A general framework for predicting the transcriptomic consequences of non-coding variation

Moustafa Abdalla[1,2,3], Mohamed Abdalla[4], Mark I. McCarthy[1,2,5]*, Chris C. Holmes[1,3,6]*

[1]Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, United Kingdom; [2]Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, United Kingdom; [3]Computational Statistics and Machine Learning, Department of Statistics, University of Oxford, United Kingdom; [4]Department of Computer Science, University of Toronto, Canada; [5]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, United Kingdom; [6]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, United Kingdom

*Correspondence should be addressed to M.I.M. (mark.mccarthy@drl.ox.ac.uk) and C.C.H. (cholmes@stats.ox.ac.uk).

16 **ABSTRACT**

17

18 Genome wide association studies (GWASs) for complex traits have implicated thousands of genetic
19 loci. Most GWAS-nominated variants lie in noncoding regions, complicating the systematic translation
20 of these findings into functional understanding. Here, we leverage convolutional neural networks to
21 assist in this challenge. Our computational framework, peaBrain, models the transcriptional machinery
22 of a tissue as a two-stage process: first, predicting the mean tissue specific abundance of all genes and
23 second, incorporating the transcriptomic consequences of genotype variation to predict individual
24 abundance on a subject-by-subject basis. We demonstrate that peaBrain accounts for the majority
25 (>50%) of variance observed in mean transcript abundance across most tissues and outperforms
26 regularized linear models in predicting the consequences of individual genotype variation. We highlight
27 the validity of the peaBrain model by calculating non-coding impact scores that correlate with
28 nucleotide evolutionary constraint that are also predictive of disease-associated variation and allele-
29 specific transcription factor binding. We further show how these tissue-specific peaBrain scores can be
30 leveraged to pinpoint functional tissues underlying complex traits, outperforming methods that depend
31 on colocalization of eQTL and GWAS signals. We subsequently derive continuous dense embeddings
32 of genes for downstream applications, and identify putatively functional eQTLs that are missed by high-
33 throughput experimental approaches.

Most reported disease-associated variation for complex traits lies in non-coding regions of the genome[1]. Despite advances in discovery and annotations of functional non-coding elements across the genome[2-5], characterising the consequences of non-coding variants remains a major challenge in human genetics. Prediction of the transcriptomic consequences of non-coding variation represents one solution[6-10]. Current methods of variant-expression prediction can be broadly divided into two classes: (a) methods that predict alterations in epigenetic and transcription factor binding sites (TFBS), such as DeepSEA[8] and Basset[10]; and (b) methods that directly predict RNA abundance from genotype or sequence data, such as PrediXcan[6] and TWAS[9]. Methods in the former category do not capture differences in transcript expression as a result of genotypic variation[8,10] and are relatively poor predictors of alterations in the histone code[8]; methods in the latter category are not able to identify which of the variants detected within an eQTL association locus are functional[6,9].

To address these concerns, here, we introduce a single framework, called **p**romoter-and-**e**nhancer-derived **a**bundance (**pea**Brain) model, which consolidates both of these approaches. Within the peaBrain framework, the transcriptional machinery of a tissue is modelled computationally as a two-stage process. Stage 1 is a single model in which peaBrain predicts the mean abundance of each gene in a given tissue from DNA sequences, optionally annotated with epigenetic and genomic annotations. Stage 2 incorporates the transcriptomic consequences of genotype variation to predict individual abundance of any given gene; that is, it generates a gene- and tissue-specific model sensitive to individual variation.

We demonstrate that the convolutional neural networks (CNNs) underlying this framework can capture the majority of variance (>50%) in the mean abundance of genes across most GTEx tissues (Stage 1), with utility in a diverse set of tasks (such as identifying somatic mutations with high-impact consequences or pinpointing the functional tissues underlying GWAS signal from complex traits). We further show that CNNs outperform linear models in predicting the consequences of genotype variation (Stage 2). In EBV-transformed lymphocytes (LCLs), we demonstrate that the estimated peaBrain variant effects correlate more strongly with coefficients from the univariate eQTL analysis, compared

to log-skew effect estimates obtained from massively parallel reporter assays (MPRAs)[11] and bi-allelic targeted STARR-seq (BiT-STARR-seq)[12], or log fold changes (logFC) of perturbed epigenetic states from DeepSEA[8]. To highlight the utility of the Stage 2 models, we identified putatively functional eQTLs in LCLs that are missed by experimental high-throughput approaches that characterise variant function, such as MPRAs, BiT-STARR-seq, and high-definition reporter assays (HiDRA)[13].

RESULTS

peaBrain captures >50% of the variance in mean gene abundance.

To predict the tissue-specific mean abundance of genes (Stage 1), we leveraged the reference genome[14]. For each gene, as input, we generated a 1-dimensional (1D) matrix centred on the region around the annotated transcription start site (TSS). By varying the length of the input sequence, the 4kbps promoter (2kbps upstream and 2kbps downstream of the annotated TSS) was determined as the best-performing length for predicting the tissue-specific mean gene abundance in the GTEx dataset, outperforming 2kbps and 6kbps promoter sequences (see **Online Methods** and **Supplementary Figure 1**). We used one-hot encoding (four channels) to represent the four DNA letters (A, T, C, G) in the reference genome (4 channels) (see **Online Methods**). The model output was the corresponding predicted mean RNA abundance of that gene, after rank-transformation to normality.

We applied this framework to all tissues from the GTEx dataset[15], constructing three classes of models: (a) using DNA sequence alone **(**class-A**)**; (b) using DNA plus epigenomic annotations not specific to any tissue or cell type (i.e. non-specific annotations) **(**class-B**)**; and (c) using DNA combined with both non-specific tissue-specific annotations **(**class-C**)**. For class-B models, we incorporated 28 channels of binary sequences that represent epigenomic (and related) annotations that are not specific to any cell type or tissue (curated by the authors of LD Score Regression[16];see **Online Methods** for details). For class-C models, we added additional channels corresponding, for those tissues where such data were available, to the consolidated epigenomes from the Epigenomics Roadmap, including tissue-specific peaks from H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27me3, and H3K36me3 ChIP-seq experiments, and experimentally-derived DNase hotspots[17].

We observed that DNA-only (class-A) models captured nearly a fifth of the variance in mean gene abundance across all GTEx tissues (10-fold cross-validated median out-of-sample-r^2 [oos-r^2] values across all tissues = 17%). Addition of non-specific regulatory annotations (class-B models) markedly

improved model performance across all tissues (median cross-validated oos-r^2 = 45%; **Figure 1**). (We average the oos-r^2 across all 10-folds within a tissue and use the median across all tissues to assess global performance; see **Online Methods**.) For example, for EBV-transformed lymphocytes, the 10-fold cross-validated average oos-r^2 is 56% for the class-B model compared to the 15% in the corresponding class-A model. Addition of tissue-specific annotations further improved model performance, such that class-C models captured more than half the variance for almost all GTEx tissues where such data were available (**Figure 1**).

These results are suggestive that differences in mean abundance between genes are largely encoded in differences between core promoter elements and interacting regulatory factors encoded in the model weights, rather than a consequence of non-transcriptional downstream regulation (e.g. silencing by small non-coding RNAs). This is broadly consistent with anecdotal experimental evidence[18]. Explicitly incorporating experimental transcription factor binding site (TFBS) annotations has limited effect on performance (median cross-validated oos-r^2 = 23%), when compared to the complete class B model with epigenetic/histone marks and chromatin annotations (median cross-validated oos-r^2 = 46%; **Supplementary Note 1**). This suggests explicitly encoding TFBS annotations is largely redundant and that epigenetic and genomic annotations add information to that contained in the DNA sequence to substantially improve predictive performance. Importantly, this performance was only accomplished using the convolutional neural network architecture of peaBrain: experimental models that we generated in skeletal muscle using regularized linear models fitted with stochastic gradient descent exhibited poor performance. In fact, for these linear models the 10-fold average oos-r^2 was negative, indicating that the out-of-sample predictions of the model fitted on the training data are worse than predicting the mean of the test set (see **Online Methods** and **Supplementary Note 1**). We also describe comparisons of the peaBrain CNN approach with other methods in **Supplementary Note 1**.

peaBrain score outperforms existing measures in predicting disease-associated variants and in predicting allele-specific transcription factor binding.

Having demonstrated the predictive ability of the model (Stage 1), we were interested in using peaBrain to generate a non-coding impact metric, which captured the impact of each position in the core promoter sequence on the expression of each gene. We defined the impact of each position as the absolute difference in abundance between the original promoter sequence and a modified promoter sequence where all the information for that site (including epigenetic and genomic annotations) is set to zero. To facilitate comparison across tissues, we performed this analysis using the class-B models, since the non-specific epigenetic and genomic annotations were, by definition, available for all tissues. Across all GTEx tissues, the non-coding impact metric correlated with variant-specific conservation scores derived from multiple alignments of 99 vertebrate genomes to the human genome[14] and represented by phylogenetic p-values (phyloP) (see **Online Methods**). Briefly, these phyloP nucleotide conservation scores are based on an alignment and a model of neutral evolution[14]: a more positive value indicates conservation or slower evolution than expected, with the magnitude of the phyloP score corresponding to the -log p-values under the null hypothesis (i.e. neutral evolution). For every unit of absolute magnitude increase in impact, we observed an average increase of 8.95 in phyloP scores, indicating increased conservation (8.95 order-of-magnitude difference in the –log10 p-value; **Supplementary Table 1**). Equivalently, for every unit increase in phyloP, we observed an approximately 0.1 absolute magnitude change in the average normalized expression of the affected gene (i.e. peaBrain impact score); again indicating that if a site is more conserved, it has a larger impact on expression. While this positive trend between conservation and impact on expression was consistent across most GTEx tissues, there were exceptions: in the nucleus accumbens (basal ganglia)**,** noncoding transcriptomic impact was correlated with accelerated evolution (**Supplementary Table 1**). These results were consistent, albeit weaker, after rank-normalization of both the phyloP and peaBrain scores (**Supplementary Table 1**).

This overall positive correlation between peaBrain impact and phyloP represents a direct equivalence between evolutionary conservation and impact on gene abundance. Most well-established non-coding impact measures (e.g. CADD[19] and Eigen[20]) indirectly capture transcriptomic consequences by modelling evolutionary conservation measures, allele frequency, and/or functional non-coding consequence annotations. However, the peaBrain-derived impact metric directly assesses the

contribution of a genomic position on mean expression. Importantly, since the metric is independent of curated consequence and disease annotation databases – as it is trained solely on expression from "healthy" tissues – it provides an unbiased estimate of the information content and deleterious impact of variation at any genomic position in the core 4kbps promoter sequence. The peaBrain impact scores, for all tissues, have been made available (see **URLs**).

Having established the correlation between peaBrain impact and evolutionary constraint, we were interested in assessing the utility of peaBrain-derived scores to interrogate disease-associated variants. We compared the performance of the non-tissue-specific peaBrain score (see **Online Methods**) to two other non-coding metrics (CADD[19] and Eigen[20]) across a series of tasks (**tasks A-C**; all tasks are summarized in **Supplementary Table 2**).

First, we made use of data on disease-related variation from the Catalogue of Somatic Mutations in Cancer [COSMIC][21] , limited to the census gene set which defines a set of genes with somatic mutations causally implicated in human cancer (see **Online Methods**)**.** In **task A,** we assessed the predictive capacity of the non-coding metric to identify positions with non-zero incidence of cancer-associated somatic mutation (n=5268), among all genomic positions within the 4kbps core promoter sequences of COSMIC census genes (approximately 2.15 million positions), using a simple logistic model. The logistic coefficients give the change in the log odds of the outcome (i.e. presence or absence of somatic mutation) for a one-unit increase in the non-coding score. In **task B,** we similarly assessed the predictive capacity of the non-coding metric to identify, using the same COSMIC data set, positions with recurrent cancer-associated somatic mutations (n=544) when contrasted to positions with non-recurrent cancer-associated somatic mutations (n=4724). The focus on cancer-associated somatic mutations allowed us to circumvent linkage disequilibrium (LD) confounding. Patterns of recurrent non-coding somatic mutations, across all tumours in these genes, provide a coarse indicator of the functional transcriptomic impact of non-coding genomic positions. Both tasks were modelled with the allele frequency and phyloP conservation incorporated as covariates (see **Online Methods**). We subsequently assessed the significance of the logistic model coefficients for each of the non-coding metrics across the two tasks

(**Table 1**). Only the non-specific-peaBrain score, derived from scores across all GTEx tissues (average across all tissues and positions), was positively and significantly predictive for both tasks (**Table 1**). Significance was assessed using the default two-tailed p-value corresponding to the z ratio based on the Normal reference distribution (**Table 1**; see **Online Methods**). The non-specific peaBrain-derived metric was useful in isolating genomic positions with non-zero incidence of somatic mutations across all positions in the promoters of COSMIC consensus genes **(task A;** coefficient point estimate=29.36; 95% confidence interval [ci] (16.63, 41.97)**)**, and could further delimit positions with recurrent somatic mutations **(task B**; coefficient=102.96 [64.58, 140.72]). Eigen was significantly predictive for **task A** (coefficient = 0.10 [0.08, 0.12]), but not for **task B** (0.08 [-0.01, 0.17]). CADD exhibited the opposite trend between **tasks A and B**: negatively predictive of genomic positions with non-zero incidence of somatic mutations (coefficient = -0.05 [-0.08, -0.02]), but positively predictive of positions with recurrent somatic mutations (coefficient = 0.17 [0.06, 0.28]; **Table 1**). Thus, the non-coding peaBrain-derived metric appears to better characterize the pathogenicity and putative functionality of non-coding variants with transcriptomic consequences in the core-promoter sequences, providing additional information to that found in allele frequency or evolutionary constraint metrics and with performance better than other established non-coding impact scores.

Having demonstrated the utility of peaBrain impact in predicting disease-associated variants, we were interested in investigating the discriminative ability of peaBrain-derived scores in identifying allele-specific transcription factor binding sites (**task C**). We hypothesized that allele-specific binding will have downstream transcriptional consequences that could be identified from directly modelling gene expression. For brevity, we briefly note that only the peaBrain impact score was significantly predictive of allele-specific binding sites (coefficient = 35.38 [12.00, 58.67]; p = 0.003; **Table 1**) and neither CADD nor Eigen achieved nominal significance. We provide a more detailed analysis in **Supplementary Note 1**. We also highlight, in **Supplementary Note 1**, how characterizing TF motifs is not necessary to understand the consequences of sequence variation on TF binding by comparing peaBrain with methods specifically designed to predict TFBS, including two neural-network methods

(DeepBind[22] and DeepSEA[8]), two kmer-based variant scoring methods (gkmSVM[23] and GERV[24]), and three position-weighted matrices (PWM)-related methods[25].

Tissue-specific peaBrain scores can identify the functional tissues underlying GWAS signals from complex traits

For **tasks A-C**, we have used the non-tissue-specific peaBrain score (average of score, per position, across all tissues) to facilitate comparison with the other tissue-agnostic impact metrics. However, we sought to investigate advantages of tissue-specific impact scores. In particular, we wanted to highlight how tissue-specific scores could allow us to identify functional tissues associated with GWAS signal from complex traits (**task D**). We hypothesized that the "true" functional gene(s) downstream of a GWAS locus ("hit") would have, on average, higher peaBrain impact scores for the tissue in which the gene is likely to act, given that >50% of the variance in mean gene abundance can be explained by the promoter sequence. In other words, we hypothesized that genes associated with a given phenotype (e.g. total cholesterol) are also likely to be transcriptionally perturbed in the underlying functional tissue (e.g. liver), which we can detect with tissue-specific peaBrain scores.

We selected 4 quantitative traits (total cholesterol[26], LDL[26], HDL[26], and triglycerides[26]) for which the (primary) putatively causal tissue is well-established and included in the GTEx dataset. Using HESS[27], we calculated the local SNP-heritability from the relevant GWAS summary statistics, while accounting for linkage disequilibrium. For European populations, HESS partitions the genome into 1703 approximately-independent LD blocks (average length = 1.6Mb)[27]. For each block (or "locus"), we calculated the tissue-specific peaBrain impact score for each GTEx tissue; the locus peaBrain score is defined as the average of the tissue-specific peaBrain scores at all positions (with a score) within that locus. We subsequently performed a regression of the rank-transformed local SNP-heritabilities as a function of the rank-transformed peaBrain locus scores to minimize bias caused by outlying loci and assessed significance for the linear model coefficient (n = 45 tests for each GTEx tissue per phenotype; see **Online Methods**). As a baseline benchmark, we compared our results to tissue predictions made

using the tissue trait concordance (RTC) score[28], which was adapted to calculate the probability that a GWAS-associated variant and an eQTL are co-localized and weighted by the extent of tissue sharing for the given eQTL to obtain tissue-causality profiles for each trait. Across all tested traits, we noted the peaBrain framework was better at identifying putatively causal tissues than simply using the RTC-/eQTL-based method (**Supplementary Tables 3 and 4**). For LDL, using the peaBrain framework, the top five tissues (ranked by nominal p-value) were: EBV-transformed lymphocytes, visceral adipose, fibroblasts, liver, and terminal ileum (small intestine); all were significant after Bonferroni adjustment with p-values tabulated in **Supplementary Table 3**. In contrast, with the RTC-based method, the top five tissues were: sun-exposed skin (from lower leg), pancreas, fibroblasts, tibial nerve, and cerebellar hemisphere (brain). This was consistent across all tested traits (e.g. for HDL, liver ranked 3rd using the peaBrain framework and 32nd using the RTC-based method; **Supplementary Tables 3 and 4**). The superior peaBrain performance suggests inherent limitations to eQTL-based methods that are sidestepped by the Stage 1 peaBrain framework, which depends only on the average expression of all genes in a single tissue and the reference genome. Notably, peaBrain is independent from the number of eQTLs identified per tissue and the number of genome-wide significant hits for a given trait, which are both limitations for eQTL-GWAS co-localization methods (such as the RTC-based framework).

Having validated the peaBrain Stage 1 approach and its utility in a diverse set of tasks, in **Supplementary Note 1,** we highlight how activations of the penultimate layer of the peaBrain model can be used as a continuous and compressed representation (i.e. embedding) of genes. These embeddings, or equivalently, neural activations, capture both the annotated DNA (input) and its additive contributions to tissue-specific abundance (output) in a compressed form amenable to downstream analyses (such as network-based analyses). These embeddings display interesting properties (see **Supplementary Note 1**), including the encoding of correlation information and membership to pathways/curated gene sets. Importantly, these embeddings are in a linear space, such that the pairwise cosine similarity between these dense gene representations is proportional to the measured RNA-seq correlation between the gene pair. In other words, co-regulation and co-expression may be discovered

by leveraging linear structure within the embeddings (e.g. adding embeddings of two genes to discover their co-expression with a third).

peaBrain can predict transcriptomic consequences of individual variation.

Having shown the utility of the Stage 1 peaBrain model, we extended the peaBrain model to incorporate the transcriptomic consequences of individual genotype variation (**Stage 2**). Given whole genome sequencing data of a group of individuals (such as GTEx participants), we sought to assess the ability of this extended peaBrain model to predict the tissue-specific expression profile of each individual, and to identify putatively functional variants within the sequence.

To do this, we constructed, for each gene and in each tissue, an extended peaBrain model that takes individual genome sequence as input and predicts the tissue-specific expression of the corresponding gene as output. (For stage 2 analyses, we did not make use of individual level epigenomic and regulatory annotations as these were not available.) More concretely, unlike stage 1 models, for a single gene, stage 2 models predict the difference between the expression of two individuals as a function of the difference in the sequences between the two individuals (for the given gene; see **Online Methods**). By jointly modelling the input "difference" sequence in a non-linear manner, we hypothesized that we would capture information relevant to *cis*-heritability missed by linear models (such as distance to TSS sites and the pairwise relationships between variants), and be able to prioritize functional variants with transcriptomic consequences solely from the DNA sequence. This additional information is modelled by using the "difference" sequence as input, rather than the dosage in variation. (Stage 2 peaBrain models were trained separately from Stage 1 models, but share similar architectures; see **Online Methods**.)

Consistent with evidence from eQTL studies[29], we noted the 4kbps core promoter used in Stage 1 did not capture enough cis-heritability as estimated by constrained GCTA[30] and thus was not sufficiently informative for this predictive task. In LCLs, for example, using the 4kbps core promoter, genes with

significant non-zero heritability ($p < 0.01$; n = 1066) had a median heritability of 0.136. We selected a 1Mbps input length, centred on the annotated TSS (0.5Mbps upstream and 0.5Mbps downstream), as a compromise between computational tractability (extending the sequence entails more computational expense) and biological relevance (the potential to capture additional narrow-sense heritability with extended intervals). Using the 1Mbps input sequence, genes with significant non-zero heritability ($p < 0.01$; n = 816) had a median heritability of 0.270; nearly twice the heritability captured with the core promoter 4kbps sequence. Importantly, our symmetric 1Mbps window likely contained >95% of cis-eQTLs; in the GTEx dataset, the 95th percentile for absolute distance of cis-eQTLs from their target transcript TSS was 441,698bps[15]. Complete analysis of a 1Mbps interval (including 5 different train/test splits) for a single gene in a single tissue and 94 individuals, if run sequentially on a CPU, required 15 days with 14 GB of memory. Limited to the genes with significant non-zero heritability in LCLs (n = 816), on 600 cores, the complete analysis took approximately a month. (Stage 1 peaBrain models only required several hours.) Prior to training, for each individual, we re-constructed the 1Mbps input sequence from the variants called from whole genome sequence (WGS) data (see **Online Methods**). Exploring the peaBrain architecture, fine-tuning the model parameters, and deploying the models was conducted on NVidia's P100 GPUs (see **Online Methods** for details); the bulk of the training, however, was run on CPUs.

To assess peaBrain's performance in predicting individual variation in RNA expression levels in comparison to other widely-used *in silico* methods and experimental assays (elastic net[6], DeepSEA[8], MPRA[11], BiT-STARR-seq[12], and HiDRA[13]), we designed four tasks (**tasks E-H**; described below and summarized in **Supplementary Table 2**). For the comparison with elastic net, in line with other recent studies in the field[9], we restricted performance analyses to a set of genes with significant non-zero narrow-sense *cis*-heritability (henceforth, simply referred to as heritability) in LCLs as estimated by constrained GCTA[30] (limited to the 1Mbps input sequence; $p<0.01$; see **Online Methods**). By limiting analysis to genes with detectable *cis*-heritability, we can make more meaningful conclusions about the comparative performance of the different methodologies. We restricted analysis to LCLs to enable comparisons with empirical data (**tasks F-H**) and to reduce the compute burden.

316 **peaBrain identifies functional architecture that is inaccessible with current high-throughput**
317 **experimental assays.**
318
319 First (**task E**)**,** we compared the predictive performance of peaBrain to that of a regularized linear model
320 (an implementation of elastic net identical to that used in PrediXcan). RNA-seq samples from the GTEx
321 dataset (n = 94 individuals after filtering) were pre-processed, residualised to account for cryptic
322 relatedness, biological confounders, and technical variance, and rank transformed to normality (see
323 **Online Methods**) before modelling. Model performance for both linear models and peaBrain was
324 assessed by generating oos-r^2 for 5% of individuals randomly withheld from training and unrelated to
325 individuals in the training set (repeated 3-10 times, depending on the how quickly the model reached
326 the exit criteria and the performance of earlier repeats; see **Online Methods**). For each of the 816 genes
327 with non-zero heritability (GCTA $p < 0.01$), we calculated the 95% confidence interval for the oos-r^2,
328 defining a gene as successfully predicted if the entire oos-r^2 confidence interval exceeded zero to ensure
329 we only consider genes with high-confidence models. Whilst regularized linear models were able to
330 capture cis-heritability for 28 of the 816 genes, the equivalent number for peaBrain was 113. *Cis*-
331 heritability for 3 genes was captured by both models, with the oos-r^2 confidence interval largely
332 overlapping (**Supplementary Table 5**). **Supplementary Table 5** also tabulates the performance
333 metrics (confidence and point estimates for oos-r^2 from both classes of models) and estimated GCTA
334 heritability for all genes.
335
336 Having established the predictive ability of peaBrain, we were interested in whether we can use the
337 best-performing peaBrain models to measure the impact of single variants, compared to DeepSEA log
338 gold change (logFC) estimates and experimental log skew estimates from MPRA and BiT-STARR-seq
339 (**Task F**). For all "captured" genes (n = 113), we selected all variants identified as significant eQTLs in
340 the GTEx v6p univariate eQTL analysis (n = 16,019 variants; **see Methods**) and replicated the analysis
341 with the Geuvadis dataset[31] (n = 17,279 variants for the EU population and n = 1601 variants for the
342 YRI [Yoruba from Ibadan, Nigeria] population). For each eQTL (including indels), we created pairs of
343 artificial sequences that only differed at the corresponding snp/indel position and predicted the

difference in expression between the alternate and reference alleles from the difference between the two artificial sequences. (We used only a single model of the those several trained during cross-validation for simplicity, but incorporating results from additional models may improve results; see **Online Methods**.) For brevity, we briefly note that only the peaBrain predictions were significantly and positively correlated with the univariate eQTL coefficients from the GTEx analysis (Spearman's rho = 0.09; $p = 3.02 \times 10^{-32}$; **Supplementary Figure 2**), from the EU-Geuvadis analysis (rho = 0.10; $p = 9.60 \times 10^{-38}$; **Supplementary Figure 3**), and from the YRI-Geuvadis analysis (rho = 0.18; $p = 8.64 \times 10^{-13}$; **Supplementary Figure 4**). Neither the DeepSEA logFC (for lymphoblastoid cell line annotations), nor the log skew estimates for the MPRA or BiT-STARR-seq assays correlated with the univariate eQTL coefficients from any of the three datasets. **Supplementary Note 2** includes a more detailed analysis and interpretation of the results. Both the MPRA and BiT-STARR-seq experimental assays were run in lymphoblastoid cell lines. Importantly, we did not have any variant-level filters for any of the methods (e.g. using a p-value threshold for the experimental assays or any significance cut-off for the peaBrain estimates); thus, our comparison was not biased towards any method and assessed the utility of the method estimate across the range of variant effects.

Next, we sought to evaluate the performance of peaBrain at identifying putatively-functional eQTLs against empirical data from MPRA, BiT-STARR-seq, and HiDRA (**Task G**). Like MPRA, HiDRA is an extension of the classical reporter gene assay, adapted for sequence constructs derived from accessible DNA regions via ATAC-seq[13]; MPRAs leverage shorter synthesized DNA sequences[32]. BiT-STARR-seq is an extension of self-transcribing active regulatory region sequencing (STARR-seq), which like HiDRA involves fragmenting the genome and cloning fragments 3' of a reporter gene. We considered whether variants with the larger estimated effects from each of the three experimental approaches and peaBrain were preferentially located in sequences with known functional relevance (e.g. accessible DNA or transcriptionally active chromatin) and depleted from quiescent or repressed regions. The sequence annotations were derived from the Roadmap's GM12878 lymphoblastoid cell line 15-state ChromHMM model; the same GM12878 cell line was also used for both experimental assays (MPRA and HiDRA). BiT-STARR-seq was also performed in a lymphoblastoid cell line, but

372 the exact cell line was not specified[12]. For each chromatin annotation, we assessed significance using a
373 simple logistic model after rank-transformation of all estimates to normality (to ensure coefficients were
374 comparable; see **Online Methods**). The coefficient of the model corresponded to the extent to which
375 each approach was predictive of chromatin states/accessibility. More concretely, the logistic
376 coefficients give the change in the log odds of the annotation overlap for a one-unit increase in the
377 normalized score.
378
379 For peaBrain, as opposed to analysing the consequences of all possible variants/indels within 1Mbps
380 input sequences for the "captured" 113 genes (which is computationally expensive), we focussed our
381 analysis on all 23,595 univariately-significant eQTLs (from either the GTEx or Geuvadis datasets). We
382 noted that variants with higher peaBrain estimates were significantly enriched in DNase accessible sites
383 and transcriptionally active regions, and significantly depleted from heterochromatin and repressed
384 sequences (**Table 2**). In contrast, the magnitudes of the MPRA log skew estimates were not significantly
385 associated with any chromatin state or accessibility annotation after Bonferroni correction (**Table 2**).
386 This absence of enrichment/depletion was consistent whether we analysed all variants assessed on the
387 platform (n = 26,986 variants after excluding those with no match in Ensembl's VEP database; see
388 **Online Methods**) or limited our analysis to the subset of variants also present in the peaBrain analysis
389 (n = 1589 MPRA variants; i.e. univariately-significant eQTLs for the 113 "captured" genes). It is
390 important to note that variants assessed on the MPRA platform were already selected, in part, because
391 their eQTL status in the Geuvadis dataset; that is, excluding negative controls and LD-based selection,
392 all variants assessed on the MPRA assays were univariately-significant eQTLs.
393
394 Similarly, variants with high magnitudes of the BiT-STARR-seq log skew estimates were not
395 significantly enriched in transcriptionally active chromatin (or depleted from repressed/quiescent
396 intervals), irrespective of whether we assessed performance on all variants assessed on the platform (n
397 = 43,494) or limited to univariately-significant eQTLs for the 113 "captured" genes (n = 621). Using
398 nominal p-value thresholds, HiDRA performed better than either MPRA or BiT-STARR-seq when
399 looking at all variants assessed on the platform (n = 32,906 variants), but no annotation reached

significance after multiple testing correction. Even when limited to the variants present in the peaBrain analysis (n = 199 univariately-significant eQTLs for the 113 "captured" genes), no significant enrichment or depletion was discovered for any annotation.

For all four methods, we did not apply any (significance) filter at the variant-level; that is, to ensure a fair comparison between all four methods, we did not select significantly active variants/fragments. Selecting the subset of variants significant for each method (e.g. using DESeq2 for HiDRA, QuSAR-MPRA for MPRA/BiT-STARR-seq, or a simple one-sample t-test across the peaBrain model repeats) would improve the results for the corresponding method (potentially biasing the test). It is important to note that we can generate confidence intervals/test-statistics for peaBrain estimates by assessing the prediction in each of the model replicates (trained and tested on different subsets of individuals); an idea conceptually similar to biological replicates in the experimental assays. However, the performance of a single cross-validated peaBrain model was deemed sufficient and thus, this assessment was not conducted. We should also note that the authors of the three experimental assays have convincingly shown that the methods, when limited to active fragments or significant variants (specific to each method), are able to identify functional variants enriched in transcriptionally active regions and depleted from heterochromatin[11-13]. However, this enrichment/depletion is limited to the subset of variants labelled as significant by the respective methods, i.e. the allelic log skew estimates are not insightful outside this limited subset. By comparing across all variants (without any significance filtering), we are able to show that peaBrain predictions from a single gene model are more informative (across the entire range of variant effects) than allelic log skew estimates from any of the experimental assays. In other words, peaBrain estimates can side step the noise inherent in assessing variant impact with experimental assays.

Having established that variants with higher peaBrain estimates are enriched in transcriptionally active chromatin (irrespective of any variant-level filtering), we sought to subsequently evaluate the four aforementioned methods on a more granular level using RegulomeDB[33] (**Task H**). The chromatin states and DNA accessibility assessed in **Task G** are only coarse indicators of variant function. RegulomeDB

annotates variants in intergenic regions with known and predicted regulatory elements and categorizes each variant based on the evidence supporting regulatory function of the variant[33]. As RegulomeDB contains annotations from multiple tissues, we selected variants with well-established regulatory function in the GM12878 cell line ("Category 1"), which includes variants matched to known TF binding with matched TF motif and matched DNase footprint. For peaBrain and the three experimental assays (MPRA, BiT-STARR-seq, and HiDRA), we assessed significance using a simple logistic model after rank-transformation of all method estimates to normality (to ensure coefficients were comparable; see **Online Methods**). Similar to **Task G** (with chromatin states and accessibility), the coefficient of the model corresponded to the extent that each approach was predictive of variants with established regulatory function. In other words, larger coefficients indicate that the method is better able to delineate established regulatory variants from variants with minimal evidence for regulatory function. We note that only peaBrain had a significant and positive coefficient; with larger peaBrain estimates indicating variants with well-established and stronger evidence for predicted regulatory function (coefficient = 0.15 [0.04, 0.28]; **Table 3**). None of the three experimental assays had significantly positive coefficients (for all variants tested on the respective platforms and limited to the subset of eQTLs for the 113 "captured" genes). Overall, on the post-selective 113 genes, **Tasks F-H** suggest that the modelling undertaken by Stage 2 peaBrain (derived from sequence data alone) detects functional architecture that is not readily accessible with the latest high-throughput empirical approaches.

DISCUSSION

Here, we have introduced a two-stage computational framework for predicting the transcriptomic consequences of non-coding variation. Using Stage 1 class-C (tissue-specific annotated) models, we observed that the majority of variance (>50%) in the mean abundance of genes across most GTEx tissues is encoded in the annotated 4kbps core promoter sequences. Thus, the difference in mean abundance between genes appears to be largely encoded in invariant differences between core promoter elements and the interacting tissue-specific regulatory factors encoded in the model weights, rather than a consequence of transcriptional regulation by more distal sequences or non-transcriptional downstream regulation (e.g. silencing by small non-coding RNAs). Furthermore, we note that the average expression of all genes in a single tissue and the reference genome is sufficient to learn both TFBS and allele-specific binding (see **Supplementary Note 1**). Taken together, this is broadly consistent with anecdotal experimental evidence[18] and suggests that non-transcriptional downstream processes play a secondary role in regulating mean expression.

The predictive ability of Stage 1 peaBrain models allowed us to calculate a non-coding impact score for all genomic positions in the core promoter sequences, a useful metric for analysis of both common rare variants. Unlike other non-coding metrics that incorporate external consequence annotations (e.g. from Ensembl's variant effect predictor [VEP], ClinVar, and other curated databases), peaBrain impact score is derived directly from predicting expression and does not depend on curated variant annotations. The tissue-specific nature of the peaBrain impact score is useful for identifying putatively functional tissues underlying GWAS signal for complex traits, which are not readily accessible through current methods that rely on eQTL-GWAS-hit co-localization.

To incorporate the consequences of individual variation on gene abundance in Stage 2 of the framework, we extended the Stage 1 model to capture a 1Mbps window, a balance between computational tractability and biological "signal". Unlike Stage 1 models, Stage 2 peaBrain leverages differences in the sequences between individuals to predict differences in expression (rather than prediction from the

sequence directly, see **Online Methods** for implementation details); that is, the sequence arrays are subtracted from each other and the resultant "difference sequence" captures how shifted the two sequences are and the differences in alleles. Without the sequence, peaBrain would simply be modelling SNP dosage (i.e. conceptually no different from existing [linear] models) and that is not sufficient for prediction of putatively functional variants as observed. Thus, the distance information encoded implicitly by modelling the sequence appears important to peaBrain performance. However, peaBrain is a black-box approach and we must be cautious in attempting to elucidate scientific rationales for the apparent improved performance. Existing methods for peering into "black-box" approaches are not particularly useful for peaBrain as it leverages differences between individual sequences aligned to the annotated TSS, rather than conventional (reference) sequences, that are modelled with "conjoined" neural networks (see **Supplementary Table 7**). In other words, we cannot readily extract meaningful motif sequences from the input data. Reconstruction of the individual sequences to generate the difference input required that we use a quality controlled VCF to reconstruct individual sequences (see **Methods**), as opposed to directly using the originally "noisy" sequence reads. However, by leveraging differences in "TSS-aligned" sequences, peaBrain learns to map differences at each genomic position of an individual (relative to the fixed TSS landmark) to predict difference in expression. The advantage of this approach is that peaBrain must learn to pinpoint important features regardless of where they occur in the sequence and that may eschew the overfitting concern associated with *a priori* identification of eQTLs. Importantly, Stage 2 peaBrain does not directly depend on eQTLs/variation dosage, but rather focusses on how differences at each genomic position (because of differences in alleles or because of shifts due to upstream/downstream indels) perturb expression.

At this conjecture, it is important to note that, unlike many methods with similar conceptual origins, peaBrain was not designed with the sole intent of predicting gene expression abundance. Rather, one of the primary goals of Stage 2 peaBrain models is identifying putatively functional eQTLs. As a first approximation, we note that peaBrain variant effect estimates positively and significantly correlate with the coefficients from the univariate eQTL analysis on the post-selective 113 "captured" genes. In contrast, MPRA and BiT-STARR-seq allelic log skew estimates did not correlate with the

corresponding univariate eQTL coefficients. Furthermore, variants with large peaBrain estimates were significantly enriched in DNase-accessible DNA and transcriptionally active chromatin, and depleted from quiescent and repressed states. Log skew estimates, for both MPRA and BiT-STARR-seq for variants were uninformative of chromatin state for the subset of variants investigated. The poor performance of MPRA may reflect the fact that it is an episomal assay so variants are not being assessed in their regular chromatin context. Variants with large HiDRA estimates were nominally enriched in transcriptionally active regions, but did not reach significance after Bonferroni correction. Notably, however, both the MPRA and HiDRA assays were performed in the GM12878 cell line from which the chromatin and DNA accessibility annotations were also derived, i.e. there is a possibility that the results for the experimental assays are biased over-estimates of true performance. It is important to note that when limited to the subset of significant variants (as labelled by each method), the experimental assays can identify regulatory variants enriched in transcriptionally active chromatin. The log-skew estimates from any of the three experimental assays, however, cannot delineate functional variants outside this limited "significant" subset. In **Tasks F-H**, by comparing across all variants (without any significance filtering), we show that peaBrain variant predictions are more informative (across the entire range of variant effects) than allelic log skew estimates from the experimental assays. More concretely, as described above, peaBrain estimates can side step the noise inherent in variant-level measurements using *in vitro* empirical assays.

As with other deep learning approaches, there are limitations to peaBrain analysis; notably, that despite our best efforts for the rigorous quality control and model regularization, there may be some information that is biasing performance results in an intricate way (i.e. the generic problem of using black-box neural network models). To mitigate the risk of bias, we implemented dropout regularization, out-of-sample testing on unrelated individuals (after conservatively filtering for cryptic relatedness), comparison with high throughput assays (such as MPRAs and HiDRA), and validation using chromatin, TF-binding, and DNA accessibility annotations. However, without an explicit model, there is always a possibility for bias. For peaBrain, the ability of the Stage 2 analyses to identify putatively functional variants that are enriched in transcriptionally active chromatin and depleted from heterochromatin/repressed sequences

530 is encouraging evidence of model generalizability. Similarly, the correlation of the impact scores from
531 Stage 1 analyses with evolutionary constraint and their utility in predicting disease-associated mutations
532 and allele-specific binding sites further underscores the true performance of peaBrain framework.
533
534 All together, the results from the Stage 1 and Stage 2 of the peaBrain framework suggest that models
535 for understanding the effects of non-coding variation on RNA abundance (and possibly more complex
536 traits) can be built by relying more on automated machine learning, rather than hand-designed or
537 selected features. Furthermore, the results highlight the variant sensitivity of the Stage 2 peaBrain model
538 and its ability to identify putatively functional variants underlying cis-eQTL signals. More generally,
539 peaBrain's performance in predicting mean abundance and individual variation further implicates the
540 importance of the invariant genomic context and distance to the annotated TSS for interpreting the
541 effects of non-coding variation in a tissue-specific manner.

ONLINE METHODS

RPKM and gene count data, for Stage 1 and Stage 2 peaBrain models, was downloaded from GTEx (v7; see URLs)[15]. To prepare the data for Stage 1 of peaBrain, the mean abundance of each gene was obtained by averaging the RPKM across all subjects. The values were then rank transformed to normality using the rntransfrom function from GenABEL v1.8-0. The GRCh37 (hg19) reference genome was downloaded from UCSC[14]. We used the default Ensembl gene definitions to define gene borders; an Ensembl gene is defined as the collection of all spliced transcripts with overlapping coding sequences but excluding manually annotated readthrough genes. The gene start and end coordinates (from which the core promoter sequences are defined) correspond to the outermost transcript start and end coordinates. We accounted for gene strand-ness while extracting the core promoter sequences; start coordinates corresponding to the TSS for genes on the positive strand and the end coordinate corresponding to the TSS for genes on the negative strand. We further limited our analysis to protein-coding genes (n = 19,820 genes) and to autosomal chromosomes for simplicity. For all Stage 1 models, the DNA promoter sequence for each gene was one-hot encoded (also known as a one-of-k scheme); each letter represented as separate channel. One-hot encoding is a technique commonly used in natural language processing to encode categorical integer features with each channel indicating the presence (1) or absence (0) of the corresponding DNA letter. Processed genomic annotations and epigenetic markers were obtained from the LDSC[16] (see **URLs**) and similarly processed. For Stage 1 class B models and using the LDSC annotations, we incorporated an additional 28 channels of binary sequences for each base-pair, that are not specific to any cell type or tissue, highlighting: coding basepairs, conserved sites[34], CTCF sites, DGF peaks[2], DHS peaks[3], enhancers[4,35], fetal DHS peaks[3], H3K27ac peaks[5,17], H3K4me1 peaks[3], H3K3me3 peaks[3], H3K9ac peaks[3], introns[14], promoters[14], promoter flanking sequences[4], repressed sites[4], super enhancers[5], transcription factor binding sites (TFBS)[2], transcribed sequences[4], TSS[4], untranslated 3' regions (UTR3)[14], untranslated 3' regions (UTR5)[14], and weak enhancers[4]. Stage 1 class C models included additional binary channels, corresponding to the consolidated epigenomes from Roadmap (see **URLs**), as described in the main text. Transcription factor processed ChIP-seq data were also downloaded from the gene transcription regulation database (GTRD

v17.04; see **URLs**). GTRD is a database of human transcription factor binding sites identified from ChIP-seq experiments and uniformly processed. As described in **Supplementary Note 1**, for a subset of Stage 1 models, transcription factor binding sites identified using four different peak callers (MACS, SISSR, GEM and PICS) and clusters of peaks for each method (defined as overlapping peak called using the protocol, but in different tissues or under different conditions) were included as separate binary channels.

Stage 1 peaBrain model was constructed using Theano 0.9.0 and Lasagne 0.1. For a single tissue, peaBrain takes in the core promoter sequence as input and predicts the normalised mean abundance of the corresponding gene (**Figure 1**). The core promoter sequence was determined by varying the length of the promoter sequence (± 1kbps, ± 2kbps, and ± 3kbps). As highlighted in **Supplementary Figure 1**, ± 2kbps (i.e. the 4kbps core promoter sequence) was the optimal length for predictive ability as assessed using a 10-fold cross-validation scheme. The input sequence is a 1D vector with 4 channels encoding the DNA sequence and when appropriate, additional channels as binary representations of various genomic annotations and epigenetic markers (described above). The Stage 1 peaBrain model is a series of 1D convolutions and max pooling layers (**Supplementary Table 6**). In practice, a 1D convolution is implemented as a 2D convolution with width set to 1 (effectively dropping the unused dimension). Each convolutional layer was set with 11 filters of size 5 and a leaky rectify non-linearity activation function. The leaky rectify activation function for all convolutional layers has a nonzero gradient for negative input, which is useful for convergence[36]:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{if } x \leq 0 \end{cases}$$

The 0.01 corresponds to the "leakiness" of the activation function, with larger values denoting increased "leakiness". The input to the first convolutional layer is 4000 x 1 x r sequence, where 4000 corresponds to the length of the core promoter sequence and r denotes the number of channels (minimum of 4 DNA letter channels). The first convolutional layer has 11 filters (or equivalently, kernels) of size 5 x 1 x 11, where 5 denotes the sequence length of the filter and 11 denotes the number of channels for that filter. The output of each filter is a locally connected structure, convolved with the sequence, to produce 11

feature maps that are then max pooled with the output of other filters from the layer, before serving as input for the subsequent layer. Prior to the penultimate embedding layer (from which we extract the continuous vector gene representations), we placed a dropout layer with p = 0.5 of setting values to zero. The dropout layer is a regularizer that randomly zeros input values (i.e. randomly dropping units and their connections), limiting co-adaptation and improving model generalizability [37,38]. The number of units in the penultimate embedding layer determines the size (or the number of components) in the vector and was set to 1001. The last layer is a single output neuron that outputs the mean abundance of the corresponding gene (for which the promoter was input). The last two dense layers (including the final output neuron) have linear activations, ensuring that the normalized mean abundance is a linear combination of the embedding components or equivalently, the neural activations of the penultimate layer. The objective was defined using the mean squared difference (between predictions and observed mean abundances) and model weights were updated using Adam with the learning rate=0.001, beta1=0.9, beta2=0.999, and epsilon=1 $x10^{-8}$. Adam is an algorithm for gradient-based optimization of (stochastic) objective functions[39]; beta1 corresponds to the exponential decay rate for the first moment estimates and beta2 is the decay rate for the second moment estimates. The model was trained for a minimum of 100 epochs, before exiting early using a validation set (defined as 10% of the training). As is typical in neural networks, the number of layers and other explicitly-specified model variables, above, are referred to as hyperparameters; they are variables that set prior to optimization of the models parameters.

Pre-processing for heritability and variant-sensitive regression for the Stage 2 peaBrain model was performed as recommended by the authors of QTLTools[40]. For each tissue, we selected genes with non-zero RPKM values for at least 50% of samples. Per gene, RPKM values were residualised using linear regression to account for autolysis score, date of nucleic acid isolation, date of genotype isolation, RIN, total ischemic time, time spent in paxgene fixative, sex, age, Hardy score, interval of onset to death for last underlying cause, number of hours in refrigeration, ischemic time, temperature, donor status (post-mortem, surgical or organ donor), three genotype PCs and enough expression matrix PCs to explain 55% of the variance (to account for unexplained technical and biological variance). The residuals were

then rank-transformed to normality using GenABEL's rntransform function. As with Stage 1 peaBrain analyses, we further limited our analysis to protein-coding genes (number of genes differed between tissues) and to autosomal chromosomes for simplicity. For each protein-encoding gene on an autosomal chromosome, we defined the input sequence as 0.5 Mbps upstream and 0.5Mbps downstream of the TSS using default GRCh37 Ensembl gene definitions (total 1Mbps centred on the TSS). As highlighted in the main text and supplementary notes, the 1Mbps was selected as a balance between computational tractability and biological relevance. Increasing the length of the input sequence beyond the 1Mbps increases both the compute time and memory footprint. Importantly, our symmetric 1Mbps window likely contained >95% of cis-eQTLs; in the GTEx dataset, the 95th percentile for absolute distance of cis-eQTLs from their target transcript TSS was 441,698bps[15]. Incidentally, the 1Mb interval have also used by other approaches in imputing RNA expression from genotype (namely, TWAS)[9]. Using the unphased whole genome sequencing GTEx data, we reconstructed the individual's sequence from the quality controlled VCF. In other words, we generated the individual variation by substituting each individual's non-reference alleles into the reference sequence. Variants in the WGS GTEx VCF were quality controlled by GTEx LDACC at the Broad Institute. As stated in the corresponding README file, quality control was conducted using GATK, Hail, and PLINK. Notably, a variant was removed if it didn't "pass Variant Quality Score Recalibration (VQSR), had low Inbreeding Coefficient or low Quality Score, was within a Low Complexity Region (LCR), became monomorphic after applying genotype quality score (GQ) <20 or allele balance (AB) >0.8 or AB<0.2 filters or assigning male heterozygous calls in chrX nonPAR regions to missing, had missingness rate >= 15%, did not pass Hardy-Weinberg Equilibrium testing in African American or European subpopulations for autosomes or in European females for chr X, showed significant association with sequencing technology or library construction batch, or showed significant non-random missing of reference alleles."[15] For each individual, we generated two copies of the gene 1Mbps input sequence; phasing did not matter as the sequences were combined prior to modelling.

Stage 2 peaBrain models for heritability analysis and variant-sensitive prediction were similarly constructed as described for the Stage 1 models. Stage 2 models, however, are three separate

653 convolutional neural networks, connected by a dense fully-connected layer prior to the output neuron
654 (**Supplementary Table 7**). The input 1Mbps sequence is split into three inputs: 0.48Mbps upstream,
655 4kbps core promoter, and 0.48Mbps downstream sequences. The 4kbp core promoter is the input to a
656 CNN with identical structure and hyperparameters as described for Stage 1 peaBrain model (described
657 above in detail). The upstream and downstream sequences are input to networks with identical
658 architecture, but different pooling hyperparameters: a pool size of 100 for the first pooling layer, 50 for
659 the second, and 10 for the last. Number of filters was consistent between all networks (n = 11). The
660 fully connected output from each sequence is concatenated, before one penultimate fully-connected
661 layer and a single output node. Unlike the Stage 1 peaBrain models, Stage 2 models are trained to
662 predict the differences between individuals (rather than direct prediction of expression). As humans are
663 diploid, for each individual, the input sequence was the sum of the one-hot encoding of each of the
664 1Mbps sequences corresponding to the "maternal" and "paternal" sequences ; phasing did not matter
665 because the sum was consistent. A separate Stage 2 model was constructed for each gene with
666 significant non-zero heritability (see **text**). For any pair of individuals, A and B, the input sequence was
667 defined as the difference between the one-hot encoded sequences, with the corresponding output as the
668 difference between the two individuals. We included both differences, (A – B) and (B – A), during
669 training. After removing individuals with cryptic relatedness (see GCTA analysis below), the GTEx
670 dataset was randomly split into train and test individuals (95% of subjects for training and 5% for
671 testing), with the model trained on all the pairwise differences between train individuals and tested on
672 all pairwise differences between test individuals. The training set was further sub-divided into training
673 and validation sets, with the latter used to exit early after a minimum 100-epoch training. As described
674 below, overall model performance was assessed using the oos-r^2 on five to ten random repetitions of
675 95/5 train/test splits; the number of repetitions was dependent on how quickly each model reached exit
676 criteria.
677
678 **Elastic net (regularized linear) models.** We used an additive genetic model as our baseline
679 comparison as described elsewhere[6]. Briefly, for each gene, an elastic net model was used to model
680 expression (alpha = 0.5; selected to match PrediXcan[6]). As with peaBrain, the models were trained to

predict the difference in expression as a function of the difference in dosages among the variants within the 1Mb input sequence (rather than the expression directly). For a linear model, this is no different from simply predicting the expression; the constant term in this case is expected to be close to 0. The lambda (regularization) parameter was 3-fold cross-validated on the training dataset, using cv.glmnet function from glmnet v2.0-10 [41].

Model performance, for all peaBrain and linear models, was assessed using the out-of-sample-r^2 (oos-r^2), a classical machine learning metric to assessing performing of regression models (often just called r^2)[42]. oos-r^2 is defined as:

$$\text{oos-}r^2 \equiv 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y}_{test})^2}$$

where f_i denotes the predicted value using the model fitted on the training data, y_i denotes the true value for, and $\bar{y}_{test}$ is the mean value for all items in the test set. The denominator of the oos-r^2 is the total sum of squares (proportional to the variance of the data) and the numerator of the oos-r^2 is the explained sum of squares (also called the regression sum of squares). When the explained sum of squares (numerator) is larger than the total sum of squares (denominator), oos-r^2 is below zero and indicates the model does not have any predictive ability. Regression models with some predictive capacity have oos-r^2 values in the range (0, 1]. Stage 1 peaBrain model performance was assessed using 10-fold cross validation (10% of genes were withheld from the algorithm during training). Stage 2 peaBrain models and elastic net linear models were assessed using repeated random splits of 95% of subjects for training and 5% of subjects for testing. Individuals with cryptic relatedness were removed prior to the training/test split, using GCTA grm-cutoff of 0.025 (see below). 5-10 random training/test splits were used to assess model performance; 95% confidence interval was estimated using the mean and standard error, assuming the distribution of oos-r^2 was normal.

Tissue-specific peaBrain impact score for any given genomic position, as described in **text**, was defined as the absolute difference in abundance between the original promoter sequence and a modified promoter sequence where all the sequence and epigenetic/genomic annotations for that site were set to zero. The impact score is proportional to the contribution of the genomic position to the average expression of the gene; genomic positions are readily mapped to genes by virtue of the promoter definitions. If the genomic position overlapped with the promoter of multiple genes, the maximum impact across all overlapping genes was taken. Tissue-specific peaBrain impact scores were compared to phylogenetic p-values (phyloP) using simple linear models (lm base function in R). As briefly described in the main text, phyloP are nucleotide conservation scores derived from multiple alignments of 99 vertebrate genomes to the human genome phyloP scores are based on an alignment and a model of neutral evolution[14]. A more positive value indicates conservation or slower evolution than expected; magnitude of the phyloP score corresponds to the -log p-values under the null hypothesis (i.e. neutral evolution). phyloP scores were downloaded from the UCSC genome browser (see **URLs**). To compare peaBrain to other non-coding metrics, a non-tissue-specific peaBrain score was used; defined as the average impact of each position across all tissues. Non-coding impact scores (combined annotation dependent depletion [CADD] v1.3, and Eigen v1.1) were downloaded from their respective webpages (see **URLs**). CADD is a single meta-score derived from analysis of multiple annotations for variants that survived natural selection, compared to simulated mutations[19]. Eigen is an unsupervised score that synthesizes a combination of functional annotations into one meta-score[20]. The non-coding somatic mutations used to assess metric performance were downloaded from the COSMIC v82 (see **URLs**). Allele frequency was derived from gnomAD release 170228. For each genomic position, we counted the number of overlapping somatic mutations. We further limited our analysis to COSMIC census genes (as a positive gene set); COSMIC census genes possess mutations that have been causally implicated in cancer. For each task used to compare the non-coding metrics (see text), a logistic model was used (fitted using the glm function in R, family = "binomial") with the allele frequency and phyloP as covariates. Allele frequency and evolutionary conservation scores were included to assess whether the non-coding impact score adds any additional information to the model, besides that derived from allele frequency or evolutionary constraint. Positions without a phyloP conservation score were excluded

from model fitting. The confidence interval was obtained using the confint function (derived from profiling the likelihood function). For the analysis of recurrent somatic mutations, we were interested in the global performance of each metric at each autosomal chromosome and thus a simple model sufficed – isolating genes or promoters with "mutation hotspots" would require more sophisticated approaches to avoid false positives (e.g. it would be necessary to incorporate tumour type, the proportion of each tumour [sub]type, the background mutation rate at each position/tumour, and more technical variables such as sequence coverage). The published non-coding impact scores (CADD and Eigen) depend on curated non-coding annotations and indirectly predict transcriptomic consequences; that is, there is potential risk of overestimating the performance of these scores in the three tasks (see **Main Text**). Allele-specific binding site data and prediction scores for TF binding prediction algorithms were downloaded from the Supplementary Table appended to Wagih *et al.*[25] (see **URLs**). For the comparison between non-coding impact metrics, duplicate sites were filtered (selecting the one with lowest nominal p-value). For the analysis of causal tissues, we downloaded the summary statistics for the four lipid traits from the webpage of the Global Lipids Genetics Consortium (see **URLs**). Local SNP-heritability for each trait was calculated using HESS (Heritability Estimation from Summary Statistics). The linear models of local heritability as a function of average peaBrain score per locus were fitted using the base function *lm* in R.

Constrained GCTA heritability analyses. We converted the GTEx whole genome sequencing VCF to PLINK binary bed file (using Plink v1.9). Using GCTA v1.24.4[30], we calculated the genetic relationship matrix (GRM) from all the autosomal SNPs and excluded individuals with grm-cutoff of 0.025. GCTA was used to calculate heritability for similar methods, including predixcan[6] and TWAS[9]. For each gene, we subsequently limited the GRM to variants within the 1Mb input sequence (centred on the TSS) and performed constrained GCTA-GREML analysis. Genes with a significant non-zero heritability ($p < 0.01$) were included for subsequent analyses.

Predicative ability of gene embeddings was assessed using a 10-fold cross validation scheme. The hallmark curated gene sets were downloaded from Molecular Signatures Database v6.0[43]. Hallmark

gene sets represent an aggregation of many gene sets and are thought to represent coherent biological states or processes. For each set, genes were assigned a binary label (1 denoting membership). We subsequently trained a multi-layer perceptron classifier from scikit-learn v0.19.0, with three hidden layers (200, 100, and 50 neurons), to predict gene-set membership using the gene's embedding. We used a rectified linear unit function as the activation for our hidden layers, and lbfgs for weight optimization. Lbfgs is an optimizer that belongs to the family of quasi-Newton methods. Cosine similarity between any pair of embeddings was assessed using eponymous function from scikit-learn, defined as:

$$\text{similarity} \equiv \frac{X \cdot Y}{\|X\| \, \|Y\|}$$

where X and Y denote the embeddings for genes X and Y, respectively. Correlation between the RNA-seq arrays for genes X and Y were calculated using the base cor function in R.

Correlation with univariate GTEx/Geuvadis eQTL analysis, DeepSEA, and MPRA & BiT-STARR-seq log skew estimates. To calculate the effects of single variants, artificial sequences were constructed that differed only at the genomic position of the corresponding variant; with one sequence containing the reference (ref) allele and one sequence containing the alternate (alt) allele. As Stage 2 peaBrain model predicts the difference between two sequences, we used the (alt – ref) configuration to estimate an effect size for each variant. Univariate eQTL coefficients were obtained from the GTEx and Geuvadis datasets (see **URLs**). Significance of spearman (rank) correlation between the peaBrain estimate and eQTL coefficient was assessed using the cor.test function in R. Both the univariate eQTL analysis and peaBrain were obtained using expression data that was rank transformed to normality and thus are comparable in magnitude (despite slightly different pre-processing protocols). MPRA variant results were obtained from Supplementary Table 1 of Tewhey *et al.*[11] (see **URLs**); snp rs ids were translated to chromosome_position_ref_alt_build nomenclature using Ensembl's GRCh37 biomaRt and a simple python script. Any variant that intersected with the peaBrain final variant set was included in the analysis, that is, variants in the 1Mbps input sequence for genes/models with the 95% confidence interval for the oos-r^2 entirely above zero. The LogSkew.Comb column, corresponding to the log2

790 allelic skew from the combined MPRA LCL analysis (alt/ref), was used as the MPRA log skew
791 estimate. The BiT-STARR-seq data was similarly processed (see **URLs**). As with the peaBrain and
792 eQTL analysis, significance of the spearman (rank) correlation between each of the experimental assays
793 allelic log skews and the univariate eQTL coefficients was assessed using the cor.test function. To
794 obtain the logFC for GM12878 annotations, a vcf file of the corresponding eQTLs was uploaded to the
795 DeepSEA platform (see **URLs**).
796
797 **Comparison of peaBrain, MPRA, BiT-STARR-seq and HiDRA.** The core 15-state model and
798 DNAse accessibility annotations for the GM12878 EBV-transformed lymphoblastoid cell line (LCLs)
799 were downloaded from the Roadmap project (see URLs). HiDRA data was downloaded from the GEO
800 series GSE104001 (see URLs). HiDRA estimate was defined as the log fold change in average counts
801 between the alternate and reference group (after normalizing for DNA count); direction did not matter
802 as only the magnitude was used in this analysis. The MPRA and BiT-STARR-seq data was downloaded
803 and pre-processed as described above. Notably, the chromatin states/DNA accessibility annotations, the
804 HiDRA, and MPRA estimates were derived from the same cell line; that is, there is possibility of
805 overestimating the performance of either method. For any given annotation, we assessed the predictive
806 ability of the magnitude of the variant estimate (from any of the three approaches) to predict whether
807 the variant overlapped with the annotation. The magnitude of the variant estimate for each approach
808 (peaBrain, HiDRA, BiT-STARR-seq, and MPRA) corresponded to either the transcriptomic impact or
809 activity of that variant. Only the absolute magnitude, after rank-transformation to normality, of each
810 variant was used in modelling. For each approach and for each annotation, a logistic model was used
811 (fitted using the glm function in R, family = "binomial") and the confidence interval was obtained using
812 the confint function (derived from profiling the likelihood function). For the granular variant-level
813 assessment, annotations were downloaded from RegulomeDB (dbSNP 141; see URLs).
814
815
816
817

818 **Code availability statement**

819 peaBrain models are available here: http://www.well.ox.ac.uk/~moustafa/peaBrain.shtml .

820

821 **Data availability statement**

822 The primary data modelled in this study are available from the GTEx consortium. Where appropriate,
823 we have provided a minimal running example with all custom code. Data generated from the models
824 (e.g. transcriptomic impact scores) have also been made available at:
825 http://www.well.ox.ac.uk/~moustafa/peaBrain.shtml .

826

827 **URLs**

828 BiT-STARR-seq, https://www.biorxiv.org/content/early/2017/09/27/193136.figures-only

829 CADD v1.3, http://cadd.gs.washington.edu/download

830 COSMIC v82, https://cancer.sanger.ac.uk/cosmic/download

831 DeepSEA, http://deepsea.princeton.edu/job/analysis/create/

832 Eigen v1.1, http://www.columbia.edu/~ii2135/download.html

833 Global Lipids Genetics Consortium, http://csg.sph.umich.edu/abecasis/public/lipids2013/

834 GTEx, https://www.gtexportal.org/

835 GTRD v17.04, http://gtrd.biouml.org/

836 HiDRA GEO, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104001

837 LDSC Epigenetic Annotations, https://data.broadinstitute.org/alkesgroup/

838 MPRA Supplementary Table,

839 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4957403/bin/NIHMS787218-supplement-7.xlsx

840 peaBrain data, http://www.well.ox.ac.uk/~moustafa/

841 phyloPl http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP100way/

842 Roadmap Annotations, http://egg2.wustl.edu/roadmap/web_portal/

843 RegulomeDB, http://www.regulomedb.org/downloads

844 Transcription factor binding sites (allele-specificity),

845 https://www.biorxiv.org/content/early/2018/02/01/253427.figures-only

846

ACKNOWLEDGEMENTS

Moustafa A. is supported by Post Graduate Doctoral Scholarships from the Rhodes Trust and the Natural Sciences and Engineering Council of Canada.

AUTHOR CONTRIBUTIONS

Moustafa A. designed the study, with guidance and input from C.C.H. and M.I.M. Moustafa A. developed the method and test tasks, with discussion and advice from Mohamed A. C.C.H. and M.I.M. supervised the study and interpretation of the results. Moustafa A. wrote the paper, with contributions and discussion from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

863 **TABLES**

864

865

Table 1. Tabulated statistics (to two decimal places) from the logistic models for the three non-coding metrics from **tasks A-C**. **Task A** assesses the predictive capacity of the non-coding metric to identify positions with non-zero incidence of cancer-associated somatic mutations in the core promoter regions. **Task B** assesses the predictive capacity of the non-coding metric to identify positions with recurrent cancer-associated somatic mutations, among all positions with at least one somatic mutation. **Task C** assesses the predictive capacity of the non-coding metric to identify variants within the 4kbps core promoter with allele-specific binding (for a subset of positions for which data was available). All three tasks were assessed using simple logistic models, with the allele frequency and phyloP incorporated as covariates. Positions without a phyloP score were excluded from model fitting (see **Online Methods**). peaBrain is the only non-coding metric with significant coefficients for all three tasks; we used a non-tissue-specific peaBrain score to facilitate comparison with the tissue-agnostic CADD and Eigen scores (see **Main Text**). The bounds for the 95% confidence interval, obtained by profiling the likelihood function, are tabulated, with significant coefficients denoted in bold. **Abbreviations:** L, lower; U, upper.

Metric	Task	Logistic Coefficient	L Bound	U Bound	p-value
peaBrain	**A**	**29.36**	**16.63**	**41.97**	**5.56×10^{-6}**
	B	**104.50**	**66.05**	**142.31**	**7.66×10^{-8}**
	C	**35.39**	**12.00**	**58.67**	**2.95×10^{-3}**
CADD	**A**	**-0.05**	**-0.08**	**-0.02**	**1.57×10^{-3}**
	B	**0.17**	**0.06**	**0.28**	**2.83×10^{-3}**
	C	0.06	-0.03	0.16	0.20
Eigen	**A**	**0.10**	**0.08**	**0.12**	**$< 2 \times 10^{-16}$**
	B	0.06	-0.003	0.12	6.65×10^{-2}
	C	0.04	-0.002	0.08	0.07

Table 2. In the 113 "captured" genes, eQTL variants with higher peaBrain estimates (i.e. more likely to be functional and with larger predicted transcriptomic impact) tended to fall in DNase accessible sites and transcriptionally active regions, and were similarly depleted from quiescent and repressed sequences (**Task G**). This trend was not observed for variants with large MPRA or BiT-STARR-seq log skew magnitudes, irrespective of whether we assessed performance on all variants on the platform or limited to univariately-significant eQTLs for the 113 "captured" genes. HiDRA performed better than MPRA and BiT-STARR-seq when using all variants assessed on the assay (all; n = 32,906 variants); performance further dropped when limited to the variants present in the peaBrain analysis (shared; n = 199). Point estimates were derived from fitting a simple logistic model with the scores from each assay rank-transformed to normality (i.e. model coefficients are directly comparable). Nominal p-value is presented in parentheses, but only entries that are significant after Bonferroni correction are shown in bold. (Green denoting enrichment; orange denoting depletion.) It is important to note that we did not filter based on the significance of the estimate for any of the methods (see **Main Text**). By comparing across all variants (without any significance filtering), we are able to show that peaBrain predictions from a single gene model are more informative (across the entire range of variant effects) than allelic log skew estimates from any of the experimental assays.

	peaBrain	MPRA log-skew		HiDRA		BiT-STARR-seq log-skew	
		all	shared	all	shared	all	shared
DNase accessibility	**0.12 (3.76×10^{-5})**	0.01 (0.463)	0.21 (2.64×10^{-2})	0.01 (0.352)	-0.05 (0.732)	**-0.05 (2.92×10^{-13})**	0.09 (0.438)
TssA	**0.32 ($<2 \times 10^{-16}$)**	0.03 (0.218)	0.25 (2.05×10^{-2})	-0.02 (0.140)	0.01 (0.934)	0.00 (0.965)	0.09 (0.360)
TssAFlnk	0.10 (1.55×10^{-2})	0.06 (4.72×10^{-2})	0.29 (2.17×10^{-2})	0.03 (4.01×10^{-2})	0.61 (4.23×10^{-3})	-0.02 (1.28×10^{-2})	0.03 (0.788)
TxFlnk	-0.13 (7.69×10^{-2})	-0.01 (0.81)	-0.16 (0.672)	0.08 (0.141)	1.61 (5.59×10^{-2})	-0.02 (0.424)	-0.06 (0.787)
Tx	-0.02 (0.297)	-0.04 (4.91×10^{-2})	-0.05 (0.471)	-0.14 (6.13×10^{-3})	-0.80 (9.58×10^{-2})	0.01 (0.571)	-0.07 (0.325)
TxWk	0.04 (1.42×10^{-2})	-0.01 (0.662)	-0.01 (0.923)	0.01 (0.630)	0.48 (0.136)	**0.03 (4.66×10^{-4})**	0.00 (0.973)
EnhG	0.04 (0.547)	-0.12 (6.97×10^{-3})	-0.26 (0.167)	0.04 (0.529)	-0.91 (4.35×10^{-2})	-0.05 (1.85×10^{-2})	-0.21 (0.349)
Enh	0.03 (0.345)	0.01 (0.693)	-0.04 (0.757)	0.00 (0.910)	-0.71 (3.23×10^{-2})	**-0.03 (1.39×10^{-3})**	0.03 (0.872)
ZNFRpts	-0.07 (0.176)	0.05 (0.310)	0.03 (0.858)	-0.01 (0.910)	-0.06 (0.869)	-0.01 (0.724)	0.16 (0.137)
Het	**-0.14 (3.00×10^{-7})**	0.04 (0.223)	-0.22 (0.169)	0.05 (0.303)	-0.17 (0.816)	0.01 (0.761)	0.24 (0.101)
TssBiv	0.66 (0.14)	-0.04 (0.881)	1.10 (0.277)	-0.01 (0.958)	NA	-0.05 (0.381)	-0.18 (0.812)
BivFlnk	0.19 (0.549)	-0.22 (0.293)	-0.52 (0.605)	-0.24 (9.80×10^{-3})	0.15 (0.839)	-0.05 (0.356)	-0.19 (0.800)
EnhBiv	0.40 (0.117)	0.08 (0.701)	-0.46 (0.426)	0.13 (0.306)	NA	-0.05 (0.302)	NA
ReprPC	0.20 (0.129)	-0.21 (3.80×10^{-2})	NA	-0.05 (0.653)	-0.81 (0.440)	-0.03 (0.230)	-0.26 (0.735)
ReprPCWk	**-0.25 ($<2 \times 10^{-16}$)**	-0.033 (0.133)	NA	-0.05 (3.73×10^{-2})	-1.00 (6.80×10^{-2})	-0.02 (2.45×10^{-2})	-0.19 (0.159)
Quies	0.01 (0.342)	0.02 (0.192)	-0.02 (0.667)	0.00 (0.718)	-0.13 (0.564)	**0.02 (4.22×10^{-4})**	-0.01 (0.896)

894 **Table 3.** In the 113 "captured" genes, peaBrain estimates can significantly delineate variants with
895 established regulatory function (**Task H**). The log-skew estimates from the experimental assays, both
896 across all variants assessed on each platform ("all") and limited to eQTLs for the 113 "captured"
897 genes ("shared"), are uninformative. Both the peaBrain estimates and log-skew for the experimental
898 assays were rank-transformed to normality to facilitate comparison between the methods. The bounds
899 for the 95% confidence interval, obtained by profiling the likelihood function, are tabulated, with
900 significant coefficients denoted in bold. **Abbreviations:** L, lower; U, upper.

Method	Variants	Logistic Coefficient	L Bound	U Bound	p-value
peaBrain		**0.16**	**0.04**	**0.28**	**7.99×10^{-3}**
MPRA	all	0.10	-0.002	0.19	5.46×10^{-2}
	shared	0.25	-0.06	0.56	0.119
BiT-STARR-seq	all	-0.03	-0.11	0.04	0.371
	shared	0.09	-0.16	0.34	0.461
HiDRA	all	0.09	-0.02	0.20	0.107
	shared	0.11	-0.30	0.51	0.603

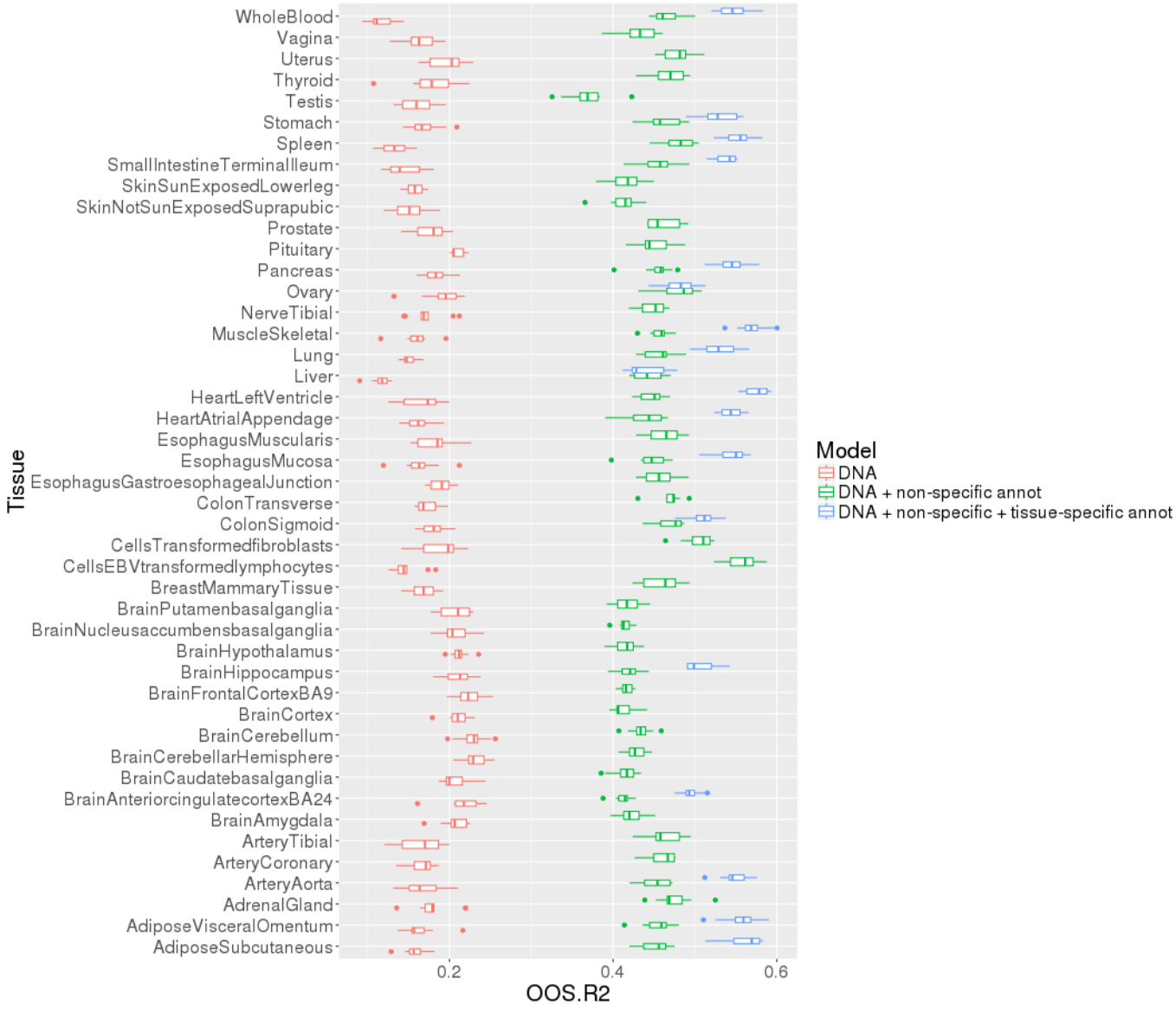
901
902

903 **FIGURES**

904

905

Figure 1. Incorporating genomic and epigenetic annotations improves the performance of peaBrain to predict the normalized mean abundance across all GTEx. The 4kbps promoter sequence, when annotated with tissue specific annotations, is sufficient to predict the majority of variance in mean expression in most tissues, ordered alphabetically from the x-axis. The boxplots highlight the distribution of the 10-folds used to cross-validate model performance. Prediction using regularized linear models performs considerably worse (10-fold cross-validated oos-$r^2 < 0$; **Supplementary Note 1**). **Abbreviations:** OOS.R2, out-of-sample r^2.



916 **SUPPLEMENTARY TABLES**

917

918

Supplementary Table 1. Tabulated summary of coefficients of the linear function modelling phyloP conservation scores as a function of tissue-specific peaBrain noncoding impact metric. Generally, across most tissues and chromosomes, the larger the impact a position has on the mean abundance of the gene (as indicated by a higher peaBrain impact metric), the more evolutionary conserved it is (i.e. a positive coefficient). The notable exception is the nucleus accumbens (basal ganglia), where the opposite trend is noted (negative coefficients; in bold). All coefficients are significant ($p < 10^{-16}$). The results were also consistent with the rank-normalized phyloP and peaBrain scores. The p-values for all Spearman correlations were also significant ($p < 10^{-16}$). **Abbreviations:** L, lower; U, upper.

	Linear Model Coefficient	L Bound	U Bound	Spearman Correlation
AdiposeSubcutaneous	15.79	15.72	15.86	0.03
AdiposeVisceralOmentum	14.32	14.25	14.39	0.03
AdrenalGland	0.33	0.27	0.39	0.00
ArteryAorta	3.51	3.47	3.56	0.01
ArteryCoronary	5.64	5.59	5.69	0.01
ArteryTibial	5.89	5.84	5.94	0.02
BrainAmygdala	9.25	9.19	9.31	0.02
BrainAnteriorcingulatecortexBA24	6.92	6.85	7.00	0.01
BrainCaudatebasalganglia	6.19	6.15	6.23	0.02
BrainCerebellarHemisphere	13.48	13.41	13.55	0.02
BrainCerebellum	4.78	4.73	4.83	0.01
BrainCortex	2.70	2.67	2.74	0.01
BrainFrontalCortexBA9	8.57	8.50	8.64	0.01
BrainHippocampus	5.09	5.03	5.14	0.02
BrainHypothalamus	5.17	5.09	5.24	0.01
BrainNucleusaccumbensbasalganglia	**-1.32**	**-1.37**	**-1.27**	**-0.01**
BrainPutamenbasalganglia	6.91	6.86	6.96	0.02
BreastMammaryTissue	4.24	4.19	4.30	0.01
CellsEBVtransformedlymphocytes	5.92	5.87	5.98	0.02
CellsTransformedfibroblasts	9.17	9.13	9.22	0.02
ColonSigmoid	7.28	7.23	7.34	0.03
ColonTransverse	3.64	3.60	3.69	0.01
EsophagusGastroesophagealJunction	7.10	7.04	7.16	0.02
EsophagusMucosa	6.37	6.30	6.44	0.02
EsophagusMuscularis	8.11	8.03	8.18	0.01
HeartAtrialAppendage	11.98	11.90	12.07	0.02
HeartLeftVentricle	7.29	7.21	7.36	0.02
Liver	2.99	2.94	3.04	0.01
Lung	7.93	7.86	7.99	0.02
MuscleSkeletal	4.27	4.23	4.31	0.02
NerveTibial	6.70	6.63	6.77	0.02
Ovary	5.39	5.33	5.45	0.01
Pancreas	11.07	11.00	11.15	0.02
Pituitary	4.67	4.62	4.71	0.01
Prostate	13.57	13.51	13.64	0.03
SkinNotSunExposedSuprapubic	7.43	7.37	7.48	0.02
SkinSunExposedLowerleg	4.06	4.00	4.12	0.01

SmallIntestineTerminalIleum	2.87	2.82	2.92	0.01
Spleen	3.70	3.63	3.76	0.01
Stomach	1.55	1.51	1.58	0.01
Testis	7.36	7.31	7.40	0.02
Thyroid	5.87	5.80	5.93	0.01
Uterus	6.54	6.50	6.58	0.03
Vagina	7.34	7.27	7.41	0.01
WholeBlood	9.56	9.50	9.62	0.02

927

928

929 **Supplementary Table 2.** Tabulated summary of all tasks used to assess peaBrain performance, for both

930 Stage 1 and Stage 2 models.

Task	peaBrain	Description	Methods (dataset)	Results
A	Stage 1	Assess Predictive capacity of the non-coding metric to identify positions with non-zero incidence of cancer-associated somatic mutations in the core promoter regions.	CADD, Eigen (COSMIC)	Table 1
B	Stage 1	Assess predictive capacity of the non-coding metric to identify positions with recurrent cancer-associated somatic mutations, among all positions with at least one somatic mutation.	CADD, Eigen (COSMIC)	Table 1
C	Stage 1	Assess the predictive capacity of the non-coding metric to identify variants within the 4kbps core promoter with allele-specific binding (for a subset of positions for which data was available).	CADD, Eigen, DeepSEA, DeepBIND, GERV, gkmSVM	Table 1 & Supplementary Note 1
D	Stage 1	Investigate how tissue-specific scores can be identify functional tissues associated with GWAS signal from complex traits	RTC (eQTL)-based method	Supplementary Tables 3 & 4
E	Stage 2	Compare predictive performance of peaBrain to regularized linear model	Elastic net	Supplementary Table 5
F	Stage 2	Assess correlation of variant estimates (for the 113 “captured” genes) with coefficients from univariately-significant eQTLS in two different populations	DeepSEA, MPRA, BiT-STARR-seq (GTEx, Geuvadis)	Supplementary Figures 2-4 & Supplementary Note 2
G	Stage 2	Assess predictive capacity of peaBrain estimates to delineate variants enriched in transcriptionally-active chromatin and depleted from quiescent/repressed chromatin states	MPRA, BiT-STARR-seq, HiDRA (Roadmap)	Table 2
H	Stage 2	Assess predictive capacity of peaBrain estimates to delineate variants with established regulatory function.	MPRA, BiT-STARR-seq, HiDRA (RegulomeDB)	Table 3

931

Supplementary Table 3. Tabulated p-values for the top five putatively functional tissues per trait (ranked in ascending order by p-value), as predicted by the peaBrain framework and the RTC (eQTL)-based methodology (**Task D**). peaBrain p-values have been Bonferroni-corrected for multiple testing; results for all tissues are available in **Supplementary Table 4**. Nominal p-values are shown for the RTC (eQTL)-methodology; obtained from Supplementary Table 8 of the corresponding manuscript[28]. Across all tested traits, the peaBrain framework identifies more relevant functional tissues per trait than the RTC-based method.

Abbreviations: LDL, low-density lipoprotein; HDL, high-density lipoprotein; RTC, regulatory trait concordance.

		peaBrain		**RTC (eQTL)-based**	
	Rank	**Tissue**	**adjusted p**	**Tissue**	**nominal p**
LDL	1	CellsEBVtransformedlymphocytes	1.81×10^{-8}	SkinSunExposedLowerleg	1.58×10^{-17}
	2	AdiposeVisceralOmentum	2.54×10^{-8}	Pancreas	1.44×10^{-9}
	3	CellsTransformedfibroblasts	3.45×10^{-8}	CellsTransformedfibroblasts	6.38×10^{-9}
	4	Liver	4.01×10^{-8}	NerveTibial	1.18×10^{-8}
	5	SmallIntestineTerminalIleum	5.06×10^{-8}	BrainCerebellarHemisphere	1.65×10^{-8}
HDL	1	ArteryTibial	9.46×10^{-8}	NerveTibial	2.36×10^{-18}
	2	Stomach	4.46×10^{-8}	AdiposeSubcutaneous	8.16×10^{-16}
	3	Liver	7.37×10^{-8}	CellsTransformedfibroblasts	5.41×10^{-15}
	4	SmallIntestineTerminalIleum	8.41×10^{-8}	SkinSunExposedLowerleg	3.54×10^{-15}
	5	AdiposeVisceralOmentum	1.07×10^{-7}	SkinNotSunExposedSuprapubic	6.39×10^{-14}
Total Cholesterol	1	Liver	4.73×10^{-11}	SkinSunExposedLowerleg	5.38×10^{-25}
	2	CellsTransformedfibroblasts	5.07×10^{-11}	Liver	2.05×10^{-13}
	3	AdiposeVisceralOmentum	8.33×10^{-10}	Pancreas	3.83×10^{-13}
	4	CellsEBVtransformedlymphocytes	2.02×10^{-9}	Thyroid	9.85×10^{-13}
	5	SmallIntestineTerminalIleum	2.99×10^{-9}	SkinNotSunExposedSuprapubic	5.70×10^{-12}
Triglycerides	1	Spleen	3.98×10^{-4}	HeartLeftVentricle	1.64×10^{-21}
	2	AdrenalGland	8.78×10^{-4}	Thyroid	4.25×10^{-21}
	3	CellsEBVtransformedlymphocytes	1.67×10^{-3}	SkinSunExposedLowerleg	1.52×10^{-20}
	4	ArteryCoronary	1.63×10^{-3}	Lung	8.04×10^{-19}
	5	AdiposeVisceralOmentum	1.69×10^{-3}	AdiposeSubcutaneous	1.15×10^{-17}

938 **Supplementary Table 4.** Causal tissue profiles for all lipid traits.

939 *[Table is too large to embed in word document and is available as a separate spreadsheet.]*

940

941

942 **Supplementary Table 5.** Performance metrics (confidence and point estimates for oos-r^2 from both
943 classes of models) and estimated GCTA heritability for all genes with significant heritability (GCTA
944 $p < 0.01$).

945 *[Table is too large to embed in word document and is available as a separate spreadsheet.]*

946

947

Supplementary Table 6. Schematic of the Stage 1 peaBrain model. The number of channels, r, is determined by the number of epigenetic and genomic annotations included in the model (minimum of 4 corresponding to the 4 DNA letter channels in class A models). The Stage 1 class B models have 32 channels, corresponding to 4 DNA sequence channels and 28 annotation channels (see **Online Methods** for details).

Input Sequence: 4000 x r channels
1st Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
1st Pooling Layer: Pool Size = 5, Pad = 1
2nd Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
2nd Pooling Layer: Pool Size = 5, Pad = 1
3rd Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
3rd Pooling Layer: Pool Size = 5, Pad = 1
Dropout Layer: p = 0.5
Dense Fully-Connected Layer: Number of Units = 1001 Linear Activation
Output Layer: Number of Units = 1 Linear Activation
Output Value: Average Gene Abundance

Supplementary Table 7. Schematic of the Stage 2 peaBrain model, which is composed of three separate networks connected by a dense layer prior to prediction. Values in red denote layers with differing values between the three networks. The network for the centre split is identical to the Stage 1 peaBrain model for the core promoter region; the networks for the upstream and downstream splits are identical to the Stage 1 peaBrain model for distal sequences. Thus, Stage 2 peaBrain can be thought of as a consolidation of the separate Stage 1 models.

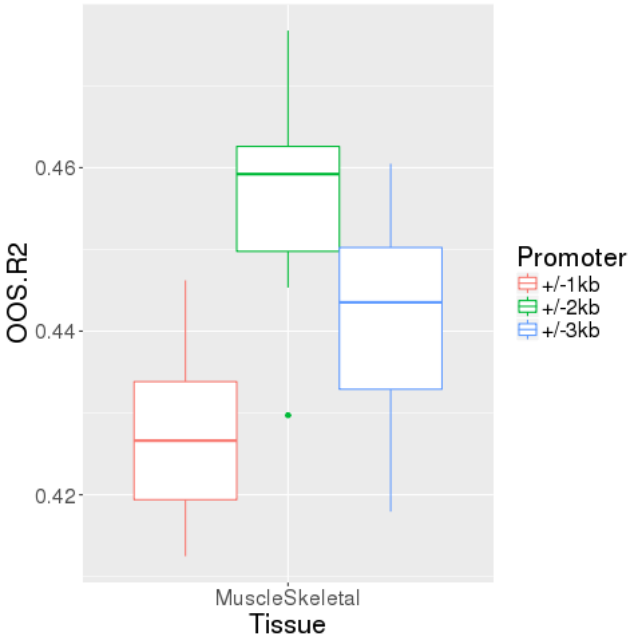
Input Sequence 1Mbps x 4 channels		
Upstream Split 0.498Mbps x 4 channels	**Centre Split** 4kbps x 4 channels	**Downstream Split** 0.498Mbps x 4 channels
1st Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation	**1st Convolutional Layer:** Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation	**1st Convolutional Layer:** Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
1st Pooling Layer: Pool Size = 100, Pad = 1	**1st Pooling Layer:** Pool Size = 5, Pad = 1	**1st Pooling Layer:** Pool Size = 100, Pad = 1
2nd Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation	**2nd Convolutional Layer:** Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation	**2nd Convolutional Layer:** Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
2nd Pooling Layer: Pool Size = 50, Pad = 1	**2nd Pooling Layer:** Pool Size = 5, Pad = 1	**2nd Pooling Layer:** Pool Size = 50, Pad = 1
3rd Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation	**3rd Convolutional Layer:** Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation	**3rd Convolutional Layer:** Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
3rd Pooling Layer: Pool Size = 10, Pad = 1	**3rd Pooling Layer:** Pool Size = 5, Pad = 1	**3rd Pooling Layer:** Pool Size = 10, Pad = 1
Dropout Layer: p = 0.5	**Dropout Layer:** p = 0.5	**Dropout Layer:** p = 0.5
Dense Fully-Connected Layer: Number of Units = 50 Linear Activation	**Dense Fully-Connected Layer:** Number of Units = 50 Linear Activation	**Dense Fully-Connected Layer:** Number of Units = 50 Linear Activation
	Dense Fully-Connected Layer: Number of Units = 50 Linear Activation	
	Output Layer: Number of Units = 1 Linear Activation	
Output Value: Individual Gene Abundance		

965 **SUPPLEMENTARY FIGURES**

966

967

968 **Supplementary Figure 1.** Using the class-B peaBrain model for MuscleSkeletal (largest tissue by
969 sample count in GTEx), the 4kbps promoter sequence (+/– 2kbps of annotated TSS) outperforms both
970 2kbps (+/– 1kbps) and 6kbps (+/– 3kbps) promoter sequences in predicting mean gene abundance.



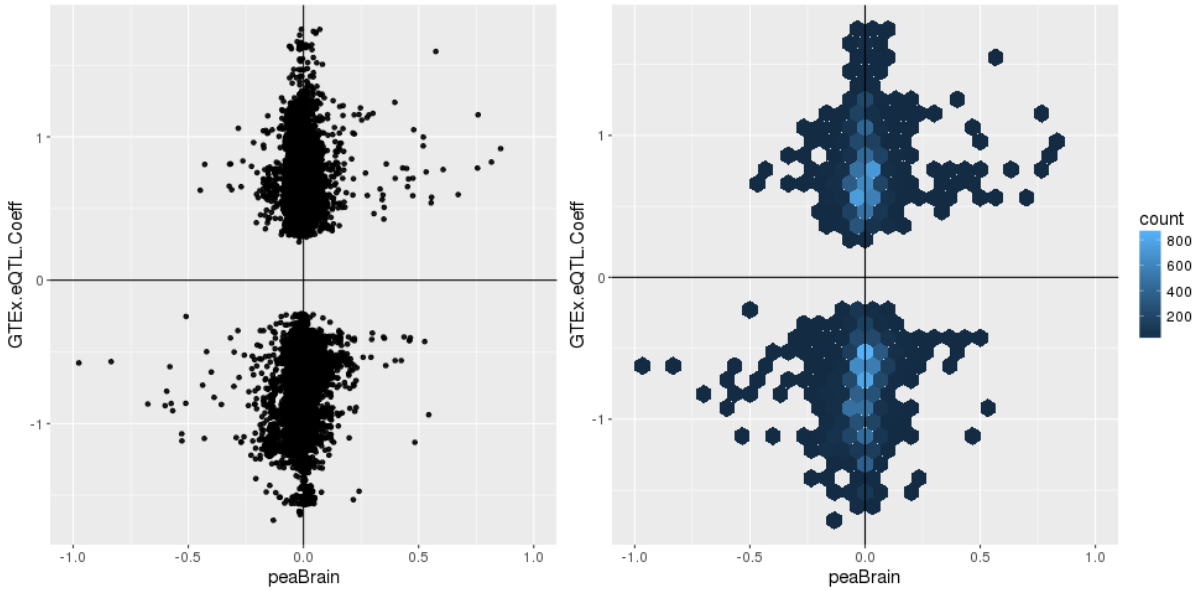
971

972

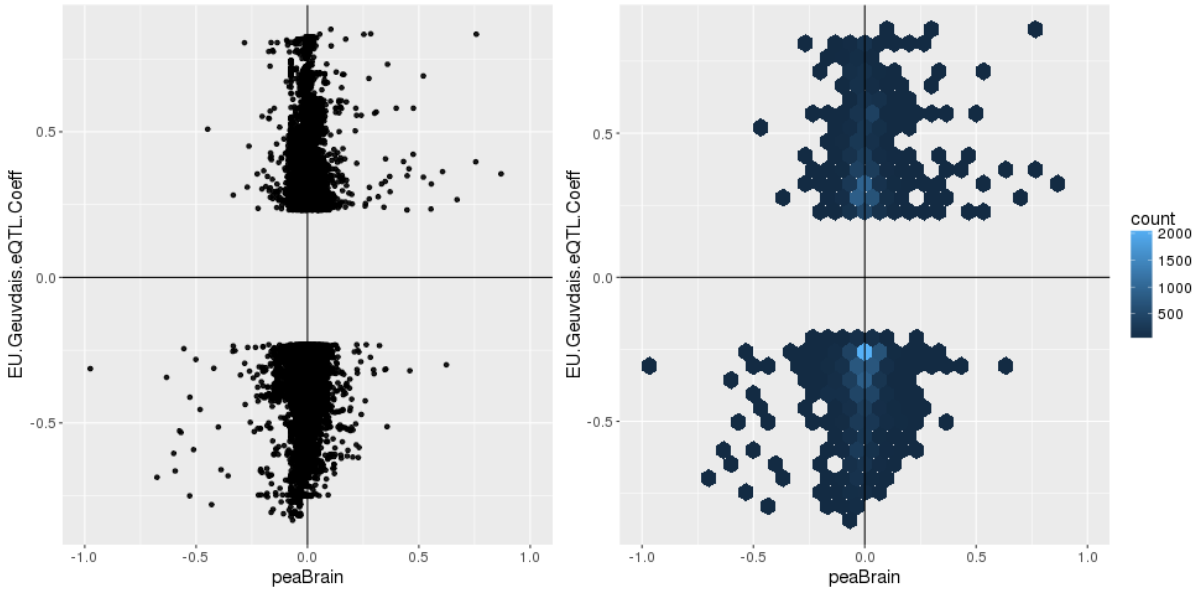
973

974

Supplementary Figure 2. Scatter (*right*) and hexa-bin (*left*) plots of variant-expression effects as estimated in LCLs by peaBrain (limited to genes whose 95% confidence interval for the oos-r^2 is entirely above 0; n = 113 genes; **Task F**). Each point corresponds to a variant that is univariately significant in the GTEx eQTL analysis (n = 16,019 eQTLs). The y-axis is the magnitude of the univariate GTEx eQTL coefficient for the corresponding variant. The correlation between the GTEx coefficient and the peaBrain prediction is positive and significant (Spearman's rho = 0.09; p = 3.02 x10^{-32}).



984 **Supplementary Figure 3.** Scatter (*right*) and hexa-bin (*left*) plots of variant-expression effects as
985 estimated in LCLs by peaBrain (limited to genes whose 95% confidence interval for the oos-r^2 is
986 entirely above 0; n = 113 genes; **Task F**). Each point corresponds to a variant that is univariately
987 significant in the EU-Geuvadis eQTL analysis (n = 17,279 eQTLs). The y-axis is the magnitude of the
988 univariate EU-Geuvadis eQTL coefficient for the corresponding variant. The correlation between the
989 EU-Geuvadis coefficient and the peaBrain prediction is positive and significant (Spearman's rho = 0.10;
990 $p = 9.60 \times 10^{-38}$).

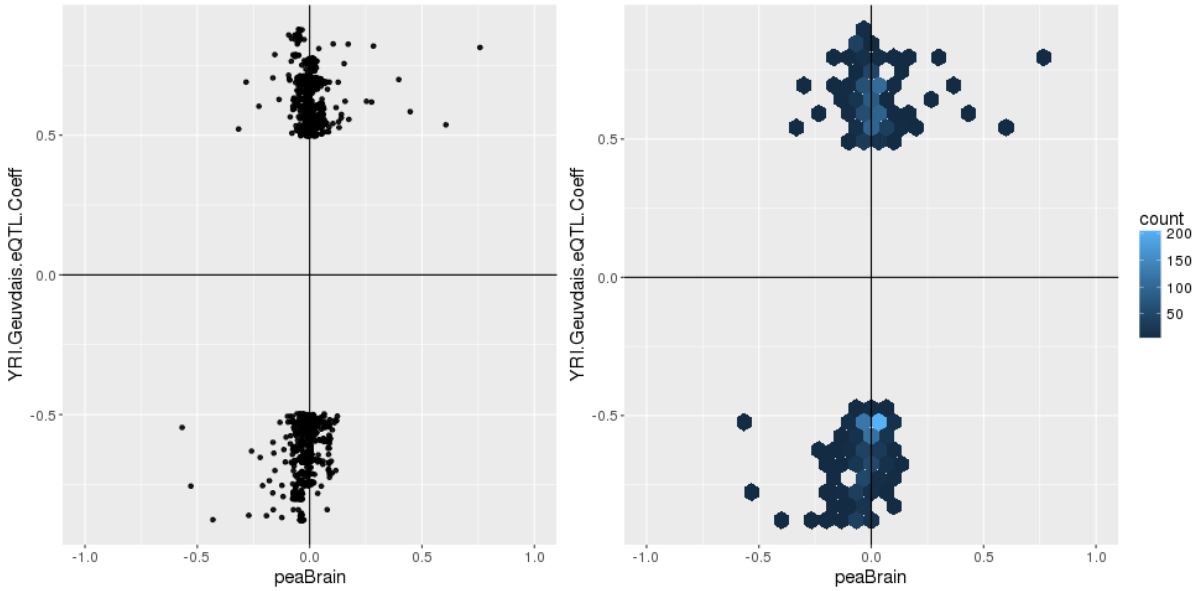


991

992

993

994 **Supplementary Figure 4.** Scatter (*right*) and hexa-bin (*left*) plots of variant-expression effects as
995 estimated in LCLs by peaBrain (limited to genes whose 95% confidence interval for the oos-r^2 is
996 entirely above 0; n = 113 genes; **Task F**). Each point corresponds to a variant that is univariately
997 significant in the YRI-Geuvadis eQTL analysis (n = 1601 eQTLs). The y-axis is the magnitude of the
998 univariate YRI-Geuvadis eQTL coefficient for the corresponding variant. The correlation between the
999 YRI-Geuvadis coefficient and the peaBrain prediction is positive and significant (Spearman's rho =
1000 0.18; $p = 8.64 \times 10^{-13}$).



1001
1002

REFERENCES

1 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362-9367 (2009).

2 ENCODE Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *nature* **447**, 799 (2007).

3 Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics* **45**, 124-130 (2013).

4 Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods* **9**, 473-476 (2012).

5 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).

6 Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091-1098 (2015).

7 Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* **99**, 1245-1260 (2016).

8 Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods* **12**, 931 (2015).

9 Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**, 245-252 (2016).

10 Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research* **26**, 990-999 (2016).

11 Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519-1529 (2016).

12 Kalita, C. A. *et al.* High throughput characterization of genetic effects on DNA:protein binding and gene transcription. *bioRxiv*, doi:10.1101/270991 (2018).

13 Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions in human. *bioRxiv*, 193136 (2017).

14 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome research* **12**, 996-1006 (2002).

15 Consortium, G. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).

16 Finucane, H. K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *bioRxiv*, 014241 (2015).

17 Bernstein, B. E. *et al.* The NIH roadmap epigenomics mapping consortium. *Nature biotechnology* **28**, 1045-1048 (2010).

18 Gasperini, M. *et al.* Paired CRISPR/Cas9 guide-RNAs enable high-throughput deletion scanning (ScanDel) of a Mendelian disease locus for functionally critical non-coding elements. *bioRxiv*, 092445 (2016).

19 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315 (2014).

20 Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics* **48**, 214 (2016).

21 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**, D805-D811 (2014).

22 Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* **33**, 831 (2015).

23 Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics* **47**, 955 (2015).

24 Zeng, H., Hashimoto, T., Kang, D. D. & Gifford, D. K. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* **32**, 490-496 (2015).

25 Wagih, O., Merico, D., Delong, A. & Frey, B. J. Allele-specific transcription factor binding as a benchmark for assessing variant impact predictors. *bioRxiv*, 253427 (2018).

26 Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature genetics* **45**, 1274 (2013).

27 Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics* **99**, 139-153 (2016).

28 Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nature genetics* **49**, 1676 (2017).

29 Grundberg, E. *et al.* Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics* **44**, 1084-1089 (2012).

30 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).

31 Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nature Genetics*, doi:10.1038/ng.3979 (2017).

32 Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology* **30**, 271-277 (2012).

33 Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22**, 1790-1797 (2012).

34 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476 (2011).

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

36 Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* (2015).

37 Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

38 Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research* **15**, 1929-1958 (2014).

39 Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

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

41 Friedman, J., Hastie, T. & Tibshirani, R. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version* **1** (2009).

42 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

43 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545-15550 (2005).