approach to generate genome-wide maps of local gene co-expression from cross-individual gene expression quantifications (e.g. a gene expression matrix from multiple RNA-seq experiments). For this, we exploited the principle that nearby genes exhibiting significant expression correlation between each other across individuals can be considered as co-expressed (Fig. 1). Briefly, for each gene, our method identifies all genes in a cis window of 1Mb around the gene transcription start site (TSS) displaying substantial inter-individual expression correlation (Pearson correlation). Since the correlation value of truly correlated genes is unknown, we compare the observed values of each gene/cis-gene pair to expected correlation values under the null obtained by shuffling expression values, effectively ensuring that the number of nearby genes is accounted for (Methods). In this way, high observed correlation values can be pinpointed and sets of *bona fide* co-expressed genes extracted, while controlling for false discovery rate (FDR). Contrary to previous studies, this approach ensures that local gene co-expression can be identified regardless of gene order or transcriptional direction, which is particularly important given that gene regulation often occurs in a 3D manner. In addition, we extensively accounted for known (e.g. sex, subpopulation structure) and unknown confounding factors (using PCA or PEER[14]; Methods).

We first applied our approach to a dataset of gene expression profiles from 358 lymphoblastoid cell line (LCL) RNA-seq samples from the Geuvadis project[15], composed of European individuals genotyped in the 1000 Genome project[16] (Methods). We identified all correlated LCL-expressed gene pairs among 15,059 protein-coding and 1,781 long intergenic non-coding RNA (lincRNA) genes (Methods). At 1% FDR, we found 9,384 significantly co-expressed gene pairs (COPs) within 1MB of each other (8,716 correlated positively and 668 correlated negatively), which correspond to 9,030 distinct genes. A higher proportion of protein coding genes (54.8%) is co-expressed compared to lincRNAs (43.5%, Supplementary Fig. 1). Importantly, regardless of the FDR threshold used, we

observe that COPs are well spread across all chromosomes (Supplementary Fig. 2), with the percentage of genes co-expressed falling between 46% and 62% across chromosomes (FDR 1%). Altogether, this indicates that local gene co-expression is highly prevalent in human LCLs.

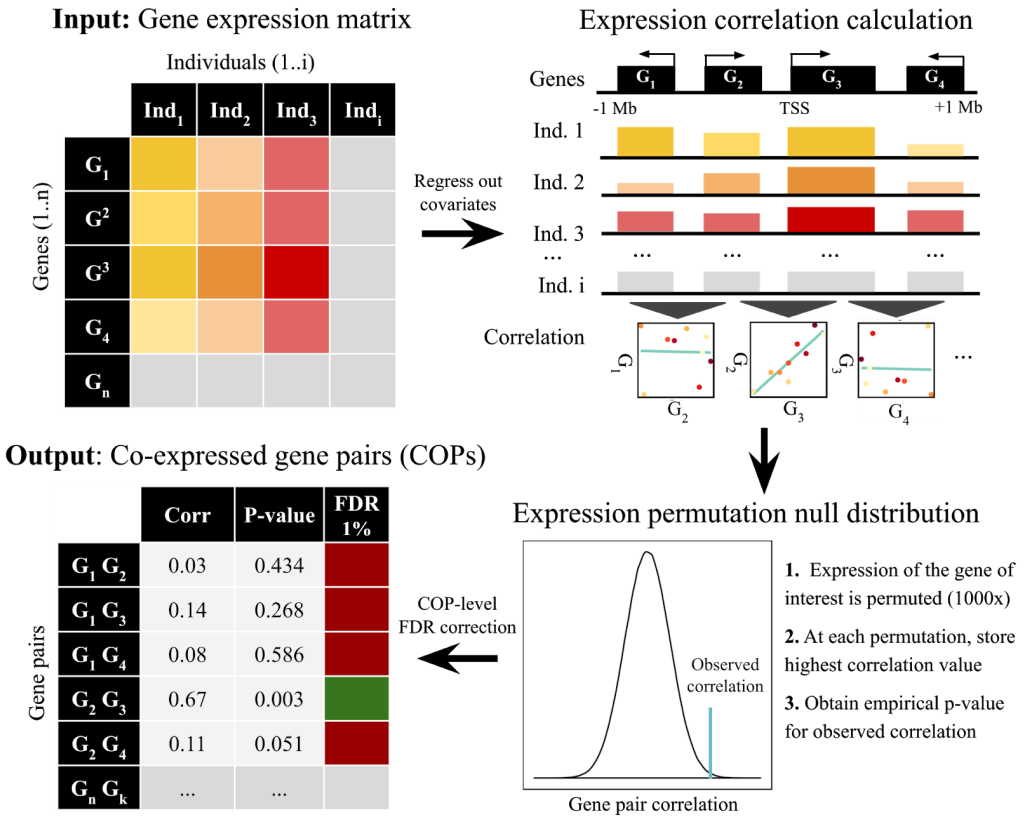


Fig. 1 Local gene co-expression identification approach. Inter-individual gene expression correlation is calculated for each gene pair whose TSSs are within 1Mb of each other. For each gene, observed correlation values are compared to expected correlation values when shuffling expression values across individuals. Gene pairs passing FDR threshold are considered COPs (Methods).

When grouping COPs into network components – nodes being genes and edges being the co-expression between genes – we found 2,545 unconnected groups of co-expressed genes (Fig. 2a). The majority of these groups are composed of only 2 or 3 genes (55.1% and 19.1% of the groups, respectively), yet, groups with more than 10 genes are also observed (3.5% of the groups). Several of these larger co-expressed gene groups can be readily assigned to known human gene clusters, such as the Hox A, B and C gene clusters[17], as well as a protocadherin cluster[18] and the histone H4 cluster[19] (Fig. 2a). To further understand how local gene co-expression is linked to shared functionality, we gathered data from several datasets of functional-relatedness (Methods). When compared to all gene pairs tested, COPs are significantly enriched for *(i)* genes annotated with the same exact function, based on the Gene Ontology (GO)[20] (Fisher's exact test p-value < $2.2e^{-308}$, odds ratio (OR) = 3.68) *(ii)* genes encoding proteins belonging to the same protein complex, based on CORUM[21] and hu.MAP[22]

databases (p-value = $2.1e^{-74}$, OR = 12.54), *(iii)* genes belonging to a same biological pathway, taken from KEGG[23] and Reactome[24] (p-value = $2.8e^{-166}$, OR = 2.52; Supplementary Fig. 3). Substantial functional-relatedness in local gene co-expression could be due to local duplication of genes. In fact, we also found that COPs connect known paralogs more often than what is expected by chance (Fisher's Exact test p-value < $2.2e^{-308}$, OR = 11.20, Supplementary Fig. 3, Methods). However, while these findings attest to the ability of our method in identifying known and expected groups of correlated genes, functional or evolutionary relatedness between gene pairs by no means explain all the local co-expression observed, as we could not find evidence of this link for the majority of COPs (71%, 6,698 out of 9,384 gene pairs).

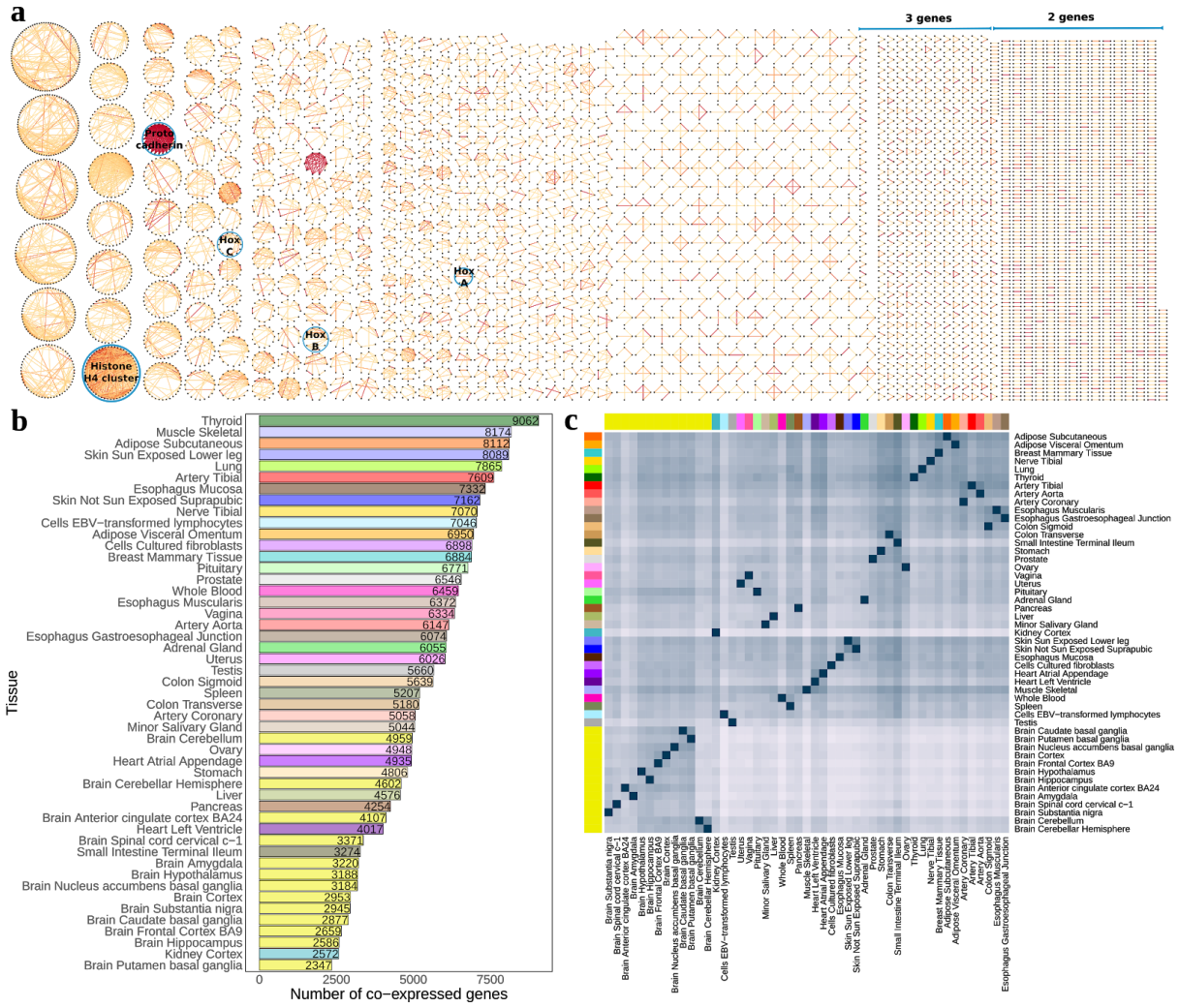


Fig. 2 Local gene co-expression discovery across tissues. a grouping of Geuvadis LCL COP discovery into 2,545 network components (genes as black nodes, correlation as edges), edge color is mapped to correlation strength and several known gene clusters are highlighted; **b** distribution of co-expressed genes across chromosome 1 for each GTEx tissue; **c** clustering of COP sharing across GTEx tissues. For each pair of tissues

(tissue A, tissue B), the percentage of COPs of tissue A that are also COPs in tissue B is calculated and used for clustering (hierarchical clustering).

To estimate the prevalence of local gene co-expression across human tissues, we applied our identification method to 49 human tissues from the GTEx v8 project[25] (Methods). We identified 64,320 distinct co-expressed gene pairs across tissues (FDR 1%). These amount to 18.3% of all gene pairs assessed for co-expression (350,564 gene pairs), which clearly demonstrate that local gene co-expression is highly prevalent. For individual tissues, the number of co-expressed genes ranged from 2,347 (Brain Putamen basal ganglia; 13% genes tested) to 9,062 (Thyroid; 50% genes) (Fig 2b). The discrepancy in the numbers of co-expressed genes may be due to distinct biological regulation of gene expression (e.g. 22,610 genes were found expressed in Testis against 15,051 in Whole Blood). However, we also found a strong correlation with the RNA-seq sample sizes available per tissue (Spearman rho = 0.75, p-value = $7e^{-10}$, Supplementary Fig. 4). This indicates that with large sample sizes more co-expression can be detected, as seen for LCLs from GTEx (147 samples, 44.5% genes) and Gevaudis (358 samples, 53.6% genes), even though, as expected, a substantial proportion of COPs overlap (29.2% gene pair overlap, Fisher's Exact test p-value $<2.2e^{-16}$, OR = 38.4; Supplementary Fig. 4). Importantly, we observe that the patterns of gene co-expression sharing reflect well our expectations in terms of biological similarity across tissues. For instance, by measuring the percentage of COPs in a tissue that are also COPs in another tissue, we find that related tissues such as artery aorta and coronary artery or various brain subregions are concertedly grouped (Fig. 2c). Tissue grouping by gene expression of single genes had previously produced similar results[2], which indicates that the local gene co-expression observed here encompasses relevant gene pair links and biological meaning. Our results about local gene co-expression across tissues, their correlation values, statistical significance and genomic information are all readily available in the LoCOP public database (http://glcoex.unil.ch).

Molecular features associated with local gene co-expression

Gene expression is known to be regulated by several genomic elements such as promoters, enhancers and insulators[26,27]. Here, we assess whether regulatory elements and mechanisms (termed here as 'molecular features') known to be associated with gene expression regulation can explain the observed inter-individual local gene co-expression. For this purpose, we train logistic regression models (on 80% of the gene pairs) and measure the area under the curve (AUC) on predictions of a test set (20% of the gene pairs) to compare how well local gene co-expressed genes pairs (i.e. positive cases) can be distinguished from non-co-expressed pairs (i.e. negative cases) for each assessed molecular feature (Methods). For instance, as reported in previous studies[11,26], genes that are co-expressed locally are more likely to be found in close proximity. Indeed, even though we searched

for co-expression of genes within 1Mb of each other, we found that most COPs are found in relatively closer distances, with 71.7% COPs found within 200kb from each other (Fig. 3a; median distance = 78,543 bp. Geuvadis LCLs). Predicting local gene co-expression through genomic distance alone provides an AUC of 0.81 (Supplementary Fig. 5). While clearly important, distance alone does not fully determine local gene co-expression, as only a fraction of nearby gene pairs are found co-expressed (e.g. 11.8% of gene pairs apart for <200 kb are co-expressed; 17.1% for 100 kb; 34.4% for 10 kb). This hints for the presence of fine-detailed regulatory mechanisms that determine which gene pairs are co-expressed and the presence of genomic structure (e.g. chromatin boundaries) which effectively separate in space and time the expression of nearby genes.

Given that many molecular features – such as DNA contacts and gene-enhancer interactions – are dependent on genomic distance, we compare COP's molecular features to those observed in a set of non-co-expressed gene pairs (non-COPs) that very closely matches the distance distribution observed in COPs (Fig. 3a; Methods). This ensures that the following results are independent of distance (i.e. distance has an AUC of 0.5). In addition, since paralog genes may display co-expression due to particular circumstances, such as the use of the same or a duplicated regulatory element, human paralog genes were excluded from this and subsequent analyses (Methods). Likewise, due to low sample size for cases of negative correlation, we focused on COPs with positive expression correlation.

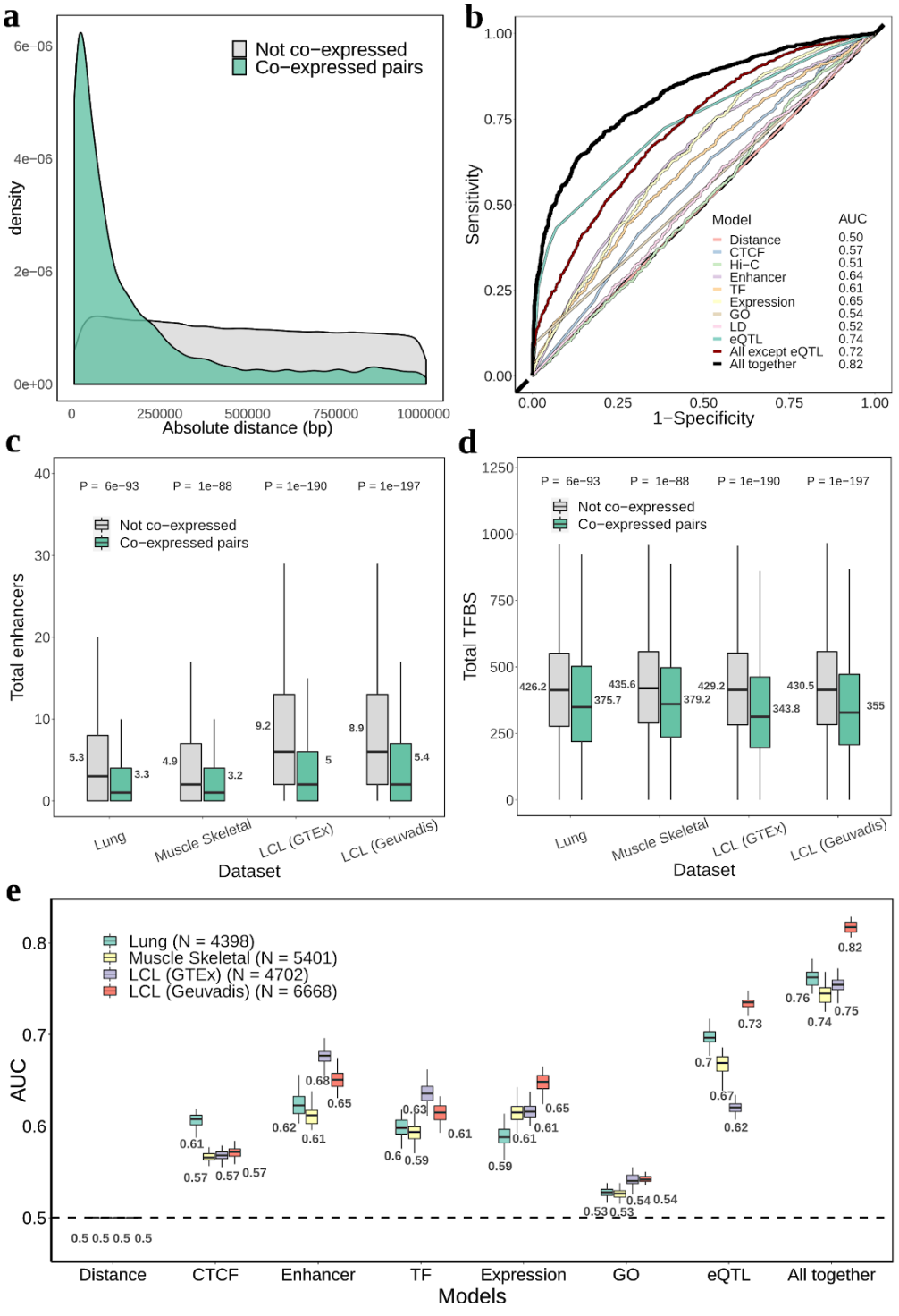


Fig. 3 Molecular features of COPs. a distribution of the absolute distance between TSS for Geuvadis LCL COPs and non-COPs. Gene pairs before applying paralog and positive correlation filters. **b** Receiver operating characteristic (ROC) curve of predicting Geuvadis LCLs COPs for several molecular features (logistic regression; N = 6668 for COPs and for non-COPs; see Methods for molecular feature descriptions). **c-d** boxplots of total enhancers and total transcription factor binding sites (TFBS), respectively, between COPs and non-COPs across 4 datasets: Geuvadis LCLs, GTEx LCLs, Lung and Muscle Skeletal. Values next to the boxplots represent the mean. P-values were obtained from two-tailed Wilcoxon signed-rank tests. **e** boxplots of

the AUC values obtained for each molecular feature and dataset across the 50 training-test set randomisations. Values below each boxplot represent the mean AUC.

We first analysed COPs derived from Geuvadis LCLs, for which there is a wealth of molecular annotations. When measuring the presence of total and inverted CTCF binding sites between gene TSSs (suggestive of insulation) we found them to be clearly depleted in COPs when compared to distance-matched non-COPs, in particular when distance increases (Supplementary Fig. 6; Methods). This makes CTCF binding site presence reasonably predictive of local gene co-expression (mean AUC = 0.57 for all COPs, 0.63 for those >200kb apart; Fig. 3b; Supplementary Fig. 7), which indicates possible constraints in the chromatin structure in which some gene pairs may be hindered from interacting. Looking more closely at chromatin interactions, we found that chromatin contacts between gene TSSs (Hi-C data, 5kb resolution; Methods) are higher in COPs than non-COPs when these are separated by more than 200kb (mean AUC = 0.59), whereas at closer distances such contacts occur with similar intensities in COPs and non-COPs (mean AUC = 0.50; Supplementary Fig. 6 and 7). These results indicate that chromatin structure may play a role or be a consequence of gene co-expression, yet, these do not seem determinant in selecting which local gene pairs are co-expressed or not.

Next, we assessed whether enhancer targeting and transcription factor binding explain local gene co-expression. Interestingly, COPs were found to be targeted by lower numbers of enhancers than non-COPs (LCL-active enhancers; Fig. 3c). Moreover, a similar pattern was observed when analysing transcription factor (TF) binding in the promoter/enhancer region of genes (50kb around gene TSS, Methods), where again we found that COPs have less TF binding sites (TFBS) than non-COPs (Fig. 3d). These results were readily replicated in LCLs from the GTEx project (Fig. 3c,d), as well as in two other GTEx tissues for which tissue-specific enhancer datasets were available – Skeletal Muscle and Lung (mean AUCs between 0.59 and 0.68, Fig. 3c-e). Interestingly, when we split COPs into four categories based on their tissue prevalence across 49 tissues (Supplementary Fig. 8; Methods), we observed a clear trend in which conserved COPs generally have higher AUCs across molecular features than specific or unique COPs (Supplementary Fig. 9). Indeed, the depletion of TF binding sites and enhancer targeting is more pronounced the more conserved the COPs are (Supplementary Fig. 9). Importantly, these results are consistent for COPs and annotations from Muscle Skeletal, Lung and LCL (Supplementary Fig. 9-11). An hypothesis for this finding is that there is a driving force for COPs to reduce the total number of expression regulators, i.e. a pressure to keep regulatory machinery relatively simple in order to achieve similar levels of expression between genes. Indeed, another feature of COPs is that the expression level and variation between the two genes in the pair tend to be more similar than when compared to non-COPs (mean AUC between 0.59 and 0.65, Fig. 3e,

Supplementary Fig. 12). Importantly, this is observed even when matching the expression levels between COPs and non-COPs (Supplementary Fig. 12).

We confirmed that the molecular features observed for COPs are largely independent of the linkage disequilibrium (LD) between the genes' promoters (mean AUC = 0.52, Geuvadis LCLs, Fig. 3b; Methods). Indeed, the genetic linkage of COPs and non-COPs does not substantially differ, as measured by their TSS centimorgan distance (Supplementary Fig. 13). Moreover, since paralog genes were excluded from analysis, the sharing of function between gene pairs also had low AUCs (0.53 to 0.54 for Lung, Muscle Skeletal and LCL, Fig. 3e), as measured by the two genes in the pair sharing the same exact Biological Process GO term (Methods). Interestingly, when combining all the molecular features assessed in Geuvadis LCLs together into the same regression model we obtain a global mean AUC of 0.72 (Fig. 3b). This AUC is higher than any individual molecular feature, indicating that each feature explains local gene co-expression in a complementary way. Indeed, except for Hi-C contacts and redundant metrics, each distinct molecular feature significantly explains some of the variation in local gene co-expression independently of the other features (Supplementary Table 1) and no combination of two features reaches AUC levels above 0.69 (the highest being total enhancers in combination with expression level similarity; Supplementary Fig. 14). Albeit reasonable, a global AUC of 0.72 also indicates that the molecular features or the datasets used here are yet far from explaining all the observed local gene co-expression.

One way to assess the effect of known and unknown molecular features and mechanisms implicated in gene regulation is to exploit the natural genetic variation observed in human populations (e.g. eQTLs). To assess how well genetic variation can provide information about gene co-expression, we identified expression Quantitative Trait Loci (eQTLs) in cis (+/- 1Mb) for all expressed human genes, i.e. discovered associations between genetic variants and gene expression levels for all genes across all tissues (FDR 5%; Methods). To discover putative cases of co-regulation of gene pairs by the same genetic variant(s), for each gene pair, we tested whether the lead eQTL of each gene (i.e. the genetic variant most strongly associated to the gene, if any) is also associated to the other gene in the pair (nominal p-value < 0.05; Methods). For simplicity, we name such associations of a variant to gene pairs as 'eQTL sharing'. Through this approach, we found that as many as 41.9% of all COPs display eQTL sharing in Geuvadis LCLs (358 genotyped individuals), compared to only 5.9% in distance-matched non-COPs (Supplementary Fig. 15). Given that eQTL discovery power is largely dependent on sample size and only 46.2% genes (7,805 out of 16,906 expressed genes) are eGenes (i.e. significantly associated with a cis variant) in Geuvadis LCLs, the high number of co-regulated COPs is striking. In fact, if only gene pairs composed of eGenes are considered, the percentage of COPs in eQTL sharing increases to 85.4%, compared to 20.0% in non-COPs (Supplementary Fig. 15).

Moreover, we observed that eQTL sharing occurs more often for local co-expressed genes than genes that are co-expressed in trans at similar correlation values (trans-COPs, Methods). Indeed, only 8.9% of trans-COPs were found in eQTL sharing (compared to 41.9% cis-COPs; Supplementary Fig. 15), suggesting that most eQTL sharing found for cis-COPs may represent co-regulation of multiple genes by the same variant rather than simple transient correlation between expression values and genotypes.

In the regression model, eQTL sharing leads to mean AUCs between 0.62 and 0.73, depending on the tissue assessed (Fig. 3e). Importantly, by combining eQTL sharing and all previously assessed molecular features, we increase the global mean AUC to between 0.74 (Muscle Skeletal) and 0.82 (Geuvadis LCL, Fig. 3b,e). This increase in the global model AUC indicates that by measuring eQTL sharing we are able to capture gene regulation that was not yet assessed with the other molecular features. Overall, these results highlight the complexity and diversity of mechanisms involved in the regulation of gene co-expression, and indicate a high level of control of co-expression regardless of the genomic distance between genes.

Genetic regulation of local co-expressed genes

The integration of QTL analysis and functional annotations provides insights into the molecular mechanisms of transcriptional regulation and their phenotypic consequences. We thus next investigated whether shared eQTLs are more likely to fall in regulatory regions of the human genome than lead eQTLs that associate with a single gene. First, as expected, we found that all Geuvadis LCL lead eQTLs (shared or not) are more likely than expected by chance to fall within gene body regions as well as various LCL-specific functional regions from Encode[28] such as enhancers (OR = 1.8-2.7, p-value = $1e^{-4}$), DNAse sensitive regions (OR 1.8-2.4, p-value = $1e^{-4}$) and protein binding regions (OR = 1.5-1.9, p-value = $1e^{-4}$, Fig. 4a; Methods). However, we found that shared eQTLs in LCL are more highly enriched in these three functional annotations than other lead eQTLs, particularly for enhancers (Fisher's Exact test OR = 1.7, p-value = $3e^{-7}$, Fig. 4b). Remarkably, an enrichment for shared eQTLs overlapping enhancers more than other eQTLs can also be observed for GTEx LCLs (Fisher's Exact test OR = 1.96, p-value = $4e^{-3}$, Fig. 4c), as well as for functional annotations from the Roadmap Epigenomics project[28,29] for LCL, Muscle and Lung tissues (Fisher's Exact test OR = 1.3-1.4, Supplementary Fig. 16-19). Crucially, discovering genetic variants affecting regulatory regions such as enhancers, which are known to target several genes, can provide a mechanistic explanation of how such variants regulate multiple genes. To further evidence that COPs are co-regulated by the same genetic variants, we performed colocalization analysis between the eQTL signals of the two genes in the pair using COLOC [30] (Methods). As expected, a large portion of COPs in eQTL sharing show evidence of common eQTL signals (45.6% COPs, coloc H4 posterior probability (PP4) > 0.5), whereas COPs that are not in eQTL sharing show very low evidence of common eQTL signals (0.9%

COPs with PP4 > 0.5, Supplementary Fig. 20). Importantly, COPs (in eQTL sharing or not) also evidence more common eQTL signals than non-COPs, even when considering only pairs of eGenes in both categories (PP4 > 0.5 in 42.4% of COPs compared to 5.4% for non-COPs, Supplementary Fig. 20).

To study the sharing of genetic regulation across all 49 GTEx tissues, we extended our approach to identify all lead eQTLs affecting the expression of COPs to all tissues (5% FDR, Methods). We found that 63% (85,460 out of 135,662) COP-tissue pairs have a lead eQTL that is significantly associated with both genes, illustrating the widespread sharing of genetic effects through local co-expressed genes. Per tissue, between 6.2%-89.3% (median = 46.2%) COPs have a shared lead eQTL (Fig. 4d), the large spread between tissues being largely due to sample size differences (Spearman rho = 0.78, p-value = $3.5e^{-11}$, Supplementary Fig. 21). As a whole, these results indicate that the co-regulation of multiple genes by single genetic variants is widespread and often made through shared regulatory elements.

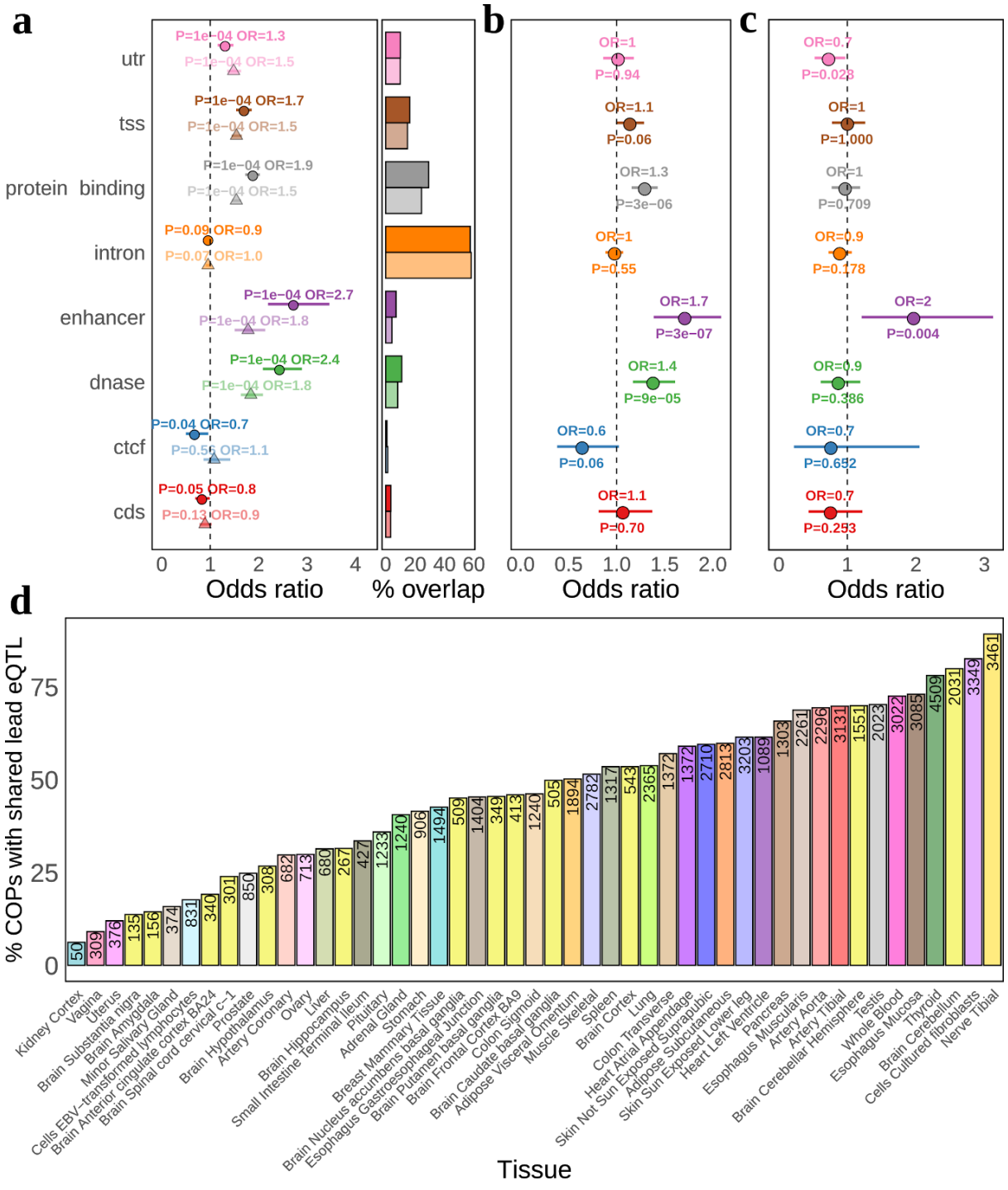


Fig. 4 Discovery and functional enrichments of shared eQTLs. a overlap enrichment of Geuvadis LCL shared lead eQTLs (solid color, round points) and other lead eQTLs (pale color, triangles) in Encode LCL functional annotations and Gencode gene bo-dy categories (Methods). Odd ratios are calculated based on the observed versus expected overlap between eQTLs and each functional annotation. Error bars are from 10000 QTLtools *fenrich* permutations. The right part of the plot denotes the percentage of overlap between eQTLs and each functional annotation; **b,c** Fisher's exact test odds ratio and p-value for the enrichment of shared lead eQTLs in each functional annotation, compared to other lead eQTLs, for eQTLs in Geuvadis LCLs and GTEx

LCLs, respectively. Error bars are 95% confidence intervals; **d** percentage of COPs with a shared lead eQTL per GTEx tissue. Numbers inside bars denote the number of COPs with eQTL sharing.

Shared gene regulation and trait pleiotropy

Recent GWAS and PheWAS studies have uncovered extensive pleiotropy of complex trait associations, finding that a high proportion of variants are associated with multiple traits[31]. A plausible explanation for this is the association of variants to multiple genes. Indeed, cis-eQTLs that affect the expression of multiple genes have been recently shown to have high complex trait pleiotropy[25]. Given the abundance of genetic variation co-regulating multiple genes across tissues found in our study (shared eQTLs), we sought to investigate how pleiotropy can be driven by gene co-regulation. For this, we first collected GWAS summary statistics for 35 traits from the PanUK Biobank including binary and continuous data types, from basic anthropometric measurements to frequently studied clinical and complex traits with more than 15,000 cases (Supplementary Table 2; Methods). Association summary statistics from the PanUK Biobank are provided for the same set of 28,987,534 variants across traits and these include the vast majority (96.8%) of all distinct lead eQTLs found across tissues. Then, for each trait we only retained genome-wide significant associations (p-value < $5e^{-8}$), resulting in 373,206 variants, of which 109,167 (29.3%) are associated with more than one of the 35 traits. Considering the 32,636 unique lead eQTLs across tissues associated with COPs (eQTL sharing), we found that 4,301 (13.2%) are associated with a trait, with 1,274 (29.6% of those) associated with several traits (Table 1). This compares to 10,633 (9.4%) of other lead eQTLs (not shared) being associated with a trait, 2,647 (24.9% of those) with multiple traits. Shared eQTLs are thus significantly more likely to be associated with at least one trait (Fisher's Exact test OR = 1.5, P-value < $2e^{-16}$), as well as with more than one trait (OR = 1.7, P-value < $2e^{-16}$). Indeed, out of all lead eQTLs associated with traits, those that are shared are more pleiotropic than other lead eQTLs (mean 1.5 for shared, 1.38 for other lead eQTLs, Wilcoxon rank sum test p-value = $3e^{-11}$, Fig. 5a). We confirmed that shared eQTLs have increased pleiotropy independently of their increased presence in multiple tissues (two-way ANOVA p-value = 1.4e-11). Indeed, the presence of the same lead eQTL across multiple tissues, previously found to have higher pleiotropy than tissue-specific eQTLs[25], results only in a mild increase of pleiotropy in our dataset (mean 1.45 for multiple tissues, 1.4 for tissue-specific, Wilcoxon rank sum test p-value = 0.06, Supplementary Fig. 22). This suggests that pleiotropy can arise due to the co-expression and co-regulation of nearby genes, for example, through the effect of several genes implicated in diverse physiological processes.

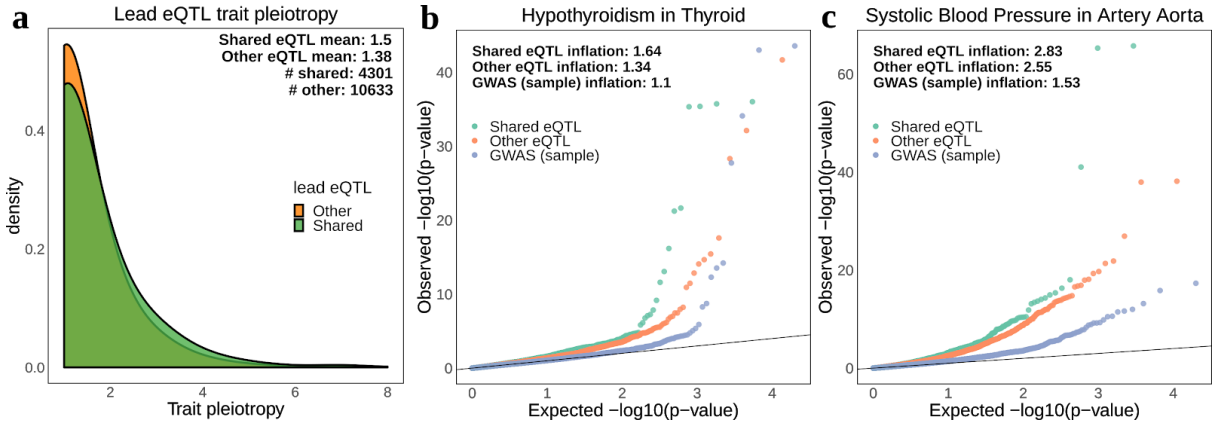


Fig. 5 Comparison of trait pleiotropy and association p-values between shared and other lead eQTLs. a distribution of trait pleiotropy (association p-value < $5e^{-8}$) across 35 UKBB GWAS traits. If a lead eQTL is shared in at least one tissue (even if not in other tissues) it is considered as "shared". The remaining lead eQTLs compose the "other" category. Only lead eQTLs associated with at least one trait were considered; **b,c** Q-Q plots of associations p-values (-log10 scale) for known tissue-trait matched pairs (hypothyroidism in thyroid tissue and systolic blood pressure in artery aorta), comparing shared (green) and other (orange) lead eQTLs. GWAS (blue) is a sample of 10000 variants (randomly and independently picked for each trait) shown only for comparison purposes.

Table 1. Number of lead eQTLs associated with one or more traits. Considering variants associated with traits with association p-value < $5e^{-8}$. A lead eQTL is defined as 'shared' if in at least one out of 49 GTEx tissues it is found shared.

Lead eQTL category	Total # eQTLs	Associated to a trait	Associated to >1 trait
All lead eQTLs	145327	14934 (10.3%)	3921 (26.3% of previous)
Shared lead eQTLs	32636	4301 (13.2%)	1274 (29.6% of previous)
Other lead eQTLs	112691	10633 (9.4%)	2647 (24.9%)

Next, we investigated whether shared lead eQTLs are strongly associated with the 35 traits analysed here, by comparing the probability distributions of all variants prior to any p-value filtering. First, as found in previous studies[5,32], we observed that lead eQTLs (shared or other) are more likely to display stronger associations (i.e. lower association p-value) than other variants, as observed in quantile-quantile plots (Q-Q plots) and measured by the genomic inflation factor (ƛ metric, Fig. 5b, Supplementary Fig. 23). Notably, we found that shared lead eQTLs often have stronger associations than other lead eQTLs, as it can be observed in known tissue-trait pairs, such as hypothyroidism-thyroid (ƛ = 1.64 for shared, ƛ = 1.34 for other eQTLs) and systolic blood pressure - artery aorta (ƛ = 2.83 for shared, ƛ = 2.55 for other eQTLs) (Fig. 5b). Moreover, considering all shared and other lead eQTLs across the 49 GTEx tissues, we found stronger association p-values (higher ƛ) for shared lead eQTLs across for 34 out of the 35 traits assessed, which is a remarkably

consistent result (Supplementary Fig. 23). Together, these findings suggest that eQTLs affecting more than one gene play an important role in (multiple) complex traits and disease, and can help reveal novel links between traits which may explain disease and trait comorbidity.

Through our analysis, we identified a total of 1,274 unique variants (in 3,201 eQTL-COP-tissue combinations) that affect multiple genes as well as multiple traits (Supplementary Table 3). Of these variants, 219 fall in promoters, and 26 in enhancer regions, a higher overlap proportion than found for other lead eQTLs that are also pleiotropic (Fisher's exact test p-value = $1e^{-5}$ and 0.05, respectively, Supplementary Fig. 24, Ensembl v101 Regulatory Build[33]). This overlap to functional regions denotes plausible mechanistic explanations linking genetic variation to the expression of several genes which in turn affect multiple traits. Overall, our approach allows us to gather novel information into the genetic architecture of disease loci, and its exploration can open the way to comprehend trait pleiotropy and comorbidity, aiding the interpretation of QTL and GWAS findings.

Discussion

Prevalent genome-wide local gene co-expression had been previously reported in human, but limited to a few tissues or cell types[10,11]. In this study we confirm that local gene co-expression is highly prevalent and we provide the community with a database containing an extensive catalog of local gene co-expression and shared eQTLs across 49 human tissues. Previous studies reported a genome-wide level of 37.4% co-expressed genes (out of 20,502 genes assessed) in 100 normal breast tissue samples[10]. This is remarkably consistent with the 38.1% of co-expressed genes observed here for GTEx breast mammary tissue (with 396 samples), even though distinct samples, approaches (e.g. correcting for confounding factors) and filters (FDR control) were applied in our study.

In this study we attempted to discover the relevance of several molecular features for local gene co-expression, replicating results in various tissues and datasets. As expected from the perceived complexity of gene regulation, we found that several factors contributed to local gene co-expression and provided novel insights into cis-regulation. Of particular relevance was the finding that high regulatory complexity, e.g. how many nearby TFBS and enhancers, seems to hinder gene co-expression, particularly in COPs prevalent in many tissues. In agreement with our findings, a recent study identifying cohesin chromatin loops also evidenced the usage of a simpler circuitry in constitutive genes which have a steady level of expression, whereas more extensive regulatory architectures were found in dosage-sensitive genes[34].

A limitation of our approach rests in our inability to disentangle whether the observed molecular features (e.g. enhancer targeting, CTCF binding) actually cause the observed gene co-expression or

whether they are its consequence. To discriminate these, molecular experiments would have to be performed (e.g. massively parallel reporter assays and CRISPR/Cas9-based technologies[27,35]). The similar expression level and reduced expression variation found between the two genes in a COP is hard to discriminate between cause or consequence. On one hand, selection for reduced expression noise has been previously proposed as a cause for the grouping of genes in the genome, particularly for essential genes[36]. On the other hand, local gene co-expression can simply be a consequence of stochastic chromatin fluctuations (i.e. similar expression patterns can be a consequence of genes being close together). Indeed, Kustatscher *et al.* 2017 has found evidence for both cases, although most local gene co-expression was shown to be buffered at the protein level[37].

Our approach is affected by the amount and quality of molecular data available for distinct tissues. This could explain some of the differences in AUCs observed, such as lower AUCs for enhancer features for Lung and Muscle skeletal (AUC = 0.57), as only 43,973 and 39,708 enhancer-target annotations were used for Lung and Muscle skeletal, respectively, compared to 78,796 for LCLs (AUC = 0.61 for GTEx and 0.64 for Geuvadis LCLs)[38]. Indeed, given the difficulty in obtaining complete tissue-specific regulatory maps (e.g. all enhancer-gene interactions relevant in a certain tissue) and the added complexity of regulatory mechanisms (e.g. we did not include CpG islands and promoter types in our study), we did not fully account for all possible regulatory mechanisms in this manner. Instead, we reasoned that identifying natural genetic variation and associating it with the expression of several genes (shared eQTLs) would allow us to map genomic regions pertaining both known (e.g. enhancers) and unknown features of gene co-regulation. Indeed, the AUCs for eQTL sharing shown to be higher than for any other metric assessed, which indicates the feasibility and benefit of this approach to study the regulation of local gene co-expression.

While our approach to find variants affecting multiple genes may be overly simplistic (e.g. only lead eQTLs are used and causality is not assured) and limited by sample size, we have nevertheless found extensive evidence of joint control of co-expressed genes through the same genetic variant. Indeed, we found that shared eQTLs *(i)* often overlap enhancers, *(ii)* are more strongly associated with GWAS traits and *(iii)* are more likely to affect more than one trait. These findings thus provide insights into the molecular underpinnings of pleiotropy: variants that affect the expression of multiple genes by overlapping regulatory elements display a higher degree of complex trait pleiotropy. This indicates that the pleiotropy reported in previous studies[31] may emerge from cis regulation, besides the possibility of it arising from multiple functions of genes, particularly under distinct tissues and cellular contexts. Moreover, describing COPs and their co-regulation can aid in establishing the direction of causality between traits that were previously found to be comorbid. Characterizing how variants are functionally linked to disease is still a challenge but approaches such as ours provide an

additional layer of interpretation, by linking groups of genes regulated together and describing how regulatory signals may be propagated. As current colocalization methods are only focused on the sharing of GWAS and eQTL signals for single genes, the development of methods dedicated to pinpointing causal variants co-regulating multiple genes and/or other molecular phenotypes may prove an asset for future research.

Methods

Geuvadis consortium gene expression dataset

Geuvadis consortium[15] BAM files previously mapped to GRCh37 for RNA-seq experiments on lymphoblastoid cell lines (LCL) were used for COP identification. Only a subset of 358 European (EUR) individuals from the Geuvadis study also present in the 1000 Genomes project phase were considered. Data was downloaded from the EBI ArrayExpress (accession code E-GEUV-1). Gene expression was quantified for all protein-coding and long intergenic non-coding RNAs (lincRNAs geness annotated in GENCODE v19[39] using QTLtools[40] v1.1 *quan* function with default parameters. Genes within or around the MHC complex region (chr6:29500000-33600000) and in non-autosomal chromosomes (X and Y), as well as pseudoautosomal regions (PAR1, PAR2) or unassembled regions were removed. Moreover, genes with no expression measurements across most individuals (>=50% of the individuals with RPKM = 0) were also excluded. To account for confounding factors, the following covariates were regressed out: *(i)* sample sex (as defined by the 1000 Genomes project), *(ii)* ancestry (3 first PCA principal components (PC) computed using QTLtools *pca* function on genotype data that has been trimmed for one variant with minor allele frequency (MAF) > 5% every 50kb) and *(iii)* unknown technical/experimental variables based on the first 50 PCA PCs computed using QTLtools *pca* function on the gene expression matrix. The 50 PCs were determined as the number of PCs that maximizes the number of eGenes discovered. Importantly, residualising gene expression by several principal components (or PEER factors), a practice commonly used in eQTL mapping[2] to improve discovery power, is expected to reduce trans gene co-expression effects, leading to increased power in detecting weaker effects of local gene co-expression. Finally, the expression quantifications across individuals were normalized to match a normal distribution $N(0,1)$.

GTEx project gene expression dataset

Gene expression quantifications (TPM values) from RNA-seq experiments across 49 tissues (for which genotype data is also available for >=70 individuals) processed and provided by the GTEx project v8[25] were used for COP identification. Data was downloaded dbGaP (accession:

phs000424.v8.p2). The provided quantifications had been mapped to Gencode v26[39] gene annotations on hg38 and normalised by TMM between samples (as implemented in edgeR), and inverse normal transform across samples. Moreover, only genes passing an expression threshold of >0.1 TPM in >=20% samples and >=6 reads in >=20% samples had been retained. As done for the Geuvadis dataset, only protein coding and lincRNA genes were considered in this study, and genes in the MHC complex region and in non-autosomal or pseudoautosomal regions were removed. Covariates provided by GTEx v8 for each tissue were regressed out of each expression matrix to account for potential confounding factors. These included 15-60 Peer factors (depending on tissue sample size), 5 Genotype PCA PCs as well as information about the sequencing platform, PCR usage, and the sex of the samples, all provided by GTEx.

Identification of co-expressed gene pairs

A novel framework for robustly identifying local gene co-expression was developed in this study. The input for this method is a gene expression matrix across individuals (e.g. genes as rows, individuals as columns) including gene coordinates (chromosome and TSS position). In this study the input expression matrices were previously pre-filtered, normalised and corrected for covariates (see above). Gene pair Pearson correlation across individuals is then calculated for all gene pairs whose TSS's are less than 1Mb apart. This window size is commonly used in genetic studies for assessing cis effects and allow to test the vast majority of genes. For instance, in Geuvadis LCLs 16,840 genes have a neighbour within 1Mb, out of the 16,907 genes assessed. The total number of gene pairs tested are thus 224,267 gene pairs for the Geuvadis dataset and 350,564 gene pairs for the GTEx dataset across all tissues (e.g. 242,084 for Lung and 192,637 for Muscle Skeletal). Of note, the gene pair correlation is calculated regardless of gene order, transcription directionality and coordinate overlap, thus accounting for all possible co-expression occurring in a certain cis window. To account for the variable number of genes neighbouring in a cis window the following procedure was applied: *(i)* for each gene, it's cross-individual expression values were shuffled and the correlation with each other gene in the cis window was recalculated 1000 times (while keeping the expression values of the other genes intact); *(ii)* in each randomisation, the highest correlation value across cis genes is kept in order to build a null distribution; *(iii)* an empirical p-value ('adjusted p-value') is obtained for each gene pair for the correlation value observed with the real data to be higher than the correlation values obtained in the randomisations. Notably, this procedure adjusts for the total number of neighbouring cis genes while ensuring that the correlation structure between them is unchanged. To control for the total number of genes tested with this approach, the Benjamini-Hochberg procedure was applied on the adjusted p-values corresponding to the highest correlation values observed for each gene. The adjusted p-value corresponding to FDR 1% was kept and any gene pair whose adjusted p-values are

below this adjusted p-value were considered significantly co-expressed gene pairs. Note that in GTEx this procedure is performed for each tissue separately. Extremely high correlation (>=0.99) was found to be mapping artefacts and were removed from analysis (164 gene pairs in Geuvadis, none for GTEx). Finally, co-expressed gene pairs were grouped into a network using igraph v0.7.1 on Python3.6 and plotted with Cytoscape 3.8.0[41].

Enrichment to functionally and evolutionarily-related datasets

The overrepresentation of Geuvadis LCL COPs (unfiltered, positive and negative correlation) as functional-related gene pairs was assessed with one-way Fisher's Exact tests to several datasets: *(i)* genes annotated with the same exact 'Biological Processes' annotation of the Gene Ontology (GO)[20], obtained from Ensembl v98 via the BioMart interface; *(ii)* genes belonging to the same biological pathway, gathered from KEGG[23] and Reactome[24] through the Ensembl v98 BioMart data mining tool[42] (25-May-2020); *(iii)* genes belonging to the same human protein complex, gathered from the CORUM 3.0 database[21] and hu.MAP[21,22] (20-April-2020). UniprotKB IDs were converted to Ensembl IDs with the Uniprot ID mapping tool[43]. Moreover, to assess overrepresentation of co-expressed gene pairs in evolutionarily-related human genes, a compendium of human paralog gene pairs was gathered from *(i)* paralog gene pairs from the Ensembl v98 BioMart data mining tool[42], *(ii)* human gene pairs belonging to the same OrthoMCL v6[44] cluster and *(iii)* gene pairs belonging to the same group of the Duplicated Genes Database[45] (last update: June 19, 2015). This compendium of 11,304 local (within 1Mb) paralog gene pairs was used to filter evolutionarily-related genes in downstream analysis.

COP dataset filtering

Given that neighbouring paralog genes may display co-expression due to usage of the same or copied regulatory elements, all 11,304 local paralog genes gathered from Ensembl, OrthoMCL and the Duplicated Genes Database (see above) were excluded from analysis (before COP discovery phase). Moreover, only COPs with positive expression correlation were used for molecular feature analysis and subsequent analysis (>90% of COPs from each dataset). This is because positive and negative gene pair correlation may be regulated in distinct manners and the number of negatively correlated genes is too low for an extensive analysis of only negative correlation (e.g. only 7.1% COPs are negatively correlated in Geuvadis LCLs). In addition, only COPs that passed significance cutoffs (FDR 1%) in both gene pair comparisons assessed (i.e. gene1-gene2 and gene2-gene1) were considered (<1% of COPs per dataset were removed in this way). These filters were applied equally to both Geuvadis LCL (from 9,384 to 6,668 COPs) and all GTEx tissues (from 64,320 to 40,999 distinct COPs across tissues).

Creation of non-COP and trans-COP controls

To control for distance effects in local gene pair co-expression, a set of non-co-expressed gene pairs (non-COPs) was built for each dataset and tissue. For this, for each COP, a gene pair that is not co-expressed (non-COP) but is matched for the distance between TSSs observed in the COP is randomly sampled. The maximum distance discrepancy allowed between a COP and a matching non-COP is 100bp. Non-COPs are picked by random without replacement. To ensure that non-COPs are not COPs that simply did not pass FDR 1%, only non-COPs with adjusted p-value > 0.5 were randomly sampled. Given the large amount of non-COPs available (~95% of all gene pairs tested), in no case a COP is lost for not having a matching non-COP in Geuvadis and in most GTEx tissues, exceptions being 4 gene pairs in Adipose_Subcutaneous, 8 in Muscle_Skeletal and 6 in Thyroid.

To compare the proportion of eQTL sharing in cis-COPs and trans-COPs in Geuvadis LCLs, a set of trans-COPs was defined, consisting of all gene pairs (regardless of genomic location) that have correlation above 0.143 (the minimum correlation value for significant cis-COPs) and are not cis-COPs. Then, for each cis-COP, a random trans-COPs with a correlation value at most 5% different from the cis-COP correlation value was selected. 6316 cis-COP/trans-COP matches were liable to be created in this manner.

Molecular feature metrics and datasets used

The following metrics were assessed for their potential to regulate local gene co-expression. These were calculated in the same way for the Geuvadis LCL dataset and 3 GTEx tissues – Muscle Skeletal, Lung and Cells EBV-transformed lymphocytes (equivalent to LCLs) – unless otherwise stated.

(a) Total CTCF sites: total number of CTCF binding sites found between the TSSs of both genes in the gene pair. Tissue or cell-line specific data was collected from the ReMap 2018 v1.2 database[46], which includes Encode and other public datasets of ChIP-seq experiments. MYOBLAST data was used as a proxy for the GTEx Muscle Skeletal tissue, whereas Lung and LCL (GM12878) datasets were readily available.

(b) Inverted CTCF motifs: number of CTCF motifs found in opposite orientations (i.e. pairs of '+' and '-' CTCF motifs) between the TSS of both genes in the gene pair. Note that this does directly count the number of CTCF motifs in a convergent orientation, but is rather intended as a proxy for the number of possibilities to form DNA loops. Transcription factor motifs mapped to hg19 were downloaded from the MotifMap database[47] in August 2019. The following CTCF motifs were used: 'LM2_CTCF=CTCF', 'M01200=CTCF', 'M01259=CTCF' and 'MA0139=CTCF'.

(c) Hi-C contacts: KR normalized Hi-C contacts between the 5kb bins encompassing the TSSs coordinate of each gene in the gene pair. Data from the GM12878 cell line (LCL) at 5kb resolution was used[48]. This metric was only calculated for the Geuvadis LCL dataset.

(d) Enhancer sharing: number of common enhancers targeting both genes in the gene pair. Enhancer-gene targeting data was downloaded from the Enhancer Atlas V2.0[49] (Mar 2020). Data from the GM12878 cell line was used for Geuvadis and GTEx LCLs as well as human 'Skeletal_muscle' and 'Lung' data for the respective GTEx tissues.

(e) Total enhancers: number of distinct enhancers targeting each gene in the gene pair. If the same enhancer targets both genes in the pair, it is counted once. Data downloaded from the Enhancer Atlas V2.0, as above.

(f) Shared transcription factors (shared TFs): number of distinct transcription factor motifs found within +/-50kb flanking the respective TSS of both genes in the gene pair (i.e. number of distinct TF motifs in common between the two genes). Transcription factor motifs mapped to hg19 were downloaded from the MotifMap database[47] in August 2019, and amount to 607 different transcription factor motifs.

(g) Total transcription factor binding sites (total TFBS): sum of number of transcription factor binding motifs (including redundant and overlapping motifs) within +/-50kb flanking of each gene TSS in the gene pair. All 607 different transcription factor motifs from the MotifMap database were used.

(h) Expression level difference (Diff. expr. level): relative difference in expression between each gene in the gene pair, calculated as the absolute difference between the average expression level of the genes in the pair, divided by the mean expression level of the pair. For the Geuvadis dataset, mean raw RPKM values were used, whereas for GTEx tissues, median raw TPM values were used.

(i) Expression coefficient of variation difference (Diff. coef. var.): absolute difference between the expression coefficient of variation of the two genes in the gene pair, divided by the mean coefficient of variation of the gene pair, calculated from cross-individual raw RPKM values. This metric was only calculated for the Geuvadis LCL dataset.

(j) GO term sharing: total number of Biological Process (BP) GO terms that are in common between the gene pair. GO ID matching is exact. Gene GO term annotations were downloaded through the Ensembl v98 BioMart data mining tool[42] (December 2019).

(k) Linkage disequilibrium (LD): a proxy for possible LD between the two genes in the gene pair was obtained as the maximum LD between the variants flanking the TSS (+/-5kb) of one gene and those flanking the TSS of the other gene in the gene pair. LD was measured as R-squared using vcftools v0.1.15[50]. This metric was only calculated for the Geuvadis LCL dataset.

(l) eQTL sharing: whether the lead cis-eQTL (FDR 5%, Benjamini-Hochberg procedure) of one gene in the pair is nominally significant (p-value < 0.05) for association with the other gene in the gene pair. This is tested two-ways and values are attributed as: 0 if no lead eQTL is shared, 1 if the lead eQTL of one gene is shared, 2 if the lead eQTL of each gene is shared in the other gene of the pair. Sharing is only considered if the effect sign (beta sign) is conserved across the pair. Only lead eQTLs of eGenes are considered. eQTL mapping was performed with QTLtools v1.1[40] on the same samples and individuals studied for local gene co-expression. Further details on the eQTL analysis are provided below.

(m) Co-expressed eGenes: whether none, one or both genes in the pair are eGenes, i.e. having a significant cis-eQTL (FDR 5%, Benjamini-Hochberg procedure), using the same cis-eQTLs as in the "eQTL sharing" metric.

For all metrics, missing data and 'NA's were replaced with zeroes. When necessary, annotation coordinates were moved to hg38 (for GTEx tissues) or hg19 (for Geuvadis dataset) using the UCSC liftOver tool[51]. In addition to these metrics, to estimate distance between gene pairs in terms of centimorgans, genetic distance maps based on recombination estimates from Auton *et al.* 2007 were used[52]. Using these maps, the absolute centimorgan distance between gene TSSs was interpolated for each Geuvadis LCL COP and non-COP.

Logistic regression models

Logistic regression for several models of features/metrics liable to explain some of the observed local gene co-expression were performed with custom scripts using R (v3.4.4) project programming language, using the *glm* function. Co-expression was encoded in a binomial format based on co-expression, with COPs regarded as positives (value 1) and distance-matched non-COPs as negatives (value 0, see above for details). The regression models were trained with 80% randomly sampled gene pairs (no replacement), keeping the same sample size between positives and negatives. The ability to correctly predict whether the remainder 20% gene pairs are positives or negatives from the learned models was evaluated for each model independently by measuring the AUC of the ROC curve. The sampling of gene pairs into training or test sets was repeated 50 times and the mean AUC reported, unless otherwise specified.

For simplicity, several related metrics were combined into the same model, and thus the AUCs reported are based on these combinations of metrics: *(i)* the 'Enhancer' model consists of 'Enhancer sharing' + 'Total enhancers' metrics; *(ii)* the 'TF' model consists of 'Shared transcription factors' + 'Total transcription factor binding sites' metrics; *(iii)* the 'CTCF' model consists of 'Total CTCF sites' and 'Inverted CTCF motifs'; *(iv)* the 'Expression' model consists of 'Expression level difference' + 'Expression coefficient of variation difference' metrics (for Geuvadis LCLs only); *(v)*

the 'eQTL sharing' model consists of 'eQTL sharing' and 'Co-expressed eGenes' metrics. All other models are based on a single metric, described above.

COP sharing across tissues

Clustering of tissues based on COP sharing was performed on R (v3.4.4) using hclust and Pearson correlation coefficient using a discovery-replication approach: for a pair of tissues (tissue A and tissue B), the percentage of COPs (FDR 1%) of tissue A that are also COPs (FDR 1%) in tissue B (and vice versa) is calculated. This provides a non-symmetrical matrix. Only COPs assessed in both tissue A and B were considered. This was performed before COP dataset filtering.

Local gene pair co-presence is defined as both genes in the pair passing an expression threshold of >0.1 TPM in >=20% samples and >=6 reads in >=20% samples (defined by the GTEx project), which is a prerequisite for COP detection. The calculation of the percentage of tissues where co-present was based on this definition across the 49 GTEx tissues. COP sharing and percentage of tissues where co-expressed was calculated as the number of tissues where the COP is found (FDR 1%) divided by the number of tissues where the pair is co-present. To be able to distinguish simple gene expression tissue-specificity from gene co-expression specificity, only COPs where both genes were present (non-zero expression) in at least 5 tissues were considered (38,196 out of 40,999). COP sharing categories (conserved, prevalent, specific and unique) were derived from this metric as follows: *i)* 'unique COPs', found in only one tissue (N = 20,781 across tissues), *ii)* 'specific COPs', found in >1 tissue but at most 15% tissues where both genes in the pair are present (range: 2-7 tissues, N = 10,111), *iii)* 'prevalent COPs', found in >1 tissue and between 15-50% tissues (N = 4,863) and *iv)* 'conserved COPs', found in >50% tissues (N = 2,441).

eQTL mapping and eQTL sharing analysis

Genotypes for the 358 European individuals of the Geuvadis dataset[15] were downloaded from the 1000 Genomes[16] FTP server (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). All variant sites with more than 2 possible alleles or with a MAF below 5% have been removed. In total 6,865,255 variants (6,050,717 SNPs, 809,705 indels, 4,833 structural variants) were used in analysis after filters. Genotypes from the GTEx v8 dataset were obtained via dbGaP accession number phs000424.v8.p1, which includes 838 subjects, 85.3% European American, 12.3% African American and 1.4% Asian American. Details regarding donor enrollment, consent process, biospecimen procurement methods and other procedures is previously described[53]. The phased version of the

genotype files was used and only variants with MAF >5% were used for analysis, totalling 6,590,509 variants (6,123,541 SNPs, 466,968 indels).

For each gene expression matrix (normalised and corrected for covariates), i.e. for Geuvadis LCL expression matrix and each GTEx tissue expression matrix, cis-eQTLs were mapped using QTLtools v1.1[40] *cis* function. The mapping window was defined as 1 Mb up- and down-stream of the TSS of each gene and in all cases the --normal option was used to enforce gene expression phenotypes to match normal distributions N(0, 1). Two runs of eQTL mapping were produced:

1) identification of eGenes and lead eQTLs: the QTLtools cis --permute 1000 option was used to obtain a lead eQTL for each gene, i.e. the variant with the lowest p-value. Of note, the --permute option takes into account the total number of variants tested in each gene by randomly shuffling gene quantifications 1,000 times and producing a null distribution of the best association p-values obtained in each randomisation. This null distribution is used to fit a beta distribution and thus adjust the nominal p-value obtained with the read data. Furthermore, to account for multiple genes being test in each gene expression matrix, the false discovery rate (FDR) was calculated on beta-adjusted nominal p-values through the Benjamini-Hochberg procedure and only genes with a lead eQTL with FDR less or equal 0.05 were considered to be eGenes and their lead eQTLs kept.
2) identification of all gene-cis-eQTL associations: the QTLtools *cis* --nominal 0.05 option was used to obtain all gene-cis-variant associations (within the defined cis-window) with nominal p-value < 0.05. In this case no adjustment or FDR threshold was applied to these nominal p-values, as these were only used to replicate the association of a lead eQTL to neighbouring co-expressed genes (eQTL sharing).

Using both runs of eQTL mapping described above and the sets of COPs (FDR 1%) for each tissue in GTEx and Geuvadis LCL dataset, the lead eQTL (FDR 5%) of each eGene was tested for their ability to associate (i.e. nominal p-value < 0.05) with other co-expressed genes (even if these are not eGenes). Lead eQTLs that associate with co-expressed genes in this way are considered as 'shared lead eQTLs". Importantly, a shared lead eQTL is only considered if the effect sign (beta sign of the regression) is the same for both genes. This definition of eQTL sharing was used for both the molecular feature 'eQTL sharing' and for subsequent functional enrichments of lead eQTLs.

For colocalization analysis, the nominal p-values of all variants 2 Mb up- and down-stream of the TSS of each gene were used (no p-value threshold applied). The COLOC [30] abf function was run for each Geuvadis LCL COP (N = 6668) and distance-matched non-COPs (N = 6668), i.e. considering each gene in the pair as a trait. Variant MAF calculated with bcftools v1.11 on the 358 Geuvadis samples

used for COP identification was used for input. Gene pairs with coloc PP4 (posterior probabilities for hypothesis 4) > 0.5 were considered to share a signal.

Functional annotations and eQTL enrichments tests

Several functional annotation datasets were collected from distinct sources:

1) ENCODE[28] functional annotations for LCL (GM12878) downloaded from the EBI FTP server. *(i) Protein binding*: Transcription Factor Binding Site (TFBS) coordinates from 74 ChIP-seq experiments across distinct transcription factors; *(ii) TSS* (promoter region including TSS); *(iii) enhancer* as well as *CTCF* (CTCF enriched element) annotations produced by ChromHMM[54] combined with SegWay[55]; *(iv) DNAse:* open chromatin regions predicted through DNAse-seq (narrowPeak, FDR 1%).
2) GENCODE[39] gene body categories of human genes: for Geuvadis LCLs, the coding sequence (CDS), untranslated region (UTR) and intron coordinates for all genes were retrieved from the GENCODE v19 GFF3 file (i.e. the annotation on which gene expression was assessed). For GTEx data, the same gene body regions were retrieved from GENCODE v26 and the 3'UTR and 5'UTR were combined into the same UTR category to reproduce the regions available for the Geuvadis dataset. In all cases, annotations were migrated from hg19 to hg38 when needed using the UCSC liftOver tool[51].
3) Chromatin state annotation from the Roadmap epigenomics Core[29] 15-state model (5 marks) was downloaded in July 2020 for Lymphoblastoid Cells (E116), Lung (E096) and Skeletal Muscle (Male (E107) and Female (E108) samples were concatenated).

The statistical enrichment of the overlap of lead eQTLs – shared lead eQTLs and other lead eQTLs – for each functional annotation was performed in two ways: *(i)* Direct enrichments: two-way Fisher's Exact test between the overlap found for shared lead eQTLs versus the overlap found for other lead eQTLs; *(ii)* Enrichments over expectation: the assessment of whether eQTLs fall within a certain functional annotations more often than expected by chance was performed with QTLtools v1.3.1[40] *fenrich* function. Briefly, for each eQTL dataset (shared or other) and for each functional annotation, the expected mean number of eQTLs falling within the functional annotations is computed from 10,000 permutations of the annotations found around +/-1Mb of the TSS of each gene. For instance, annotations around gene A are attributed to gene B, accounting for strand orientation and maintaining distances between annotations and TSS identical. In this way, the distribution of functional annotations around genes is unchanged, only their assignment to genes is shuffled through permutations. Then, the observed overlap is compared to the expected one through a two-way Fisher's Exact test. Of note, possible differences between the distribution of functional annotations around eGenes with shared lead eQTLs and those with other lead eQTLs is taken into account in this way,

since the expected distributions represent what is expected for the eGenes present in each category. The granularity of the data used consists of eGene-eQTL pairs, i.e. if the same variant is a lead eQTL for two genes, this is counted twice.

For the analysis aggregating pleiotropic lead eQTLs across the 49 GTEx tissues (shared eQTLs and other lead eQTLs) unique eQTLs were used, i.e. if the same eGene-eQTL pair is present in several tissues, this is only counted once. If a lead eQTL is shared in at least one tissue (even if not in other tissues) it is considered as "shared". Overlap of shared eQTLs and other lead eQTLs was compared across functional annotations from the Ensembl v101 Regulatory Build[33] and genic regions from GENCODE v26[39] using BEDTools[56] intersect v2.29.2.

GWAS summary statistics and trait pleiotropy analysis

GWAS summary statistics for 35 phenotypes comprising diverse categories such as disease, anthropometric, live-style, cardio-metabolic, blood and neurological traits (Supplementary Table 2) were downloaded from the UK Biobank project[57] through the Pan UK Biobank resource (Pan-UKB team. https://pan.ukbb.broadinstitute.org. 2020.). This resource provides summary statistics of 7,221 phenotypes using a generalized mixed model association testing framework and provides results stratified by ancestry groups. For each phenotype, the same set of 28,987,534 variants is used (which includes 96.8% of all distinct GTEx lead eQTLs across tissues), therefore not requiring variant imputation or harmonisation between phenotypes. Only summary statistics for the European ancestry group (EUR, total of 420531 individuals) were used for each phenotype. Moreover, all phenotypes selected had at least 15.000 cases in the EUR population. Genomic coordinates of genetic variants were lifted over to hg38 using the UCSC liftOver tool. For the pleiotropy analysis, genetic variants were considered to be associated with a phenotype if the (EUR) association p-value was below $5e^{-8}$. Quantile-quantile plots of association p-values (not filtered) were built against a uniform distribution using the *ggGWAS* R package and the genomic inflation factor was calculated as λ (lambda), defined as the median of the resulting chi-squared test statistics divided by the expected median of the chi-squared distribution with one degree of freedom.

Data availability

The code used for analysis is available under: https://github.com/Digans/LoCOP. Data on co-expressed genes and shared eQTLs discovered here are available for consultation and download through the LoCOP database (http://glcoex.unil.ch/) developed here. Geuvadis RNA-seq and genotype data are available under accession numbers: EBI ArrayExpress (accession code E-GEUV-1) for RNA-seq data and 1000 Genomes (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/)

for the genotype data. GTEx RNA-seq and genotype data are available from dbGaP (accession: phs000424.v8.p2).

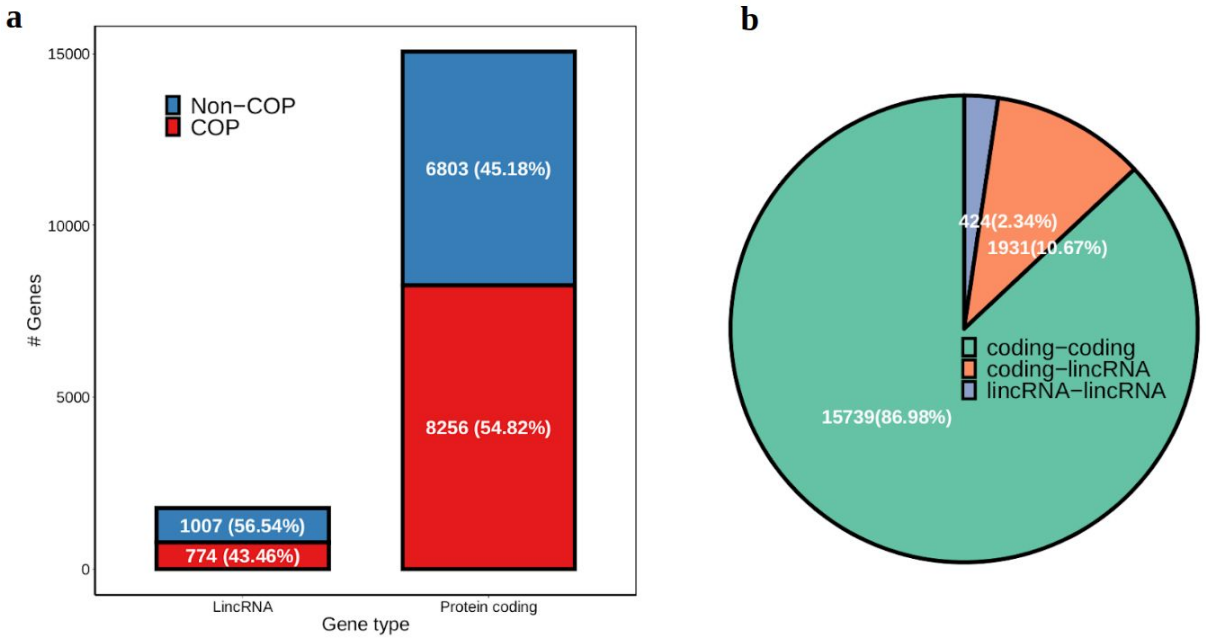
Acknowledgements

O.D. and D.M.R. have been funded by a Swiss National Science Foundation (SNSF) project grant (PP00P3_176977). D.M.R. has also been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 885998. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

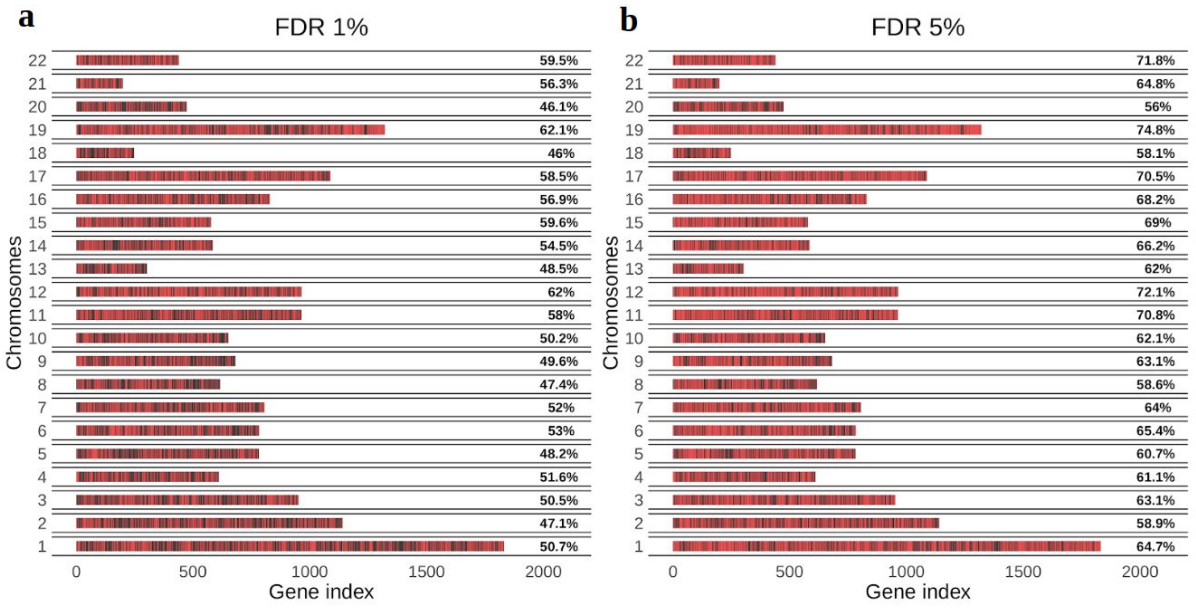
Author Contributions

D.M.R. and O.D. developed computational methods. D.M.R. performed experiments and data analysis. S.R. developed the project website. A.R. helped with ideas/feedback throughout the project. R.J.H. assisted in genetic variation analysis. D.M.R. and O.D. wrote the manuscript with input from other authors. O.D. and E.T.D. conceived and designed the study.

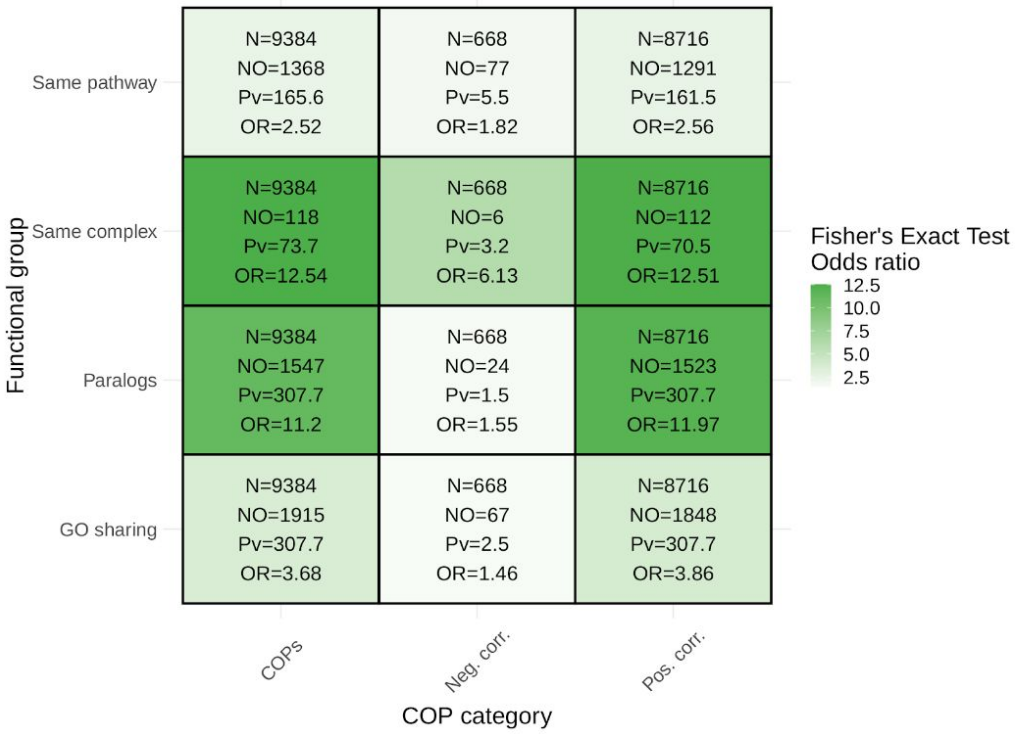
Supplementary Figures



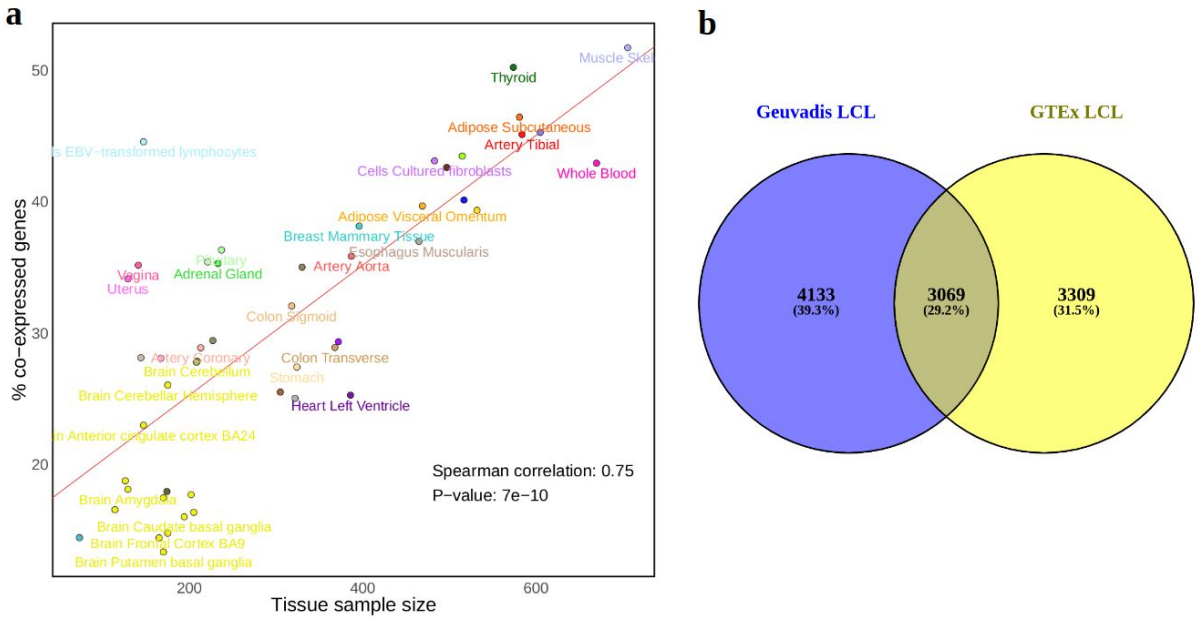
Supplementary Fig. 1 Gene type proportions of COPs. a numbers and proportions of co-expressed genes per gene type; **b** numbers and proportions of COPs per gene type.



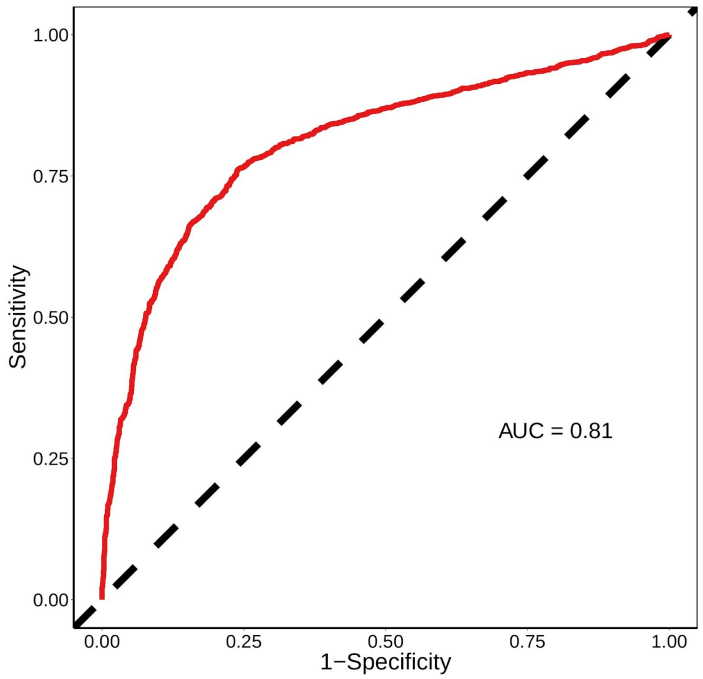
Supplementary Fig. 2 Co-expressed gene percentage and distribution across chromosomes. a COPs detected at 1% FDR; **b** COPs detected at 5% FDR. Red color represents a co-expressed gene, black a non-co-expressed gene. Values on the right side of each plot denote the percentage of genes that are co-expressed per chromosome.



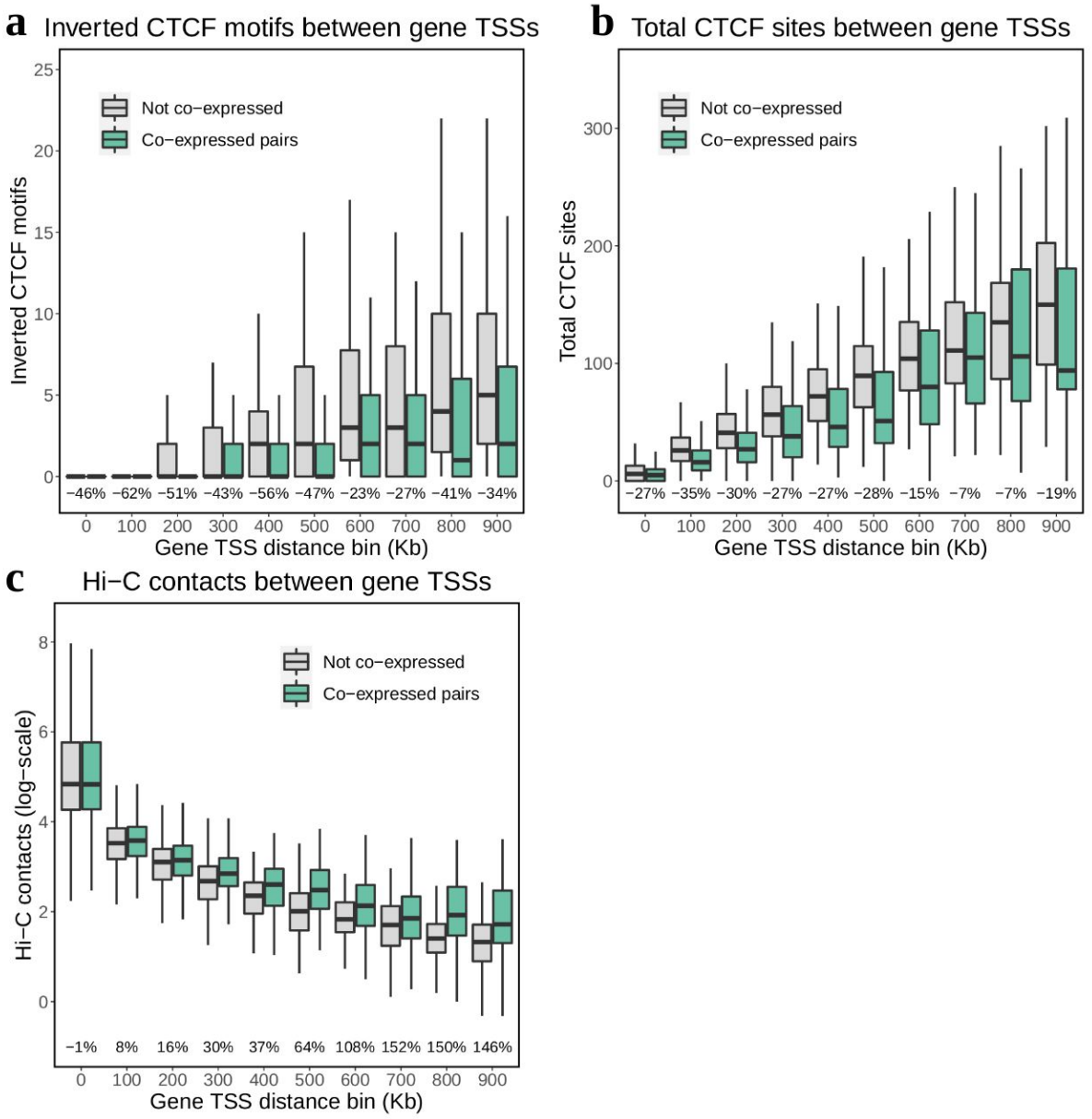
Supplementary Fig. 3 Enrichment of COPs in functionally and evolutionarily-related datasets. 'COPs' represent all COPs together, 'Neg. corr.' represents negatively correlated COPs only whereas 'Pos. corr.' represents positively correlated COPs. 'Same pathway' refers to enrichments for KEGG and Reactome pathways, 'Same complex' refers to enrichments for CORUM and Hu.MAP protein complexes, 'GO sharing' to enrichment of sharing of the same Biological Processes GO term. 'Paralogs' refers to enrichment in paralog genes (Methods). N, total number of COPs tested; NO, number of COPs overlapping with the functional group; Pv, Fisher's Exact test p-value (-log10 scale); OR, Fisher's Exact test odds ratio.



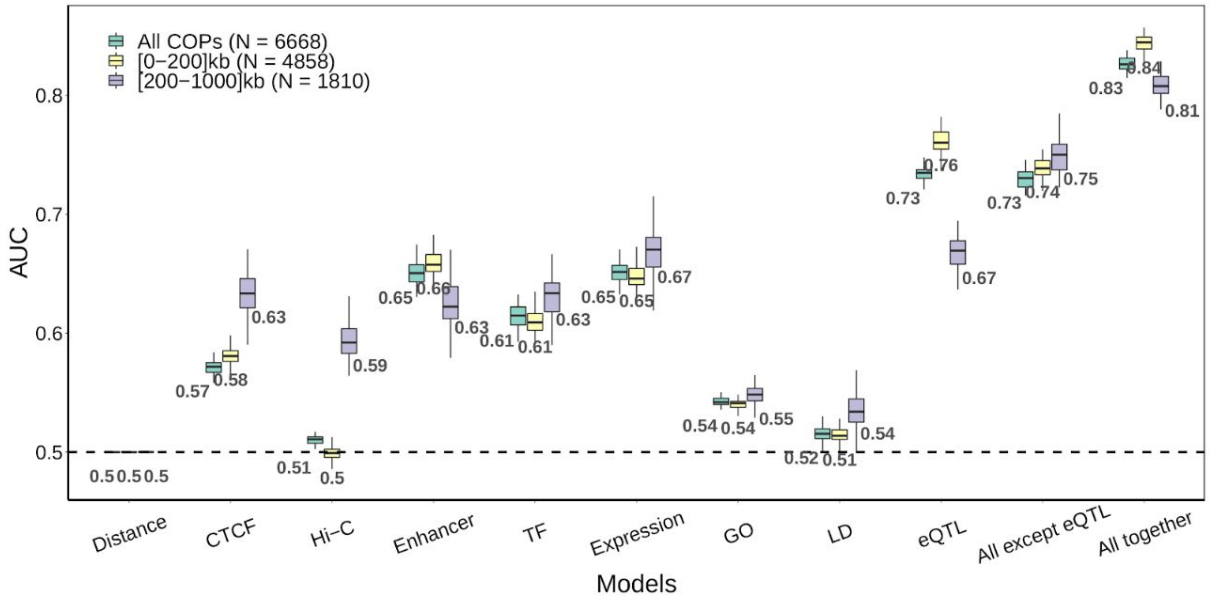
Supplementary Fig. 4 COP identification and tissue sample size. a percentage of co-expressed genes per tissue sample size. **b** Venn diagram of the overlap between Geuvadis and GTEx LCL COPs. Only COPs assessed in both datasets were evaluated. The Venn diagram was produced with Venny v2.1; Oliveros, J.C. (2007-2015) Venny. An interactive tool for comparing lists with Venn's diagrams. https://bioinfogp.cnb.csic.es/tools/venny/index.html.



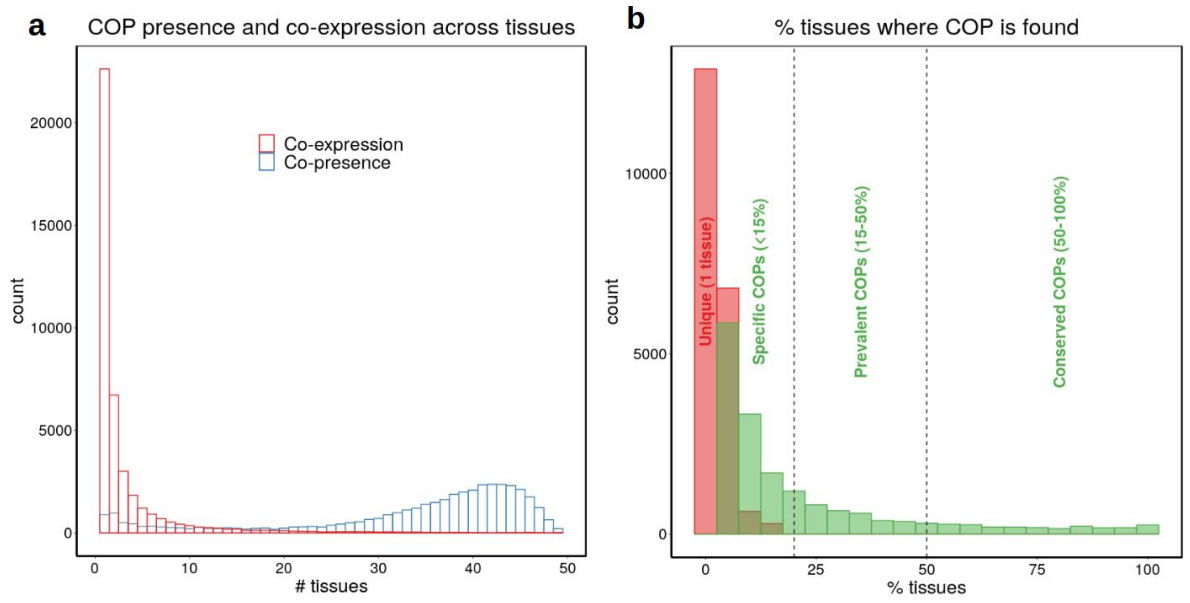
Supplementary Fig. 5 Gene pair TSS absolute distance ROC curve for Geuvadis LCL COPs. COP dataset before applying paralog and positive correlation filters, 9384 COPs (positives) and 9384 randomly sampled non-COPs (negatives).



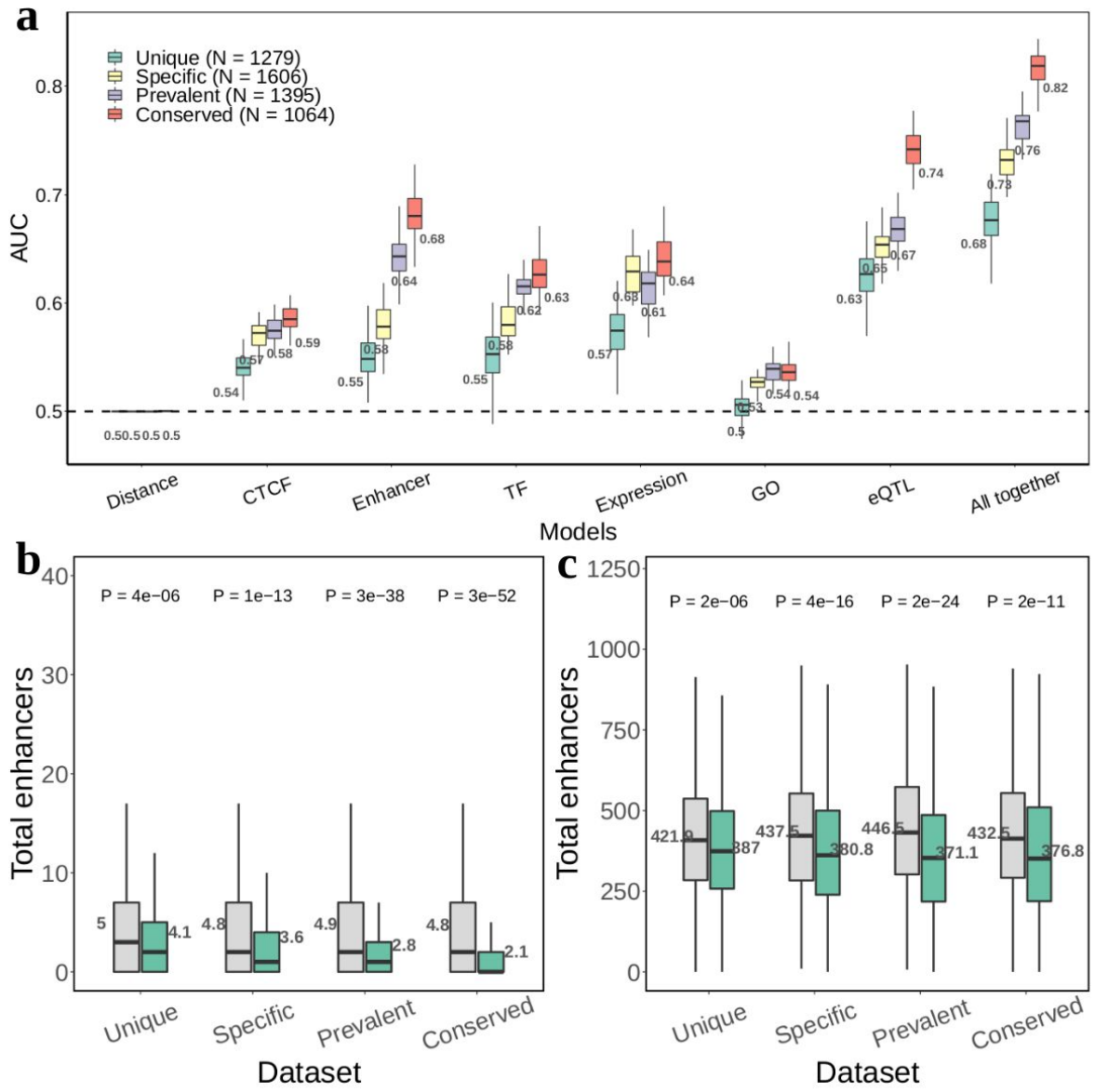
Supplementary Fig. 6 Chromatin structure features (CTCF and Hi-C) on COPs and non-COPs across distance bins. a inverted CTCF motifs (number of pairs of plus- and negatively-stranded motifs) between the two genes in the pair from MotifMap (predictions based on motif matching)**; b** total CTCF sites between the two genes in the pair from ReMap, regardless of strand (LCL-specific experimental data; Methods); **c** Hi-C contact intensities between TSS regions of the two genes in the pair (log-scale; 5kb resolution) from Rao *et al.* 2014. Percentages at the bottom refer to the difference in means between COPs and non-COPs. In all cases, 6668 COPs and 6668 distance matched non-COPs were used.



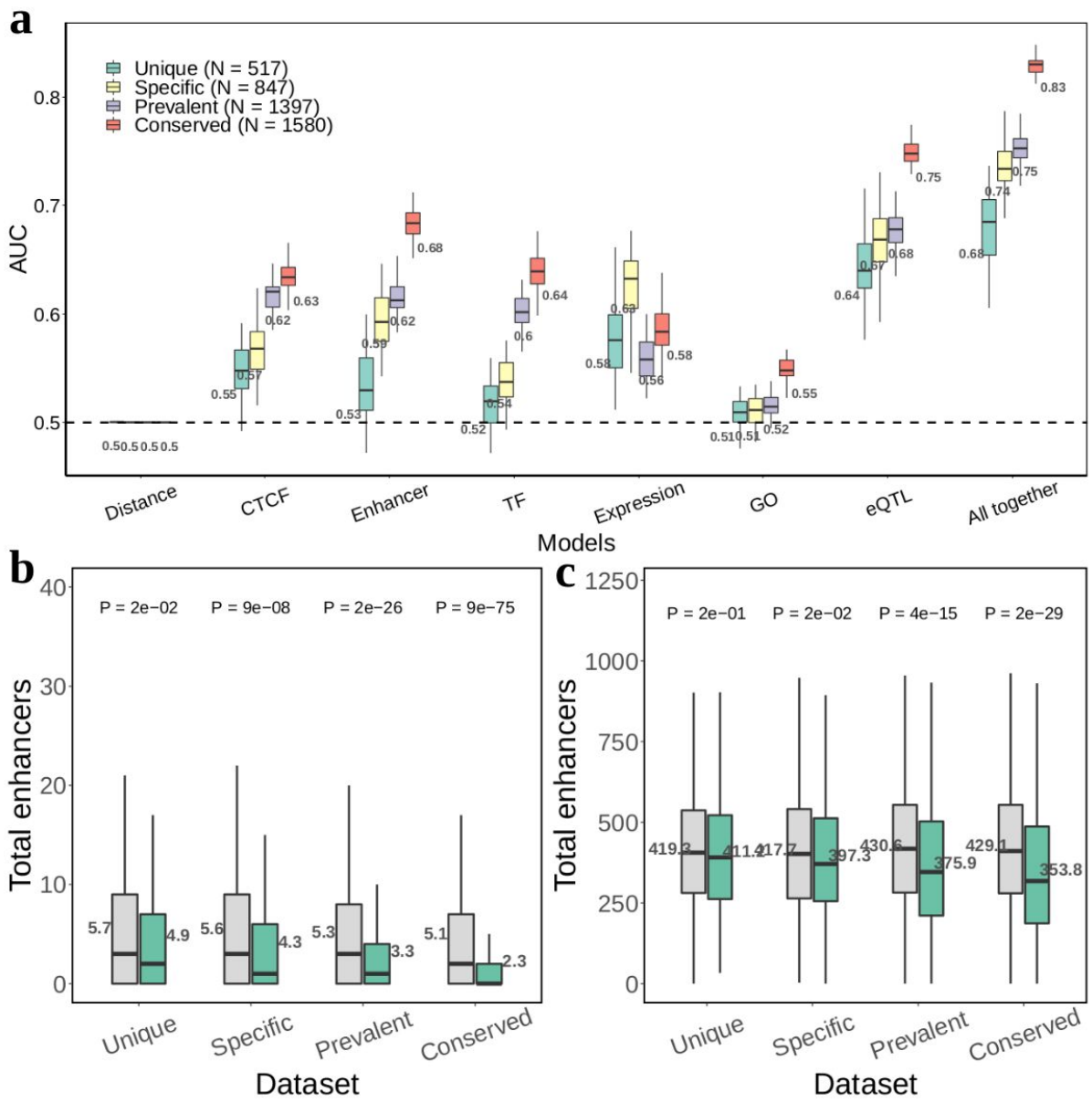
Supplementary Fig. 7 Boxplots of the AUC values obtained for each molecular feature separated by two distance bins for Geuvadis LCLs. Values below the boxplot represent the mean over the 50 randomisations.



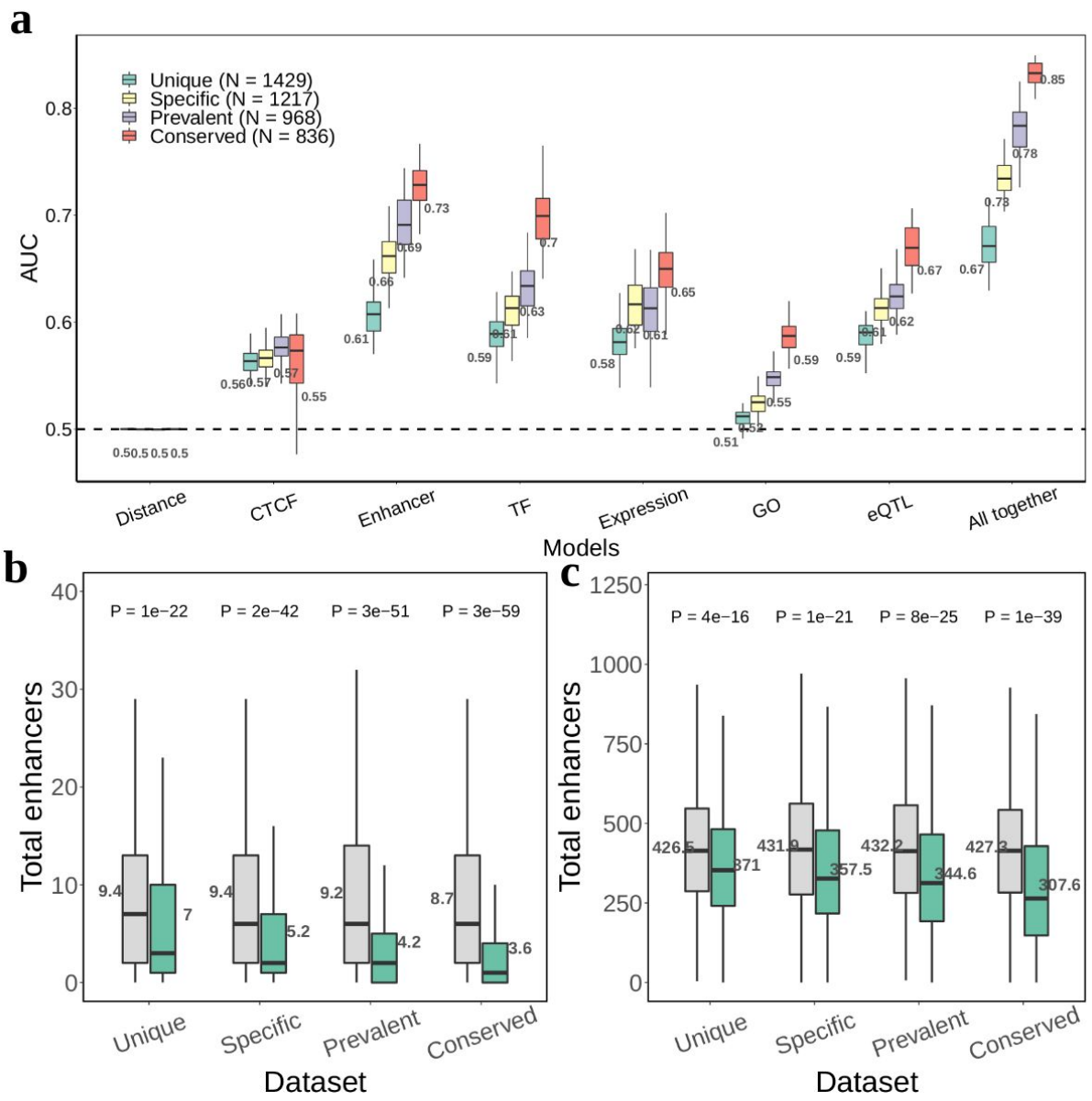
Supplementary Fig. 8 Molecular feature results separating COPs into four categories based on their tissue prevalence across 49 tissues. **a** in red, COP tissue frequency of all 40999 COPs is shown. In blue, the presence of these COPs across tissues is shown (i.e. based on gene expression of both genes, rather than significant co-expression); **b** distribution of the percentage of tissues where COPs are found (against the number of tissues where COPs are present), only COPs where both genes are present in at least 5 tissues were considered. This allowed the separation of COPs into the following categories: *i)* 'unique COPs', found in only one tissue (N = 20,781 across tissues), *ii)* 'specific COPs', found in more than 1 tissue but at most 15% tissues where both genes in the pair are present (range: 2-7 tissues, N = 10,111), *iii)* 'prevalent COPs', found in more than 1 tissue and between 15-50% tissues (N = 4,863) and *iv)* 'conserved COPs', found in more than 50% tissues (N = 2,441).



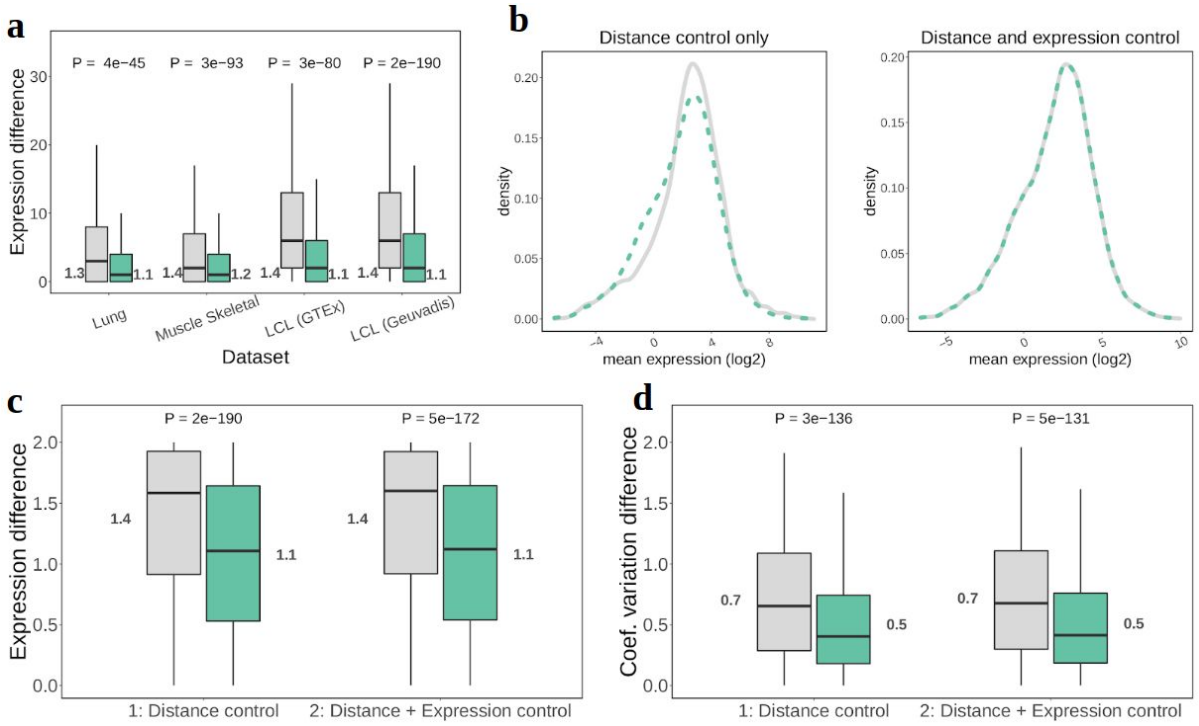
Supplementary Fig. 9 Molecular feature results separating COPs into four categories based on their tissue prevalence across 49 tissues for Muscle Skeletal. **a** boxplots of the AUC values obtained for each molecular feature on Muscle Skeletal COPs, separated by tissue prevalence categories. Values below the boxplot represent the mean over the 50 randomisations. Sample size of each category (number of positive and distance-matched negative) is found on the left corner of the plot; **b,c** boxplots of two molecular features of Muscle Skeletal COPs (green) and non-COPs (grey): total enhancers and total TFBS. Values next to the boxplots represent the mean. P-values were obtained from two-tailed Wilcoxon signed-rank tests.



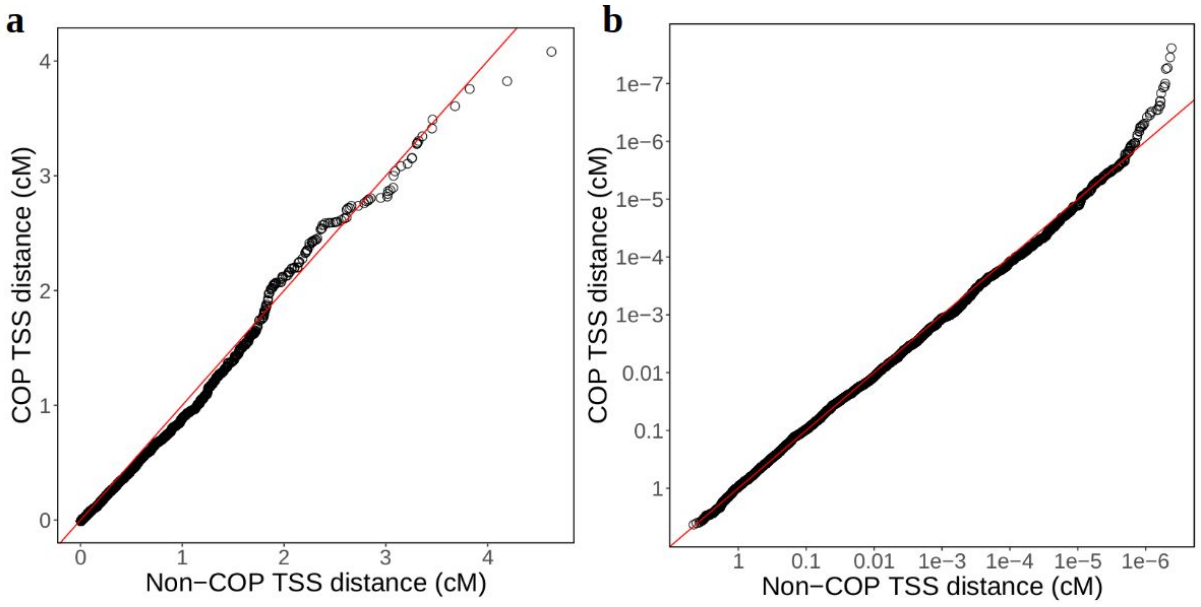
Supplementary Fig. 10 Molecular feature results separating COPs into four categories based on their tissue prevalence across 49 tissues for Lung. **a** boxplots of the AUC values obtained for each molecular feature on Lung COPs, separated by tissue prevalence categories. Values below the boxplot represent the mean over the 50 randomisations. Sample size of each category (number of positive and distance-matched negative) is found on the left corner of the plot; **b,c** boxplots of two molecular features of Lung COPs (green) and non-COPs (grey): total enhancers and total TFBS. Values next to the boxplots represent the mean. P-values were obtained from two-tailed Wilcoxon signed-rank tests.



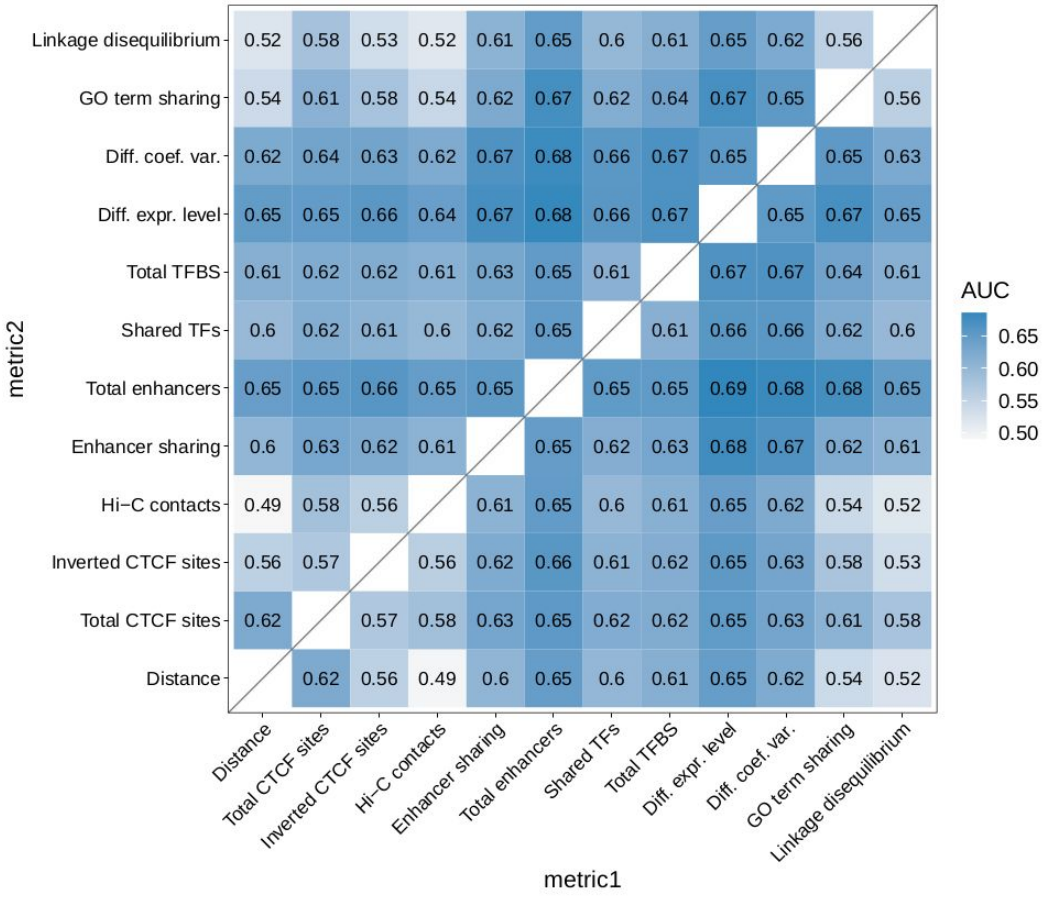
Supplementary Fig. 11 Molecular feature results separating COPs into four categories based on their tissue prevalence across 49 tissues for GTEx LCLs. **a** boxplots of the AUC values obtained for each molecular feature on LCL COPs, separated by tissue prevalence categories. Values below the boxplot represent the mean over the 50 randomisations. Sample size of each category (number of positive and distance-matched negative) is found on the left corner of the plot; **b,c** boxplots of two molecular features of GTEx LCL COPs (green) and non-COPs (grey): total enhancers and total TFBS. Values next to the boxplots represent the mean. P-values were obtained from two-tailed Wilcoxon signed-rank tests.



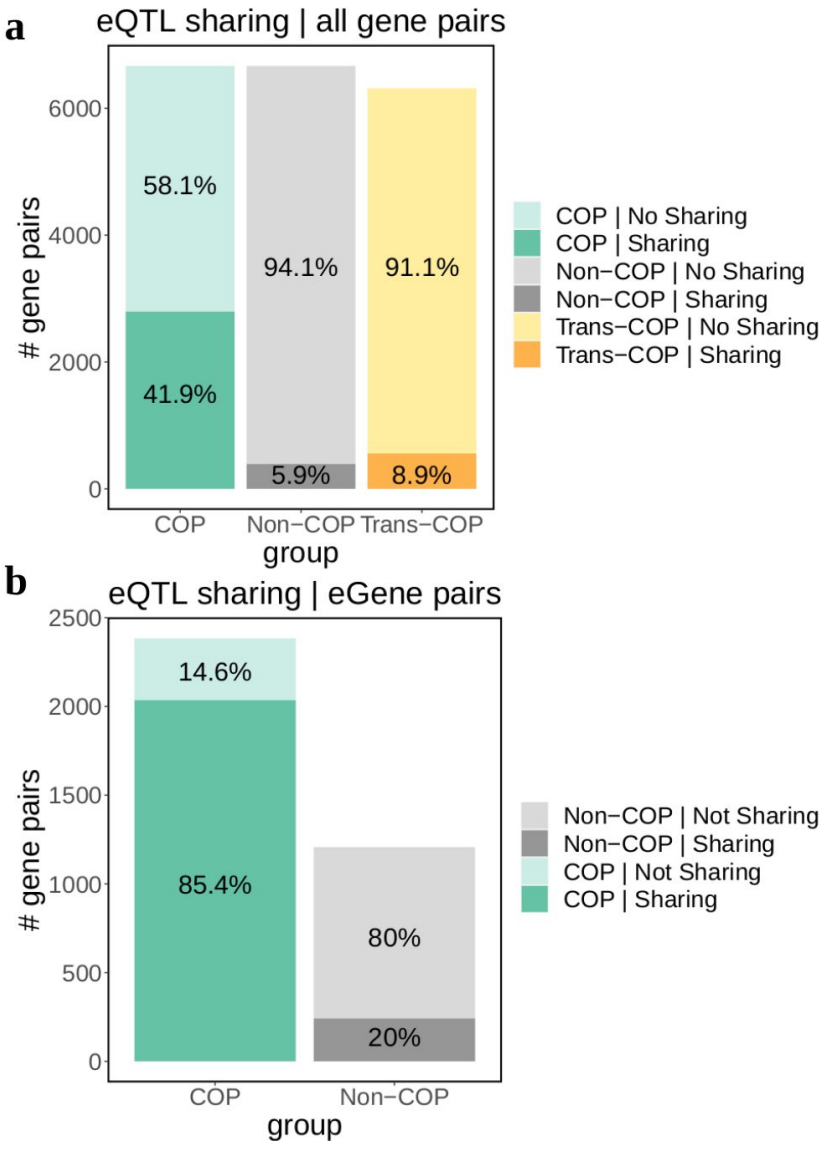
Supplementary Fig. 12 Expression level difference and coefficient of variation difference between COps and non-COPs. **a** boxplots of the expression level difference between the two genes in the pair, for COPs (green) and non-COPs (grey) across 4 datasets. Expression level difference was calculated as the absolute difference between the average expression level of the genes in the pair, divided by the average expression level of the pair. In all cases, values next to the boxplots represent means; **b** details of the mean expression level control (between the two genes in the pair) used for this specific analysis. After controlling for both distance (at most 5% difference in distance allowed between COP and matching non-COP) and mean expression level (at most 10% difference allowed), the mean expression distribution between COPs and non-COPs clearly matches (right plot). In the step of picking non-COPs matched for both distance and expression level, 545 COPs were lost; **c** boxplots of the expression level difference between the two genes in the pair for Geuvadis LCLs COPs and non-COPs, before and after controlling for mean expression level; **d** boxplots of the coefficient of variation difference between the two genes in the pair for Geuvadis LCLs COPs and non-COPs, before and after controlling for mean expression level (Methods). The difference between COP and non-COP is highly significant in all cases Wilcoxon signed-rank tests p-values $<2.2e^{-16}$. Mean RPKMs were used for Geuvadis LCLs and median TPM were used for GTEx tissues.



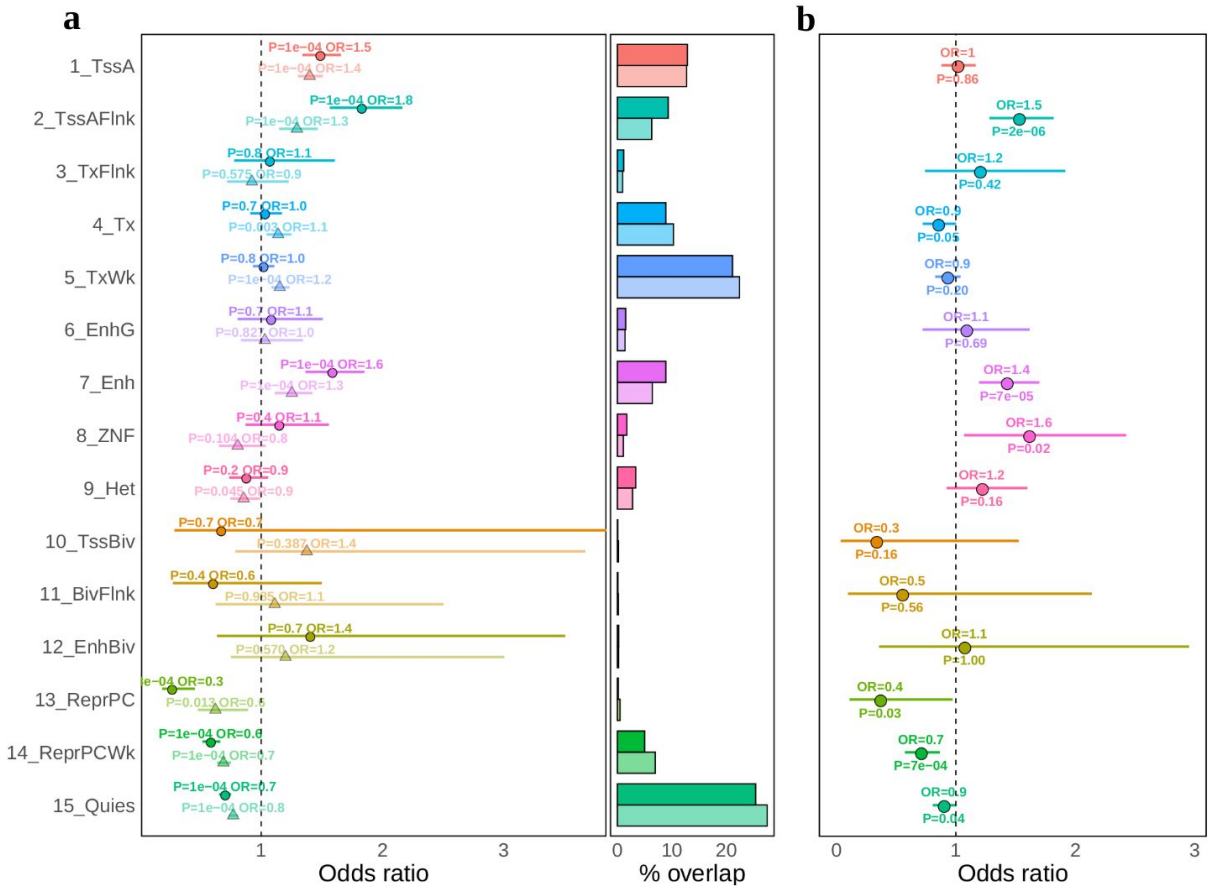
Supplementary Fig. 13 Comparison of centimorgan (cM) distance between TSSs of Geuvadis LCL COPs and non-COPs. a absolute distance in normal scale; **b** absolute distance in -log10 scale. In both cases 6668 COPs and 6668 distance matched non-COPs were used and linear regression slope = 1.02.



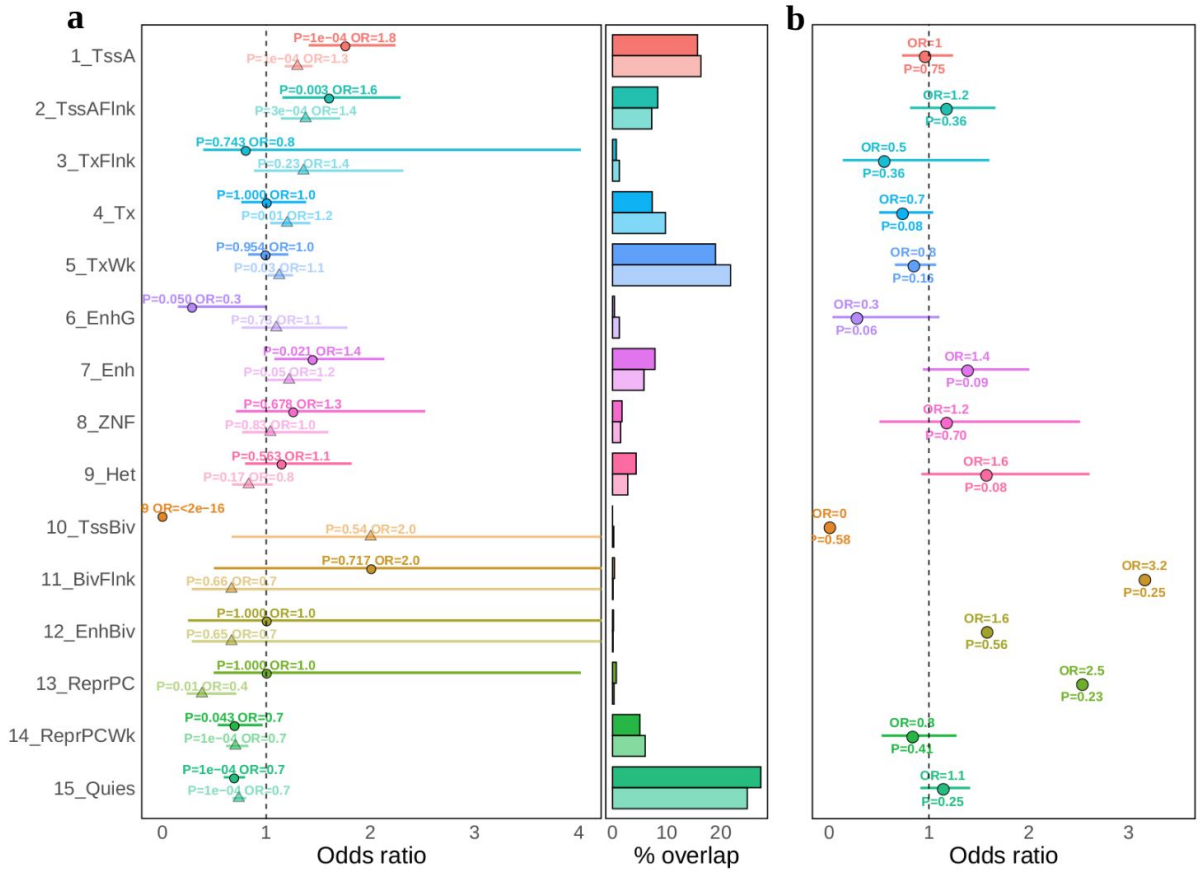
Supplementary Fig. 14 Mean AUC of pairwise combinations of molecular feature metrics in Geuvadis LCLs. AUCs are averages of 50 training-test set randomisations. The upper and lower triangles come from two separate sets of randomisations.



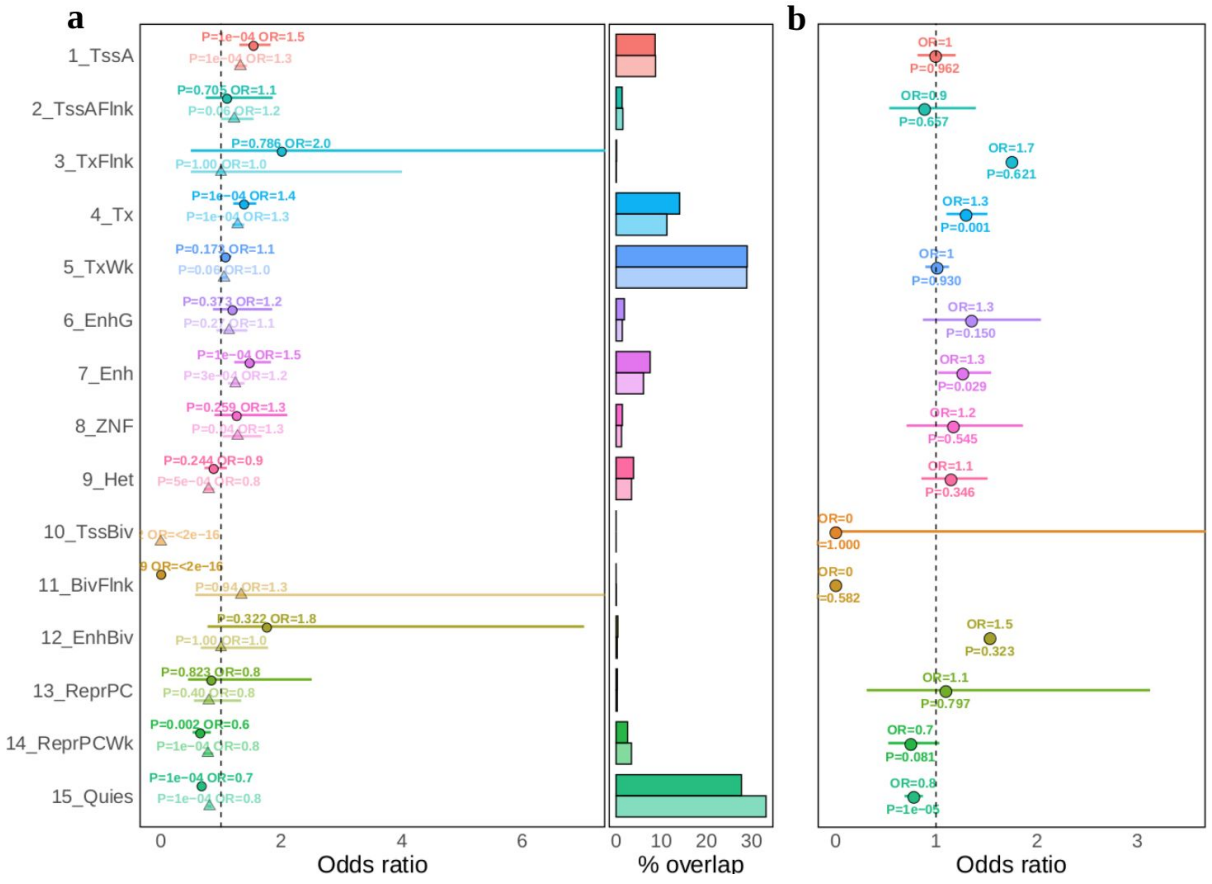
Supplementary Fig. 15 eQTL sharing in COP, non-COPs and trans-COPs. a number and percentages of Geuvadis LCLs COPs (N = 6668), distance-matched non-COPs (N = 6668) and correlation-matched trans-COPs (N = 6316) in eQTL sharing. Only 6316 trans-COPs could be matched to cis-COPs by correlation (maximum difference of 5% correlation value between cis-COPs and trans-COPs; Methods); **b** eQTL sharing in Geuvadis LCL COPs and non-COPs when only considering cases where both genes are eGenes.



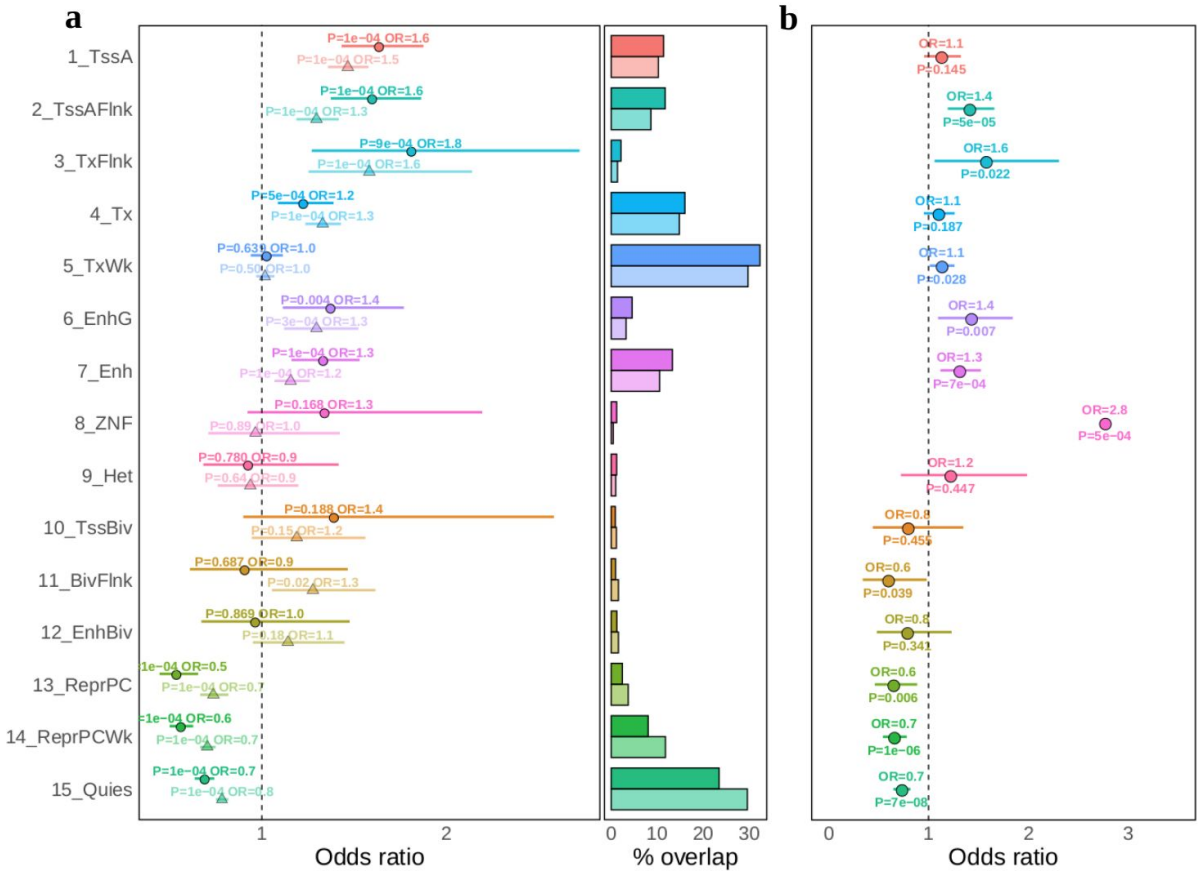
Supplementary Fig. 16 Functional enrichments of Geuvadis LCL shared eQTLs on Roadmap epigenomics annotations. a overlap enrichment of shared lead eQTLs (solid color, round points) and other lead eQTLs (pale color, triangles) in 15-state Roadmap epigenomics functional annotations for Geuvadis LCL (Methods). Odds ratios are calculated based on the observed versus expected overlap between eQTLs and each functional annotation. Error bars are from 10000 fenrich permutations. The right part of the plot denotes the percentage of overlap between eQTLs and each functional annotation; **b** Fisher's exact test odds ratio and p-value for the enrichment of shared lead eQTLs in each functional annotation, compared to other lead eQTLs, for Geuvadis LCL. Error bars of odds ratio are 95% confidence intervals.



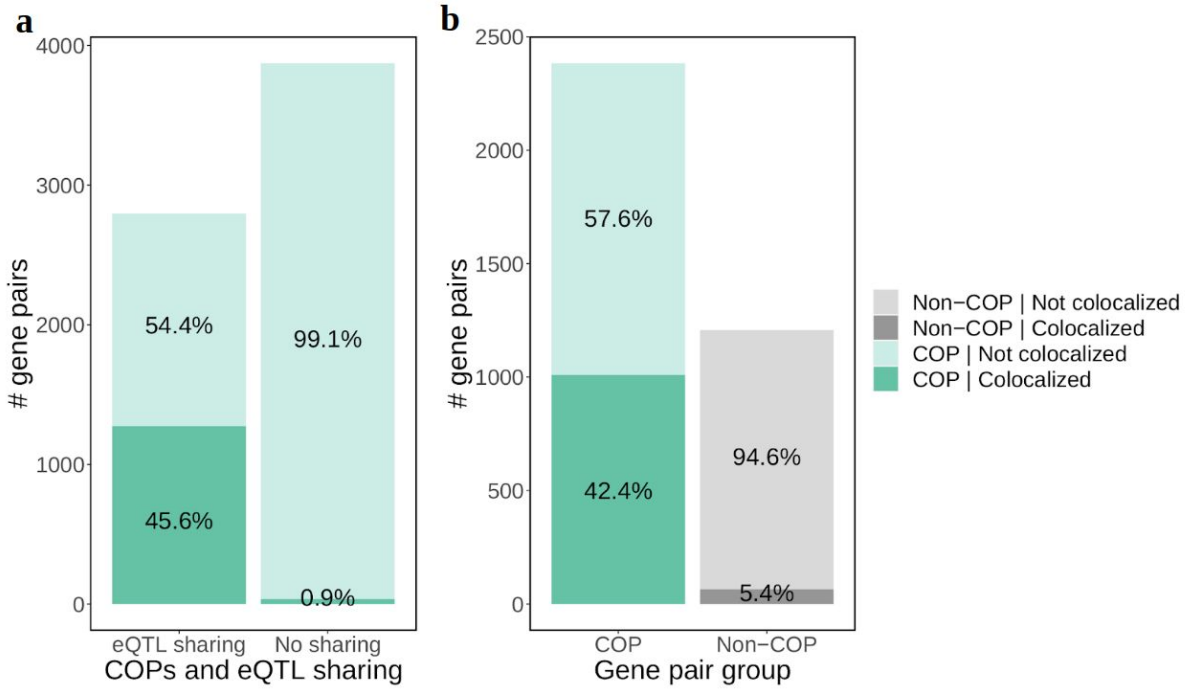
Supplementary Fig. 17 Functional enrichments of GTEx LCL shared eQTLs on Roadmap epigenomics annotations. a overlap enrichment of shared lead eQTLs (solid color, round points) and other lead eQTLs (pale color, triangles) in 15-state Roadmap epigenomics functional annotations for GTEx LCL (Methods). Odds ratios are calculated based on the observed versus expected overlap between eQTLs and each functional annotation. Error bars are from 10000 fenrich permutations. The right part of the plot denotes the percentage of overlap between eQTLs and each functional annotation; **b** Fisher's exact test odds ratio and p-value for the enrichment of shared lead eQTLs in each functional annotation, compared to other lead eQTLs, for GTEx LCL. Error bars of odds ratio are 95% confidence intervals.



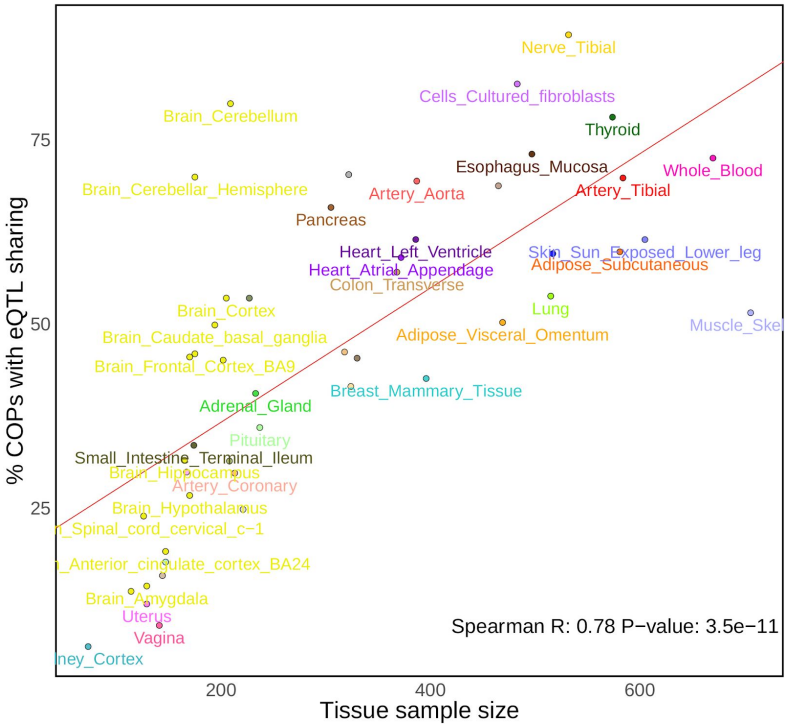
Supplementary Fig. 18 Functional enrichments of Lung shared eQTLs on Roadmap epigenomics annotations. a overlap enrichment of shared lead eQTLs (solid color, round points) and other lead eQTLs (pale color, triangles) in 15-state Roadmap epigenomics functional annotations for Lung (Methods). Odds ratios are calculated based on the observed versus expected overlap between eQTLs and each functional annotation. Error bars are from 10000 fenrich permutations. The right part of the plot denotes the percentage of overlap between eQTLs and each functional annotation; **b** Fisher's exact test odds ratio and p-value for the enrichment of shared lead eQTLs in each functional annotation, compared to other lead eQTLs, for Lung. Error bars of odds ratio are 95% confidence intervals.



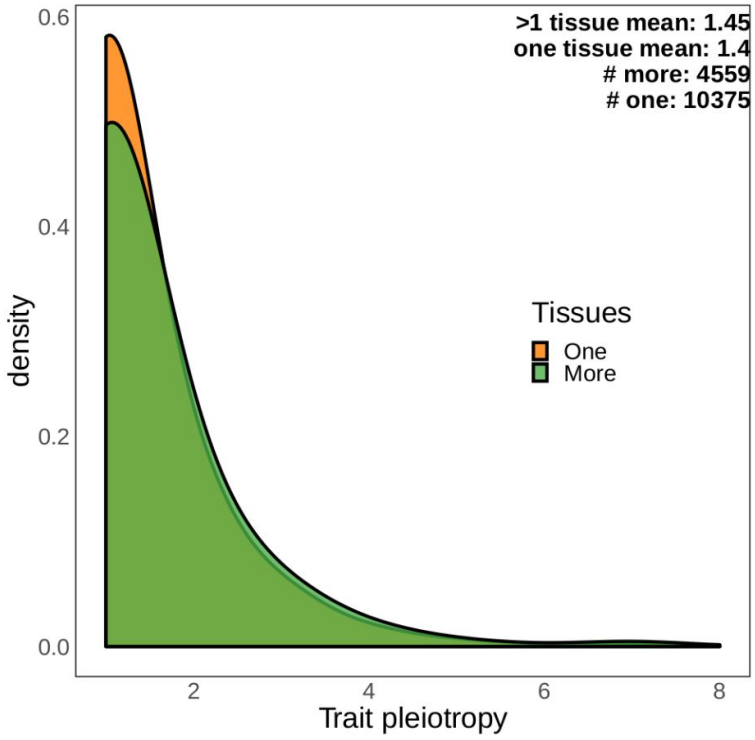
Supplementary Fig. 19 Functional enrichments of Muscle Skeletal shared eQTLs on Roadmap epigenomics annotations. a overlap enrichment of shared lead eQTLs (solid color, round points) and other lead eQTLs (pale color, triangles) in 15-state Roadmap epigenomics functional annotations for Muscle Skeletal (Methods). Odds ratios are calculated based on the observed versus expected overlap between eQTLs and each functional annotation. Error bars are from 10000 fenrich permutations. The right part of the plot denotes the percentage of overlap between eQTLs and each functional annotation; **b** Fisher's exact test odds ratio and p-value for the enrichment of shared lead eQTLs in each functional annotation, compared to other lead eQTLs, for Muscle Skeletal. Error bars of odds ratio are 95% confidence intervals.



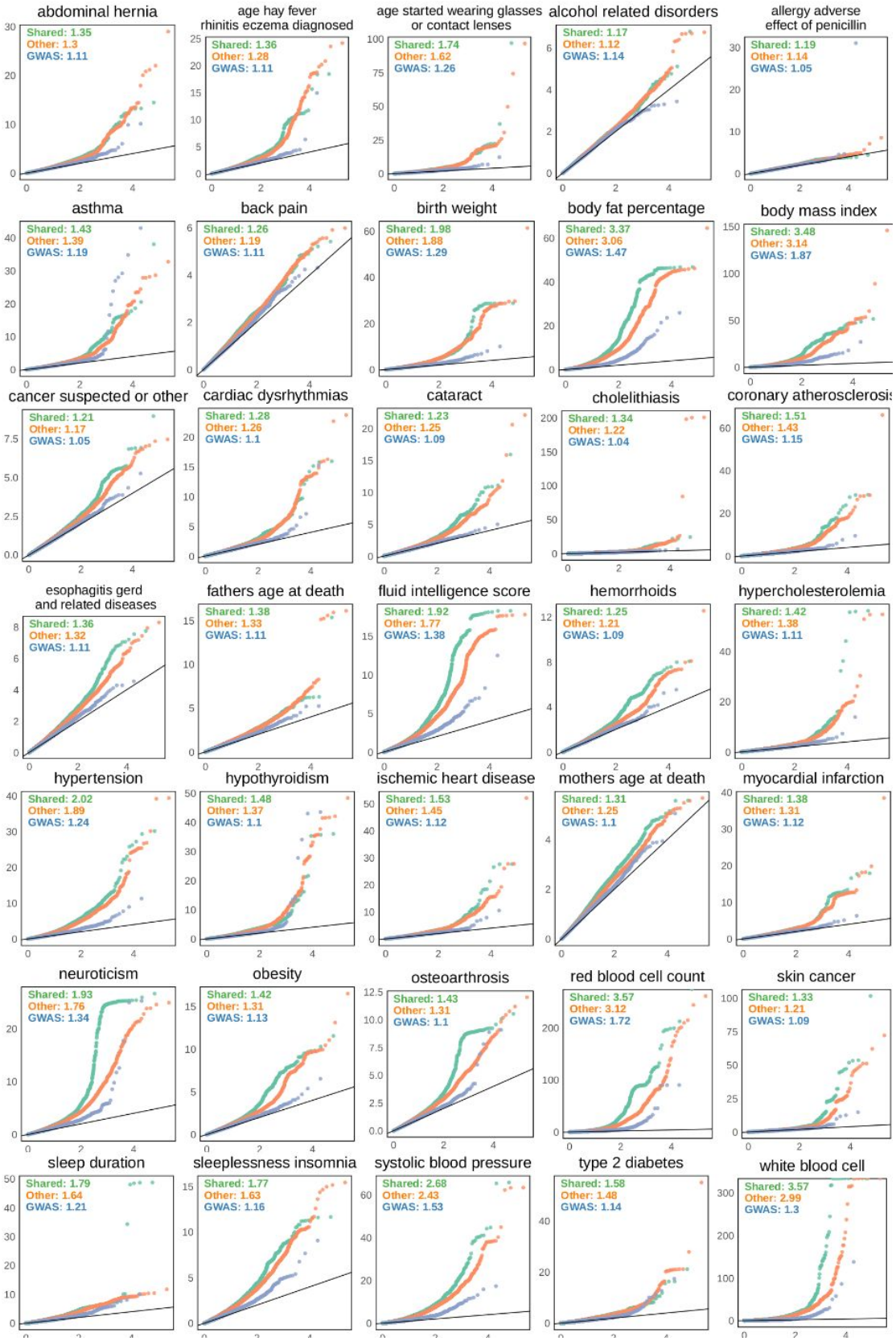
Supplementary Fig. 20 Comparison of gene pair colocalization between eQTL sharing status and COPs and non-COPs. a number of colocalized COPs (COLOC PP4 > 0.5) split by eQTL sharing status. Out of 6668 Gevaudis LCL COPs, 2796 are in eQTL sharing and 3871 are not; **b** numbers of colocalized COPs and non-COPs, only including gene pairs where both genes are eGenes, which is more likely for COPs. N = 2383 for COPs, N = 1207 for distance-matched non-COPs.



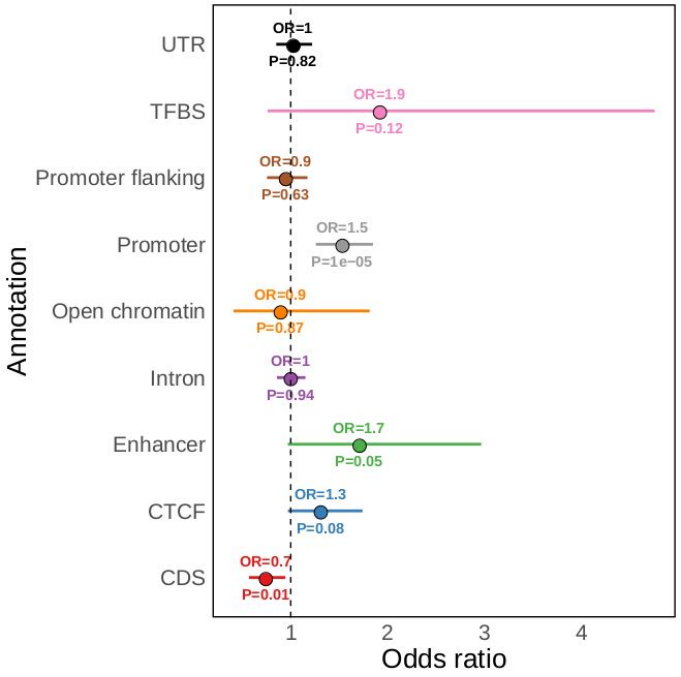
Supplementary Fig. 21 Percentage of COPs with eQTL sharing per tissue sample size.



Supplementary Fig. 22 Comparison of trait pleiotropy between lead eQTLs present in one tissue or more than one tissue.



Supplementary Fig. 23 Quantile-quantile (Q-Q) plots for shared lead eQTLs and other lead eQTLs from all GTEx tissues across 35 traits. Shared (green) and other lead eQTLs (orange) were gathered from all 49 GTEx tissues. Values on the plot denote the genomic inflation factor (Methods). Note that the inflation is higher for shared lead eQTLs than other eQTLs across all but one (cataract) of the 35 traits. GWAS (blue) is a sample of 10000 variants (randomly and independently picked for each trait) shown for comparison purposes.



Supplementary Fig 24 Functional enrichment of shared and other pleiotropic variants. Fisher's exact test odds ratio and p-value for the enrichment of pleiotropic shared lead eQTLs (N = 1274) in each functional annotation, compared to other pleiotropic lead eQTLs (N = 2647). Pleiotropic variants are defined as being associated ($P < 5e^{-8}$) with more than one of the 35 GWAS traits assessed, variants were gathered across 49 GTEx tissues. Error bars are 95% confidence intervals.

Bibliography

1. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).

2. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

3. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).

4. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).

5. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).

6. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in

regulatory DNA. *Science* **337**, 1190–1195 (2012).

7. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
8. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5**, 299–310 (2004).
9. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, (2019).
10. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V. & Reyes, J. C. Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLoS Comput. Biol.* **13**, e1005708 (2017).
11. Ghanbarian, A. T. & Hurst, L. D. Neighboring Genes Show Correlated Evolution in Gene Expression. *Mol. Biol. Evol.* **32**, 1748–1766 (2015).
12. Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat. Cell Biol.* **10**, 1106–1113 (2008).
13. Parsana, P. *et al.* Addressing confounding artifacts in reconstruction of gene co-expression networks. *Genome Biol.* **20**, 94 (2019).
14. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
15. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
16. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* vol. 526 68–74 (2015).
17. Holland, P. W. H., Booth, H. A. F. & Bruford, E. A. Classification and nomenclature of all human homeobox genes. *BMC Biol.* **5**, 47 (2007).
18. Chen, W. V. & Maniatis, T. Clustered protocadherins. *Development* **140**, 3297–3302 (2013).

19. Braastad, C. D., Hovhannisyan, H., van Wijnen, A. J., Stein, J. L. & Stein, G. S. Functional characterization of a human histone gene cluster duplication. *Gene* **342**, 35–40 (2004).

20. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

21. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).

22. Drew, K. *et al.* Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.* **13**, 932 (2017).

23. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

24. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).

25. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

26. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).

27. Miguel-Escalada, I. *et al.* Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* **51**, 1137–1148 (2019).

28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

29. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

30. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

31. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).

32. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

33. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol.* **16**, 56 (2015).

34. Grubert, F. *et al.* Landscape of cohesin-mediated chromatin loops in the human genome. *Nature* **583**, 737–743 (2020).

35. Santiago-Algarra, D., Dao, L. T. M., Pradel, L., España, A. & Spicuglia, S. Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res.* **6**, 939 (2017).

36. Batada, N. N. & Hurst, L. D. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat. Genet.* **39**, 945–949 (2007).

37. Kustatscher, G., Grabowski, P. & Rappsilber, J. Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* **13**, 937 (2017).

38. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58–D64 (2020).

39. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

40. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).

41. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

42. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).

43. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

44. Chen, F., Mackey, A. J., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363–8 (2006).

45. Ouedraogo, M. *et al.* The duplicated genes database: identification and functional annotation of co-localised duplicated genes across genomes. *PLoS One* **7**, e50653 (2012).

46. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).

47. Daily, K., Patel, V. R., Rigor, P., Xie, X. & Baldi, P. MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics* **12**, 495 (2011).

48. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

49. Gao, T. *et al.* EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* **32**, 3543–3551 (2016).

50. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

51. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).

52. Auton, A. & McVean, G. Recombination rate estimation in the presence of hotspots. *Genome Res.* **17**, 1219–1227 (2007).

53. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* **13**, 311–319 (2015).

54. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

55. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).

56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

57. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).