A general-purpose model for emulating expression-based high-throughput assays

Moustafa Abdalla

Supervisors: Mark I. McCarthy and Chris C. Holmes

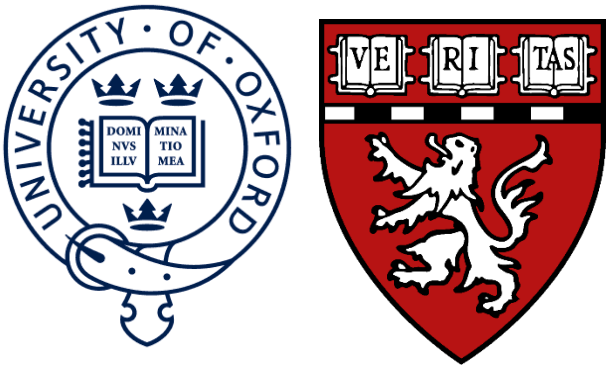
Computational Statistics and Machine Learning Group, Department of Statistics

Wellcome Centre for Human Genetics, Nuffield Department of Medicine

Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine

Brasenose College, University of Oxford

Harvard Medical School, Harvard University



A dissertation submitted in the partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Hilary Term, 2019

DEDICATION

To my patients.

ACKNOWLEDGMENTS

I would like to thank the Rhodes Trust for granting me the opportunity to pursue my DPhil and Harvard Medical School for granting me the deferral.

A general-purpose model for emulating expression-based high-throughput assays

Moustafa Abdalla, Brasenose College, *Doctor of Philosophy*, Hilary Term (2019)

High-throughput assays (HTAs) are the gold standard for scientific experimentation, providing starting points for drug design, for understanding the interactions between genes in biochemical cascades, and for identifying the role of genetic variants in the broader context of cellular processes. Here, I introduce a widely applicable and practically usable deep neural architecture that can recapitulate expression-based HTAs. Trained solely on promoter elements, the model can learn to emulate the transcriptional machinery of a given cell type or tissue – allowing us to construct *in silico* DNA reporter assays, predict the transcriptomic impact of small molecules in cell lines, and infer regulatory networks. By leveraging the tissue-specific properties of the model, I further show how model-derived impact scores can pinpoint functional tissues underlying complex traits, outperforming methods that depend on colocalization of eQTL and GWAS signals. I subsequently demonstrate the utility of this model in investigating the role of DNA and histone modifications in difficult-to-study processes, such as neural induction. Using an extension of the same architecture, I show how the framework outperforms linear models in predicting the consequences of genotype variation. I use the framework to identify putatively functional eQTLs that are missed by experimental high-throughput approaches that characterise variant function. By emulating the transcriptional machinery of human tissues/cells, this represents one of the first attempts to recapitulate expression-based HTAs using a single *in silico* framework. I also introduce a new analytic strategy that leverages coexpression differential dependence networks (i.e. correlation perturbation between gene clusters) to model the effects of genetic variation in a tissue-dependent manner, and to subsequently, discover new loci for type 2 diabetes-related traits. I demonstrate that the newly discovered risk loci are driven by tissue-specific differential dependence network annotations and as validation, highlight how they have been confirmed in (larger) genetic studies for the same or closely-related traits and/or partially verified experimentally. My results are a first step towards my long-term goal of developing a general-purpose model that can learn to predict expression under any set of arbitrary constraints (e.g. for a tissue of interest, given a particular genotype, or after a fixed-time post-drug exposure) and demonstrate the utility of such strategies in decoding GWAS signals.

TABLE OF CONTENTS

Chapter 1: Brief Introduction

In this thesis, I introduce two new methods that address the hurdle of translating molecular discoveries (e.g. GWAS signals) into functional and clinical understanding of disease pathophysiology. The first approach is a "from-first-principles" framework: constructing a model that can learn to predict tissue-specific expression from promoter and enhancer sequences (as well as the consequences of individual variation) without a priori knowledge of relevant sequences or disease phenotypes. I subsequently apply model properties and model-derived scores to understand the physiological context of genetic discoveries. The second approach is a "top-down" framework: given GWAS signals, constructing coherent gene modules/cascades that can capture some of the physiological and tissue-specific functionality of the variants (i.e. enrichment). This framework is meant to operationalize GWAS signals in the context of interactions between gene modules/cascades. These two approaches, while sharing a common goal, are fundamentally different perspectives on the same problem. I briefly introduce both methods in this chapter and present more details in their respective chapters.

1.1 Expression-based High-throughput Assays

High-throughput assays (HTAs) are the cornerstone of modern drug discovery and a useful tool to translating the hundreds of genetic discoveries associated with human traits and disease into functional understanding[3,14-17]. All high-throughput assays can be described as empirical assessments of the activity of biological entities (e.g. genetic variation, DNA sequences, small molecules) by a standardized output, usually in the form of optically detectable labels (i.e. reporters), or more rarely, using (scalable) high-dimensional measurements (e.g. L1000[i], RNA-seq). Increasingly, with mounting evidence suggesting that the causal variants at genetic loci affect non-regulatory changes (rather than protein-coding alterations)[20], many of these HTAs – including those used in drug discovery[3] – are

[i] L1000 is an expression-profiling assay that directly measures the abundance of 978 transcripts, which is subsequently followed by the computational inference of the remaining transcriptome (22,000 genes). In addition, 80 "invariant" genes are also directly measured to enable scaling and normalization of the imputed values. For this thesis, I only use the 978 directly measured genes and discard all imputed profiles (which tend to be of variable quality).

focussed on the immediate impact of biological entities on the transcriptome. In fact, these expression-centred HTAs can be described generally as:

Eq. 1: biological entity (e.g. DNA sequence or small molecule) → measured expression output

Most, if not all, current research has focussed on the development of these experimental methods and the analytic techniques necessary to analyse the generated data[3,16,17]. Here, I introduce a general, modular *in silico* framework – **p**romoter-and-**e**nhancer-derived **a**bundance (**pea**Brain) framework – with the goal of recapitulating these expression-centred HTAs as computational models on graphical processing units[ii] (GPUs) to enable inexpensive and more scalable interrogations of the functional impact of diverse biological entities on expression. The convolutional neural network[iii] architecture (see **Box 1.1**) at the core of this framework leverages promoter sequences to emulate the transcriptional machinery of human cell lines/tissues, allowing us to (**Figure 1.1**, *on next page*): (a) re-create DNA sequence reporter assays, (b) re-create small molecule high-throughput screens, (c) re-create shRNA[iv] high-throughput screens; and (d) jointly model whole genome sequences to identify putatively functional eQTLs that are missed by experimental high-throughput assays – all using small extensions of the same architecture[v] and model hyperparameters[vi]. This work is introduced in greater detail in Chapter 2 and Chapter 3.

[ii] GPUs are specialized electronic circuits designed to allow users to rapidly manipulate and alter memory to accelerate (and facilitate) the creation and manipulation of very large matrices/arrays. Since images are often represented as arrays, GPUs are frequently used for image visualization and manipulation (hence the name).
[iii] Convolutional neural networks are a class of deep neural networks. Like any deep neural network, CNNs consist of input and output layers, as well as multiple hidden layers. The hidden layers of CNNs typically include convolutional layers (where units perform a cross-correlation on the input). See **Box 1.1** for details.
[iv] shRNA, or short hairpin RNA, are artificial RNA molecules that can be used to silence target gene expression via RNA interference (RNAi). RNAi is a broad term that refers to any biological process in which RNA molecules inhibit gene expression, and include naturally occurring molecules, such as miRNA (microRNA) and siRNA (small interfering RNAs). shRNA have become less popular with the introduction of CRISPR/Cas9 genome editing. However, nearly all large, publicly available high-throughput datasets of gene knockdowns use shRNA (rather than CRISPR/Cas9 editing).
[v] The architecture of the neural network, in simple terms, can be defined as the number of hidden layers, the number of units in each layer, and the connections between them.
[vi] In machine learning, hyperparameters are parameters (or variables) whose value is set before the learning process. Simple models (e.g. ordinary least squares linear regression) have no hyperparameters. In contrast, regularized linear models (e.g. LASSO or l1-regularization) have a hyperparameter, l1, that specifies the degree of regularization. Neural networks often have many more hyperparameters, including the variables that determine the network architecture (e.g. number of layers or the number of hidden units in a layer) and the variables that influence how the network learns (e.g. the learning rate). In all cases, hyperparameters are fixed before the training process; that is, they do not change after the model is exposed to the data.

Box 1.1 Convolutional Neural Networks

Convolutional layers are the core building blocks of convolutional neural networks (CNNs). The input can be any array (e.g. 1-dimensional in the case of DNA sequences or 2- dimensional in the case of images, as depicted below). In general terms, convolution is any mathematical operation on two functions that produces a third function that expresses how the shape of one function is modified by the other. Operationally, convolution is the process of adding each element in the (input) array to its local neighbours, but weighed by the kernel/filter that is learnt by the model during the training process. Image copied from Wikimedia commons.

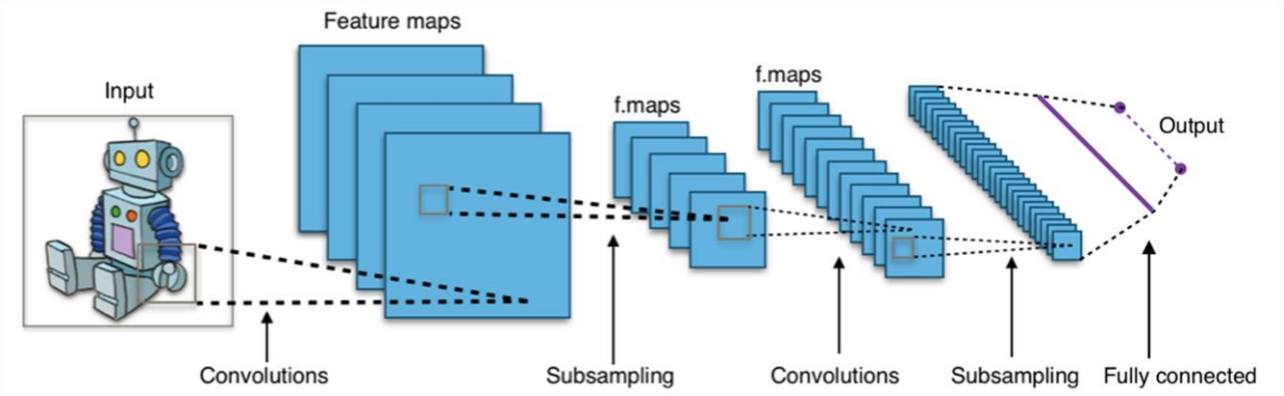
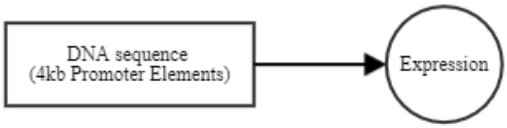
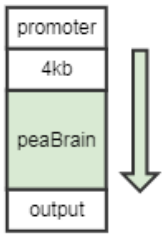


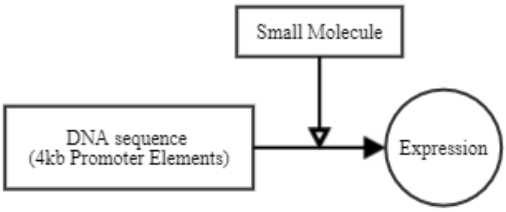
Figure 1.1 An overview of the biological architecture underlying the modular extensions and applications of the peaBrain framework. **(a)** The DNA sequence reporter assay is at the core of the peaBrain framework, which can be extended to incorporate **(b)** molecular fingerprints to model the impact of small molecules on expression or **(c)** shRNA oligo-sequences to model the impact of RNA interference on expression. **(d)** If I train the peaBrain model on larger sequences (centred on the promoter), I can jointly model the impact of genotypic variation (substitutions, insertions, and deletions) to predict individual expression and identify putatively functional eQTLs. *Abbreviations:* u/s, upstream; d/s, downstream.

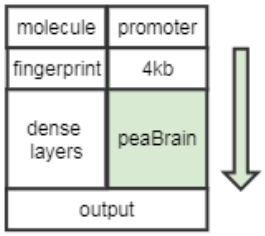
a Promoter Sequence Reporter Assay



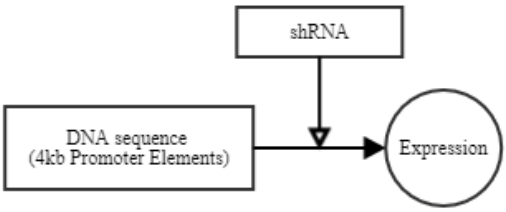


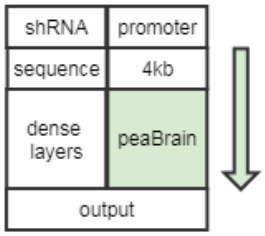
b Small Molecule Screen



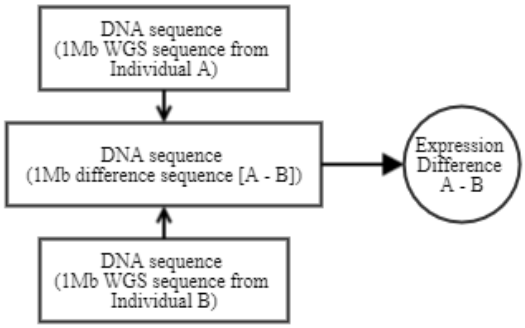


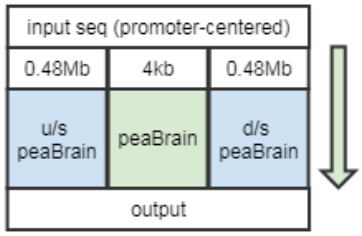
c shRNA Screen





d Functional eQTL Screening (Joint Model)





1.2 Differential dependence to identify functional mechanisms underlying GWAS signals

Translating much of the sequence variation associated with type 2 diabetes (T2D) and related traits into functional insight remains a bottleneck to mechanistic understanding of disease pathophysiology. If this genetic variation is overwhelmingly common[1] and influencing a limited subset of cascades, summarized genotype data should readily encode both a tissue-specific context and readily place variants in causal pathways. I sought to develop an analytic strategy for multiple T2D-related traits, using expression data that leverages these assumptions to: (a) decode the underlying casual tissue(s); (b) reveal tissue-specific mechanisms; and (c) discover new risk loci. This strategy is meant to parallel the aforementioned peaBrain approach and is independent of directly modelling promoter sequences.

At the heart of the analytic strategy described here, are expression-derived tissue-specific coexpression differential dependence networks (xDDNs). Briefly, xDDNs are abstract network representations of transcriptomic perturbations; generated using a Bayesian approach that detects whether modules of coexpressed genes are differentially dependent (i.e. pairs of gene modules that are dependent under one condition and independent under another condition). I used affinity propagation (which will be introduced in greater detail in chapter 4) to define gene module sets, in a manner more robust than other "state-of-the-art" coexpression methods, for two tissues (pancreatic islets and skeletal muscle) for which data was available. Here, I show variants associated with genes from these xDDNs are significantly enriched for GWAS hits from T2D and related traits; in other words, genetic variation associated with these traits can explain differential dependence in the disease state. Second, I demonstrate this enrichment is sensitive to the underlying physiological context (e.g., where only one tissue is the causal tissue, such as beta cell function) and reveal tissue-specific differences about the consequences of genetic variation (in traits where multiple tissues may contribute to the underlying phenotype, such as glycated haemoglobin). I subsequently contrast differential dependence network-derived annotations to information gleaned from histone markers and chromatin state data. Finally, I showcase the utility of these tissue-specific xDDNs to recover mechanistic information and discover multiple new loci-phenotype associations by constructing tissue-specific joint models.

Why T2D? With alarming rates of prevalence[21], understanding the pathophysiology underlying development and progression of T2D is essential for screening and developing effective, personalized treatment regimes. Recent genome-wide association studies (GWAS) have uncovered more than 100 genomic loci associated with T2D[22-24]; many of which are also implicated in fasting glucose, fasting insulin levels, and more generally, beta cell function. However, the mechanisms by which these genetic variants act remains unclear as most variants lie in noncoding regions of the genome[25]. By narrowing the focus on T2D and closely related traits, I can begin to define a framework to assess and compare the biological utility and clinical relevance of many of existing and new approaches used to gain insight from molecular and clinical data types.

Chapter 2: peaBrain DNA-reporter assay

Most reported disease-associated variation for complex traits lies in non-coding regions of the genome[20]. Despite advances in discovery and annotations of functional non-coding elements across the genome[10,11,13,19], 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[8,26-30], distinct from both colocalization-based approaches that depend on the availability of genome-wide genetic associations[31-33] and annotation-/frequency-driven prioritization of "functional" variants[34-37]. 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[28], Basset[29], and Bassenji[30]; and (b) methods that directly predict RNA abundance from genotype or sequence data, such as PrediXcan[26] and TWAS[8]. Methods in the former category poorly capture differences in transcript expression as a result of genotypic variation[28-30] and are relatively poor predictors of alterations in the histone code[28]; methods in the latter category are not able to identify which of the variants detected within an eQTL association locus are functional[8,26]. To address these concerns, I developed the **p**romoter-and-**e**nhancer-derived **a**bundance (**pea**Brain) model, which consolidates both of these approaches.

2.1 peaBrain DNA-reporter assay captures >50% of the variance in mean gene expression with promoter sequences

At the core of the peaBrain framework, the DNA-reporter assay is designed to predict tissue-specific average expression by leveraging core promoter elements (from the human genome[vii]; **Chapter 1, Figure 1.1a**)[4]. For each gene, as input, I generated a 1-dimensional (1D) matrix centred on the region

[vii] I used the human reference genome. However, any human genome could be used to train the model (provided that there are no major chromosomal abnormalities [e.g. aneuploidies] or large deletions/insertions). This is consistent with the goal of the DNA-reporter assay, which is to train a peaBrain model that can predict the differences between genes in the same tissue. To ensure generalizability, the sole condition is that the promoter of any gene is largely similar between any two individuals (which is demonstrably true; humans differ roughly 1 in every 1000 basepairs across the genome). This would not be necessarily true for other species (e.g. *Drosophila* where there are ~1 variant every 300 basepairs).

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 **Figure 2.1**, *on next page*). I used one-hot encoding[viii] (four channels) to represent the four DNA letters (A, T, C, G) in the reference genome (4 channels). The model output was the corresponding predicted mean RNA abundance of that gene, after rank-transformation to normality (see **Box 2.1: Methods - GTEx Data and Promoter Definitions** for details, *on next page*).

It is important to note that this 4 kbps interval (+/– 2kbps of annotated TSS) can be defined for genes smaller than this window (and for genes that are much longer)[ix]. The crux of this definition is to narrow the search space for the peaBrain algorithm to identify useful features that are predictive of expression. Regardless of the window defined, peaBrain will search the entire space; this underpins the rationale of using convolutional and pooling layers for following the input layer of the model as we discuss in depth below. A longer window will contain more information, but is more computationally expensive. Furthermore, there is a diminishing return-on-investment with larger windows, as we observe with the plateauing model performances. These results suggest that promoter (4kbps window) contains all the necessary information for predictive performance sufficient for the tasks highlight in this thesis.

[viii] One-hot encoding is a technique commonly used in machine learning to encode categorical variables with each channel indicating the presence (1) or absence (0) of the corresponding DNA letter.
[ix] By defining a fixed window, we simplify the problem; another possibility that we did not investigate is defining tissue-specific gene-specific windows, but that would be prone to bias and overfitting.

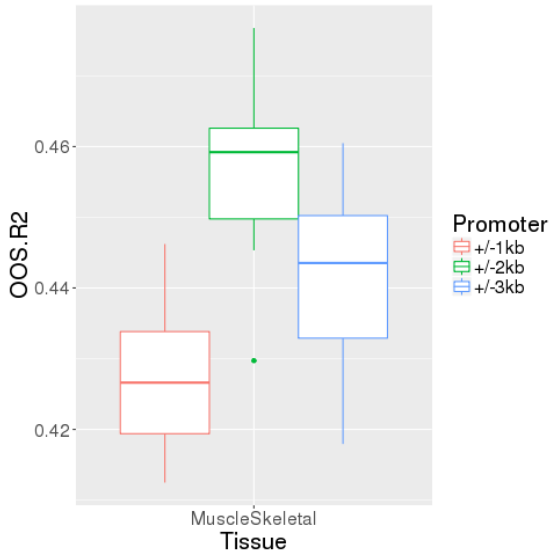


Figure 2.1 Using the peaBrain model for skeletal muscle (largest tissue by sample count in GTEx), the 4kbps promoter sequence (+/– 2kbps of annotated TSS) outperforms both 2kbps (+/– 1kbps) and 6kbps (+/– 3kbps) promoter sequences in predicting average gene expression. The results were consistent for all other tissues; skeletal muscle was selected for visualization, as it was tissue with the most samples. (Plots for all other tissues are available on the peaBrain website; see **URLs**). Boxplots from left-to-right: +/- 1kb, +/- 2kb, and +/- 3kb.

I applied this framework to all tissues from the GTEx dataset[1], 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 and tissue-specific annotations (class-C). For class-B models, I 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[7]; see **Box 2.1: Methods - GTEx Data and Promoter Definitions** for details). For class-C models, I 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[18].

Box 2.1 Methods – GTEx Data and Promoter Definitions

GTEx RPKM data, for training and testing the DNA reporter assay, was downloaded from GTEx (v7; see **URLs**)[1]. To prepare the data, 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[4].

I 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. I accounted for gene stranded-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. I further limited my analysis to protein-coding genes (n = 19,820 genes) and to autosomal chromosomes for simplicity.

For all peaBrain DNA-reporter assay 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. Processed genomic annotations and epigenetic markers were obtained from the LDSC[7] (see **URLs**) and similarly processed. For class B models and using the LDSC annotations, I 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[9], CTCF sites, DGF peaks[10], DHS peaks[11], enhancers[12,13], fetal DHS peaks[11], H3K27ac peaks[18,19], H3K4me1 peaks[11], H3K3me3 peaks[11], H3K9ac peaks[11], introns[4], promoters[4], promoter flanking sequences[13], repressed sites[13], super enhancers[19], transcription factor binding sites (TFBS)[10], transcribed sequences[13], TSS[13], untranslated 3' regions (UTR3)[4], untranslated 3' regions (UTR5)[4], and weak enhancers[13]. 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 the main text, for a subset of peaBrain DNA reporter assay 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.

The peaBrain DNA reporter assay model, itself, was constructed using Theano 0.9.0[x] and Lasagne 0.1[xi]. For a single tissue, peaBrain DNA reporter assay model takes in the core promoter sequence as input and predicts the normalised mean abundance of the corresponding gene (**Chapter 1, Figure 1.1a**). The input sequence is a 1D vector with four channels encoding the DNA sequence and when appropriate, additional channels as binary representations of various genomic annotations and epigenetic markers (described above). The peaBrain model is a series of 1D convolutions and max pooling[xii] layers (**Appendix I - Supplementary Table 1[xiii]**). In practice, a 1D convolution is implemented as a 2D convolution with width set to one (effectively dropping the unused dimension). Each convolutional layer was set with 11 filters of size 5 and a leaky rectify non-linearity activation function[xiv]. These values were determined empirically. Where appropriate, I highlight critical experiments used in their derivation (e.g. the 4kb core promoter selection). However, due to length constraints, many of the experiments are only available online or in the pre-print (see **URLs**). The leaky rectify activation function for all convolutional layers has a nonzero gradient for negative input, which is useful for convergence[38]:

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

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,

[x] Theano is a Python library developed to optimize and efficiently evaluate mathematical expressions involving multi-dimensional arrays efficiently.
[xi] Lasagne is a Python library to build and train neural networks in Theano.
[xii] Pooling is a down-sampling function that progressively reduces the spatial size of the input (i.e. "combining" inputs over an axis). This pooling reduces the amount of parameters in the model, and hence also controls/limits overfitting. Max pooling is the most common, but there are several other functions that could be used (e.g. mean/average pooling)
[xiii] Appendix I –Table 1 details the model architecture (all hyperparameters). This table is at the very end of this document (alongside the model architectures for all other work).
[xiv] These model hyperparameters were determined empirically through experiments. Some of these experiments are discussed here (e.g. selection of the 4kb core promoter sequence) when they are particularly relevant; however many are not because of length constraints for the thesis. Given the broadness of the hyperparameters space, I recognize that all hyperparameters selection - even when backed with thorough empirical work - may have some arbitrary components. The work described here (and the extensive validation) is meant to highlight that this given selection of hyperparameters is useful for the tasks described in the thesis.

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 I extract the continuous vector gene representations, discussed below), I 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 [39,40]. 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[xv]; the value was empirically chosen and must be sufficiently large enough for useful embeddings (see discussion below). 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 the Adam algorithm (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[41]; 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 set). (All code, with necessary data, has been publicly released and is available on the peaBrain website [see **URLs**]). To assess Stage 1 model performance, I used 10-fold cross-validation schema; 10% of genes (n = 2000) were with-held for testing during each fold. I assessed sequence similarity of promoters to ensure that testing set was independent of the training and validation sets.

[xv] The size of the penultimate layer was designed with the generation of "gene embeddings" in mind, which is discussed later in this chapter. While the number itself is arbitrary, experiments have shown that it must be large enough for the embeddings to be useful (minimum number of neurons in the layer is at least 2 orders of magnitude). A smaller penultimate layer may be sufficient (and preferred) if the task is only prediction. In fact, model performance is largely independent of the size of the layer (provided that it is not too small [e.g. <10 neurons]).

When implemented, I observed that DNA-only (class-A) peaBrain 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][xvi] 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 2.2**, *on next page*). (I average the oos-r^2 across all 10-folds within a tissue and use the median across all tissues to assess global performance.) 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 2.2**, *below*).

[xvi] 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, 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]. peaBrain DNA reporter assay performance was assessed using 10-fold cross validation (10% of genes were withheld from the algorithm during training).

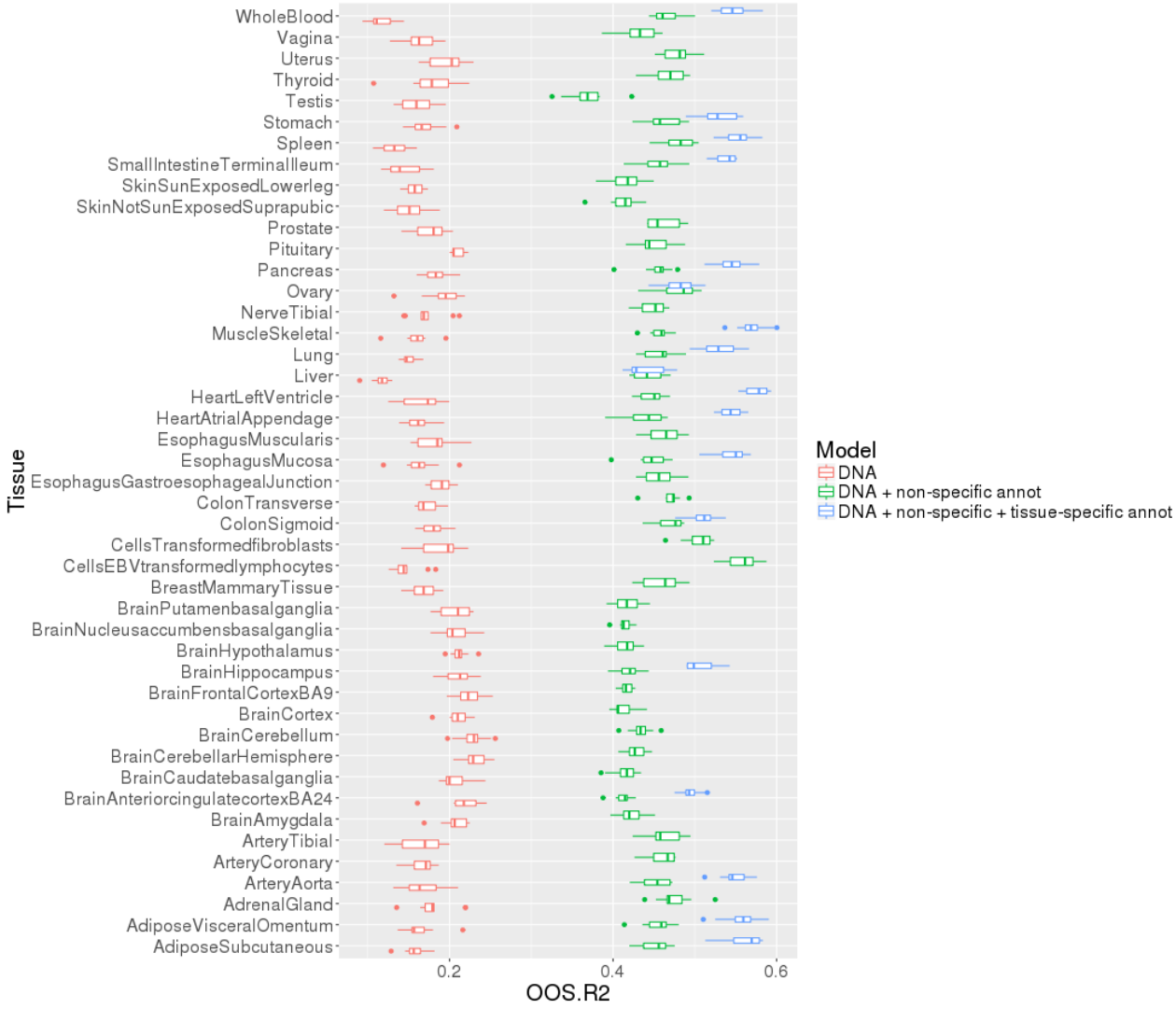


Figure 2.2 Incorporating genomic and epigenetic annotations improves the performance of peaBrain to predict the normalized mean abundance across all GTEx. The 4kbps promoter sequence, when annotated with tissue specific annotations, is sufficient to predict the majority of variance in mean expression 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; see text below). **Abbreviations:** OOS.R2, out-of-sample r^2.

Regularized linear models are poor models for predicting mean abundance from core promoter sequences (10-fold cross-validated average oos-r^2 < 0). To assess how simple linear models compare to "deep learning" models, I fitted linear models by minimizing a regularized empirical squared loss, with a L2 (squared Euclidean norm) penalty, using stochastic gradient descent from python's sci-kit learn (sklearn)[xvii]. The input of the class B models (non-specific annotated DNA model; 1D matrix with

[xvii] Scikit-learn is a Python library for machine learning.

32 channels) was flattened prior to modelling, i.e. the mean abundance each gene was modelled as a linear function of 128,000 variables (most were binary variables). The model was fit using the partial fit function from SGDRegressor with sklearn using default parameters; the same exit criteria as described above for the peaBrain models were applied. I noted that 10-fold cross-validated average oos-r^2 was consistently below zero[xviii], with performance considerably worse than the peaBrain model. The gap in performance was considerable (precludes visualization[xix]). I also repeated this analysis using a single dense neural network layer with linear activations, and noted that the 10-fold cross-validated average oos-r^2 was consistently below zero; a single dense layer with linear activations is equivalent to a simple linear model.

Fully connected neural networks are slower and more memory-intensive, with slightly worse performance than convolutional neural networks; non-linear activations for convolutional layers improve performance over linear activations. I compared fully connected dense neural networks to the convolutional neural networks at the heart of peaBrain (**Figure 2.3**, *on next page*) for class B skeletal muscle model. I replaced each pair of convolutional-pooling layers with a single dense layer; the number of neurons in the three dense layers was limited only by GPU memory. The penultimate layer and single output neuron were kept consistent between the models. Using the skeletal muscle Class B input, I noted that fully connected neural networks performed slightly worse (oos-r^2 = 0.42) than the convolutional neural networks (oos-r^2 = 0.46). Increasing the number of layers allows fully connected neural network to reach performance parity with CNNs, at the cost of increased memory and computational cost (not feasible for any of the peaBrain extensions I discuss in later chapters). I subsequently wanted to assess the importance of the non-linear activation for the convolutional layers of peaBrain. I constructed an identical model, but replacing all CNN activations with linear functions and noted that performance for this model was consistently worse (oos-r^2 = 0.43) than the classical

[xviii] Oos-r^2 compares the fit of the chosen model with that of the null hypothesis (a horizontal straight line, which is the mean of the data). If the fitted model performs worse than a horizontal line, then r^2 can be negative.
[xix] By "precludes visualization", I mean that showing both sets of results on the same plot is simply not informative. The gap in performance is too large for any meaningful interpretation or understanding of the results via boxplot. However, the text includes the range for the results, which are sufficient.

peaBrain architecture (oos-r^2 = 0.46). It is important to note that this is not an exhaustive search of the ideal set of parameters, but an exploratory analysis to begin to understand peaBrain's performance.

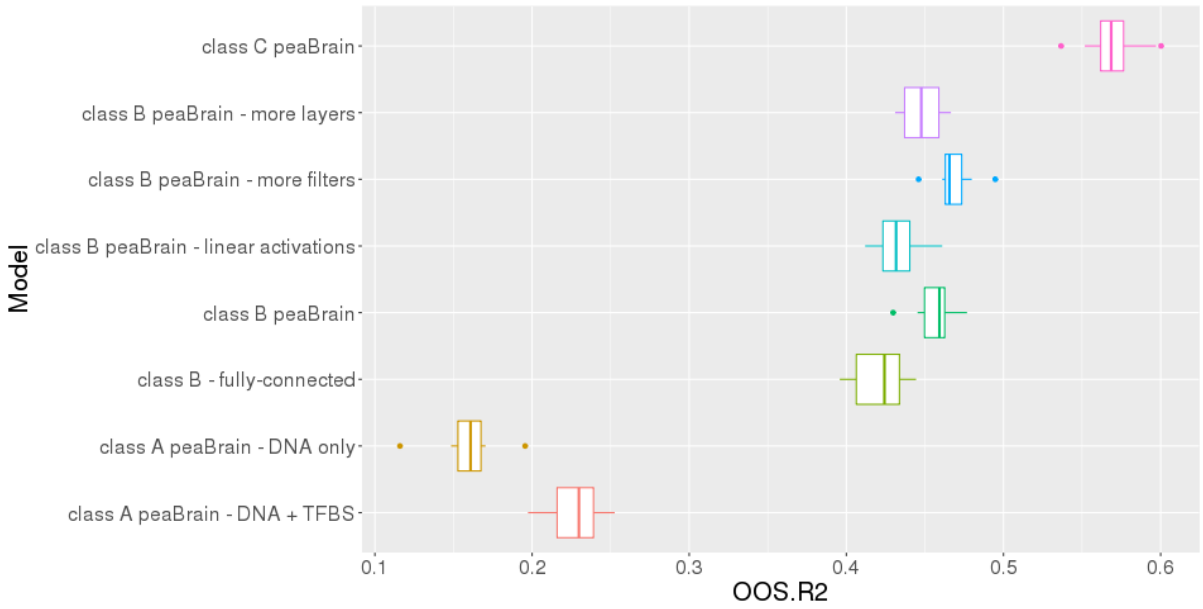


Figure 2.3 Boxplots of 10-fold cross-validated oos-r^2, as assessed in skeletal muscle[xx], for class A models (labelled as "class A peaBrain – DNA only"), class A with TFBS annotations (labelled as "class A peaBrain – DNA+TFBS"), class B models with tissue-agnostic annotations ("class B peaBrain – CNNs"), fully connected neural networks ("class B – fully-connected"), class B models with linear activation functions ("class B peaBrain – linear activations")[xxi], class B models with increased number of layers ("class B peaBrain – more layers"), class B models with increased number of filters ("class B peaBrain – more filters"), and class C models with tissue-specific annotations ("class C peaBrain"). (All of these models are sub-selections of variations on the peaBrain architecture and data input – none of these models are linear models!)

DNA sequences, annotated with experimentally-derived TFBS, from core promoter sequences are insufficient to predict mean abundance with high accuracy – epigenetic/histone markers contain the bulk of the information and are not readily accessible from the DNA sequence alone. I was interested in determining the contribution of epigenetic/histone makers, alongside more general genomic annotations (such as coding sequences), in predicting the mean abundance of genes. In

[xx] As aforementioned, due to word limit constraints, I will often highlight the results for skeletal muscle only. Similar results have been generated for other tissues and are available on the website and/or accompanying preprint.

[xxi] To clarify, linear activations does not mean the model was a linear model. In this case, the activation function for the neurons, after the convolutional layers, was linear.

particular, I wanted to explore whether the DNA sequence alone was sufficient to predict expression in skeletal muscle. I noted that increasing the number of convolutional layers or the number of filters did not improve model performance (**Figure 2.3**, *above*). Explicitly incorporating TFBS into the model (i.e. annotating the DNA only and explicitly with TFBS) only improved performance slightly (oos-r^2 = 23%), and was still considerably worse than the full class B model with epigenetic/histone marker annotations (oos-r^2 = 46%; **Figure 2.3**, *above*). (Class-A DNA-only models had an average oos-r^2 of 16% for skeletal muscle; class-C models annotated with tissue-specific information had an average oos-r^2 of 57%.) The TFBS were collected from the Gene Transcription Regulation Database (GTRD) v17.4 with data on 476 human transcription factors and included peak calling with four different software (MACS, SISSRs, GEM, and PICS). In addition to including the processed peak calls, I also incorporated clusters (i.e. peaks merged for the same transcription factor but under different experimental conditions) and meta-clusters (i.e. non-redundant peaks synthesized from all four methods). This absence of improvement suggests that peaBrain model already recognizes many of the TFBS; identified by the convolutional filters inherent to the model architecture. These results indicate that experimentally-derived epigenetic and genomic annotations add information to that contained in the DNA sequence alone. As described above, this is broadly consistent with the observation that other convolutional neural networks models like DeepSEA are better at predicting TFBS (median AUC = 0.958) than at predicting histone modifications (median AUC = 0.856)[28].

Testing on expression data from microarray platforms and using orthologous promoters in non-human Hominid primates further validates performance of the peaBrain DNA reporter assay. I further assessed performance of the peaBrain DNA-promoter reporter assay on external datasets in three human tissues (expression measured by microarray[xxii]; Spearman's rho[xxiii] between predicted and external platform = 51%-63%; **Figure 2.4**, *on next page*) and on its ability to predict expression of

[xxii] Why microarrays? One useful metric in assessing generalizability of any model is its ability to predict on gene expression measured by other technology platforms (i.e. other than RNA-seq in this case). This is a useful benchmark that indicates the model is not inherently biased towards the underlying technology from which the training data was derived.

[xxiii] Spearman's rho is a non-parametric test used to measure the strength of correlation between two variables. It is equivalent to Pearson's correlation on the ranks of the data.

orthologous promoter sequences in four non-human Hominid primates[xxiv] in two different tissues (rho = 5%-22%). For the non-human Hominid primates, I used the promoter sequences from their respective genomes and the human-trained peaBrain models. In other words, the human-trained peaBrain models were able to capture a large portion of the variance in expression, which suggests that there is conservation of transcriptional cascades across the evolutionary clade[xxv]. (Spearman's rho was used because peaBrain was trained using RNAseq and the test data sets were generated on microarray platforms and/or different species. This means that the magnitude of expression for any given gene is different; however, the ranks of the genes would be expected to be largely consistent.) All together, 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 experimental evidence[43]. As I noted above, 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%). 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 I 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. Importantly, this suggests that expression-based HTAs (e.g. small molecule screens) can be recapitulated by extending this DNA reporter assay; that is, the information regarding the transcriptional

[xxiv] Why other species? I was interested in coarsely analysing the extent to which tissue-specific expression (and transcription factor cascades) are preserved across evolutionary clades.
[xxv] This could be used to model the evolution of promoters (and expression of genes) through the primate lineage!

machinery of any given tissue or cell type is encoded in and can be extracted from the promoter sequence (and the corresponding gene expression).

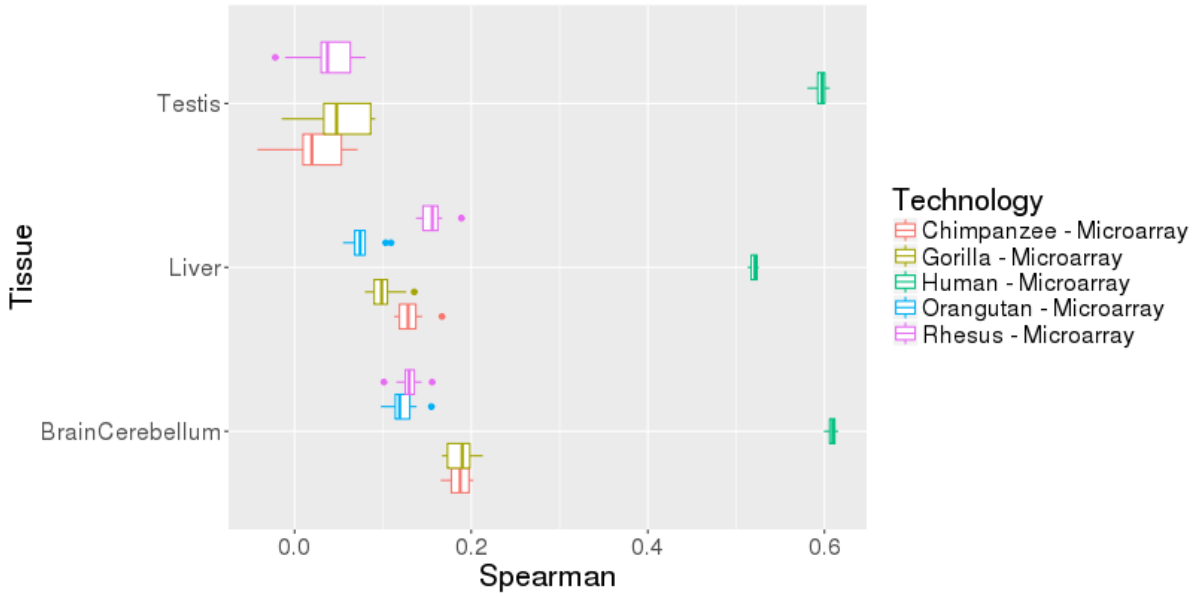


Figure 2.4 Boxplots of the cross-validated peaBrain models when trained on human GTEx RNAseq data and tested on microarray expression data from the Hominid lineage (including humans). The performance on non-human hominid species suggests that the peaBrain model – when trained exclusively on human – can also shed insight the transcriptional machinery of related species (both extant and extinct). This could be useful, for instance, when tracing the evolution of regulatory cascades in the Hominid lineage.

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

Having demonstrated the predictive ability of the peaBrain DNA reporter assay, I was 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. I 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[xxvi]. To facilitate comparison across tissues, I performed this analysis using the class-B DNA reporter assay, 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[4] and represented by phylogenetic p-values (phyloP) (see **Box 2.2 Methods – Tissue-specific peaBrain impact scores**, *two pages after next*). Briefly, these phyloP nucleotide conservation scores are based on an alignment and a model of neutral evolution[4]: a more positive value indicates conservation or slower evolution than expected, with the magnitude of the phyloP score corresponding to the negative log p-values under the null hypothesis (i.e. neutral evolution). For every unit of absolute magnitude increase in peaBrain impact score (normalized expression of gene), I observed an average increase of 8.95 in phyloP scores, indicating increased conservation (i.e. 8.95 order-of-magnitude difference in the negative log10 p-value; **Table 2.1**). Equivalently, for every unit increase in phyloP, I 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 was a single exception (the nucleus accumbens [basal ganglia])**,** where noncoding transcriptomic impact was correlated with accelerated evolution (**Table 2.1**).

Table 2.1 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 ($p < 10^{-16}$). **Abbreviations:** L, lower; U, upper.

[xxvi] Why not use real variants (with reference/alternate alleles) instead of zeroing out each site? The combinatorial number of variants makes this approach not amenable to "calculating" an impact score for each position in the promoter. I wanted a position-specific metric, rather than a variant-specific metric. I calculate the impact of individual level variation, using a slightly modified architecture, in chapter 3.

	Linear Model Coefficient	L Bound	U Bound
AdiposeSubcutaneous	15.79	15.72	15.86
AdiposeVisceralOmentum	14.32	14.25	14.39
AdrenalGland	0.33	0.27	0.39
ArteryAorta	3.51	3.47	3.56
ArteryCoronary	5.64	5.59	5.69
ArteryTibial	5.89	5.84	5.94
BrainAmygdala	9.25	9.19	9.31
BrainAnteriorcingulatecortexBA24	6.92	6.85	7.00
BrainCaudatebasalganglia	6.19	6.15	6.23
BrainCerebellarHemisphere	13.48	13.41	13.55
BrainCerebellum	4.78	4.73	4.83
BrainCortex	2.70	2.67	2.74
BrainFrontalCortexBA9	8.57	8.50	8.64
BrainHippocampus	5.09	5.03	5.14
BrainHypothalamus	5.17	5.09	5.24
BrainNucleusaccumbensbasalganglia	**-1.32**	**-1.37**	**-1.27**
BrainPutamenbasalganglia	6.91	6.86	6.96
BreastMammaryTissue	4.24	4.19	4.30
CellsEBVtransformedlymphocytes	5.92	5.87	5.98
CellsTransformedfibroblasts	9.17	9.13	9.22
ColonSigmoid	7.28	7.23	7.34
ColonTransverse	3.64	3.60	3.69
EsophagusGastroesophagealJunction	7.10	7.04	7.16
EsophagusMucosa	6.37	6.30	6.44
EsophagusMuscularis	8.11	8.03	8.18
HeartAtrialAppendage	11.98	11.90	12.07
HeartLeftVentricle	7.29	7.21	7.36
Liver	2.99	2.94	3.04
Lung	7.93	7.86	7.99
MuscleSkeletal	4.27	4.23	4.31
NerveTibial	6.70	6.63	6.77
Ovary	5.39	5.33	5.45
Pancreas	11.07	11.00	11.15
Pituitary	4.67	4.62	4.71
Prostate	13.57	13.51	13.64
SkinNotSunExposedSuprapubic	7.43	7.37	7.48
SkinSunExposedLowerleg	4.06	4.00	4.12
SmallIntestineTerminalIleum	2.87	2.82	2.92
Spleen	3.70	3.63	3.76
Stomach	1.55	1.51	1.58
Testis	7.36	7.31	7.40
Thyroid	5.87	5.80	5.93
Uterus	6.54	6.50	6.58
Vagina	7.34	7.27	7.41
WholeBlood	9.56	9.50	9.62

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[34] and Eigen[35]) 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, I was interested in assessing the utility of peaBrain-derived scores to interrogate disease-associated variants. I compared the performance of the non-tissue-specific peaBrain score (see **Box 2.2 Methods – Tissue-specific peaBrain impact scores**, *next page*) to two other non-coding metrics (CADD[xxvii,34] and Eigen[xxviii,35]) across a series of tasks (**tasks A-C**; all tasks are summarized in **Table 2.2**).

[xxvii] CADD is a single meta-score derived from analysis of multiple annotations for variants that survived natural selection, compared to simulated mutations[29].
[xxviii] Eigen is an unsupervised score that synthesizes a combination of functional annotations into one meta-score[30].

Table 2.2 Tabulated summary of tasks used to assess peaBrain performance in Chapter 2.2.

Task	Description	Methods (dataset)	Results
A	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 2.3
B	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 2.3
C	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 2.3 Figure 2.5

Box 2.2 Methods – Tissue-specific peaBrain impact scores

The 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[4]. 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**). 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, I counted the number of overlapping somatic mutations. I further limited my 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. (The binary predictive tasks were to identify positions with non-zero incidence of cancer-associated somatic mutations and to identify positions with recurrent cancer-association somatic mutations among all positions with at least one somatic mutation.) 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). 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**).

First, I made use of data on disease-related variation from the Catalogue of Somatic Mutations in Cancer [COSMIC][44], limited to the census gene set which defines a set of genes with somatic mutations causally implicated in human cancer (see **Box 2.2 Methods – Tissue-specific peaBrain impact scores**, *below*)**.** In **task A,** I 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,** I 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)[xxix]. 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 **Box 2.2 Methods – Tissue-specific peaBrain impact scores**, *above*). I subsequently assessed the significance of the logistic model coefficients for each of the non-coding metrics across the two tasks (**Table 2.3**). 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 2.3**). Significance was assessed using the default two-tailed p-value corresponding to the z ratio based on the Normal reference distribution (**Table 2.3**; see **Box 2.2 Methods – Tissue-specific peaBrain impact scores**, *above*)[xxx]. 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

[xxix] For the analysis of recurrent somatic mutations, I was 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).

[xxx] I perform only three tasks and thus calculate only three p-values. Furthermore, the tasks are correlated with each other. Thus, I do not have to correct the p-values or adjust for the false discovery rate.

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 2.3**). 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, I was interested in investigating the discriminative ability of peaBrain-derived scores in identifying allele-specific transcription factor binding sites (**task C**). I hypothesized that allele-specific binding will have downstream transcriptional consequences that could be identified from directly modelling gene expression. As with **tasks A and B** (*described above*), I compared the performance of the non-tissue-specific peaBrain score to predictions by CADD and EIGEN in predicting allele-specific binding, after accounting for allele frequency and evolutionary conservation. I assessed performance of the three non-coding metrics across 6675 sites in core promoter regions after filtering for duplicate sites[45]; 1896 of which exhibited allele-specific binding at an unadjusted binomial $p < 0.05$. I noted that only peaBrain impact score was significantly predictive of allele-specific binding sites (coefficient = 35.38 [12.00, 58.67]; p = 0.003; see **Table 2.3**, *on next page*); relaxing the binomial p-value threshold (i.e. increasing the number of sites considered as allele-specific) brings the other non-coding metrics to significance. peaBrain's discriminative ability to identify allele-specific binding sites is consistent with my earlier observation that explicitly adding TFBS annotations did not improve the model. Notably, peaBrain's ability indicates that average expression of all genes in a single tissue and the reference genome is sufficient to learn both TFBS and allele-specific binding.

Table 2.3 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 **Main Text**). peaBrain is the only non-coding metric with significant coefficients for all three tasks; I 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. peaBrain's impact score has the same "units" as normalized expression. Eigen and CADD are on arbitrarily-normalized scales (hence the difference in coefficient magnitudes). **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

2.3 peaBrain Applications

In this section, I briefly highlight a subset of applications of the aforementioned peaBrain models. The results contained herein are proof-of-principles (rather than comprehensive assessments). Because of the broad utility and flexibility of peaBrain, there is no method that can perform all the same tasks as peaBrain. Instead, where appropriate, I select the best peer-reviewed methods available for each task as the baseline comparator.

Application 1: Tissue-specific peaBrain scores can identify the functional tissues underlying GWAS signals from complex traits. For **tasks A-C**, I 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, I sought to investigate advantages of tissue-specific impact scores. In particular, I wanted to highlight how tissue-specific scores could allow us to identify functional tissues associated with GWAS signal from complex traits. I 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, I 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 I can detect with tissue-specific peaBrain scores. This approach is independent of eQTLs. In particular, since genes that matter for disease (i.e. genes for which a perturbation will increase risk of a disease phenotype) are less tolerant of coding variants and of expression changes, they are less likely to be influenced by cis-eQTLs. However, the peaBrain impact score is not a cis-QTL discovery score; rather it is a metric that quantifies the "importance" of each position in the core promoter with higher peaBrain impact score meaning less variation and more conservations, as shown in Tasks A-C above). The key element is that peaBrain scores can be calculated for positions, even if there is no (common) variation present at that position.

I selected 4 quantitative traits (total cholesterol[46], LDL[46], HDL[46], and triglycerides[46]) for which the (primary) putatively causal tissue is well-established and included in the GTEx dataset[xxxi]. Using HESS[47], I 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)[47]. For each block (or "locus"), I 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. I 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

[xxxi] For the analysis of causal tissues, I 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.

assessed significance for the linear model coefficient (n = 45 tests for each GTEx tissue per phenotype). As a baseline benchmark, I compared my results to tissue predictions made using the tissue trait concordance (RTC) score[xxxii,48], 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. (RTC framework was not an ad-hoc selection; it is the only method at the time of writing that was tested and trained on the GTEx dataset for the traits included in this analysis.) Across all tested traits, the peaBrain framework was better at identifying putatively causal tissues than simply using the RTC-/eQTL-based method (**Table 2.4**). 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 **Table 2.4**. 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; **Table 2.4**). 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, using the peaBrain scores 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).

[xxxii] Tissue trait concordance (RTC) score is one of few methods tested and trained on GTEx eQTL data to predict the tissue-causality profiles for the traits tested here. It is included as a baseline for the peaBrain comparison. There are other approaches to inferring tissue-specific function for GWAS variants (e.g. co-localization approaches and associated fine-mapping techniques). These approaches are, however, beyond the scope of this discussion as they don't directly involve modelling tissue-specific expression.

Table 2.4 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. Nominal p-values are shown for the RTC (eQTL)-methodology; obtained from Supplementary Table 8 of the corresponding manuscript[48]. 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}

Application 2: peaBrain score out-performs existing measures in predicting allele-specific transcription factor binding. To further investigate peaBrain's ability to identify allele-specific binding sites[xxxiii], I compared peaBrain impact scores to predictions by methods specifically designed to predict TFBS, including two neural-network methods (DeepBind[49] and DeepSEA[28]), two kmer-based variant scoring methods (gkmSVM[50] and GERV[51]), and three position-weighted matrices (PWM)-related methods[45]. These methods depend on modelling TF ChIP-seq data in various ways and may have multiple models for the same TF. The predictive ability of these methods to identify allele-specific binding sites had been demonstrated by the respective authors. I noted that peaBrain scores positively correlated only with GERV measures, a kmer-based variant scoring algorithm (**Figure 2.5**). Unlike the other methods, peaBrain (and GERV) do not assume the existence of canonical motifs and learn TFBS by modelling sequences (or kmers) directly (i.e. not simply by modelling the absence or presence of a ChIP-seq peak). In contrast, for both DeepBind and DeepSEA, I noted positive correlation with at least one PWM-method. These methods generally assume the existence of canonical TF binding sites and predictions are based on the extent of perturbation of those motifs. While this comparison is limited to variants for which data was available, the peaBrain results suggest that explicitly characterizing TF motifs is not necessary to understand the consequences of sequence variation on TF binding and transcriptional dysregulation.

Figure 2.5 Rank correlation plot for TF-binding algorithms and the peaBrain impact score. JASPAR, MEME_1 and MEME_2 are PWM-approaches.

[xxxiii] Allele-specific binding site data and prediction scores for TF binding prediction algorithms were downloaded from the Supplementary Table appended to Wagih *et al.* (see **URLs**). For the comparison between non-coding impact metrics, duplicate sites were filtered (selecting the one with lowest nominal p-value).

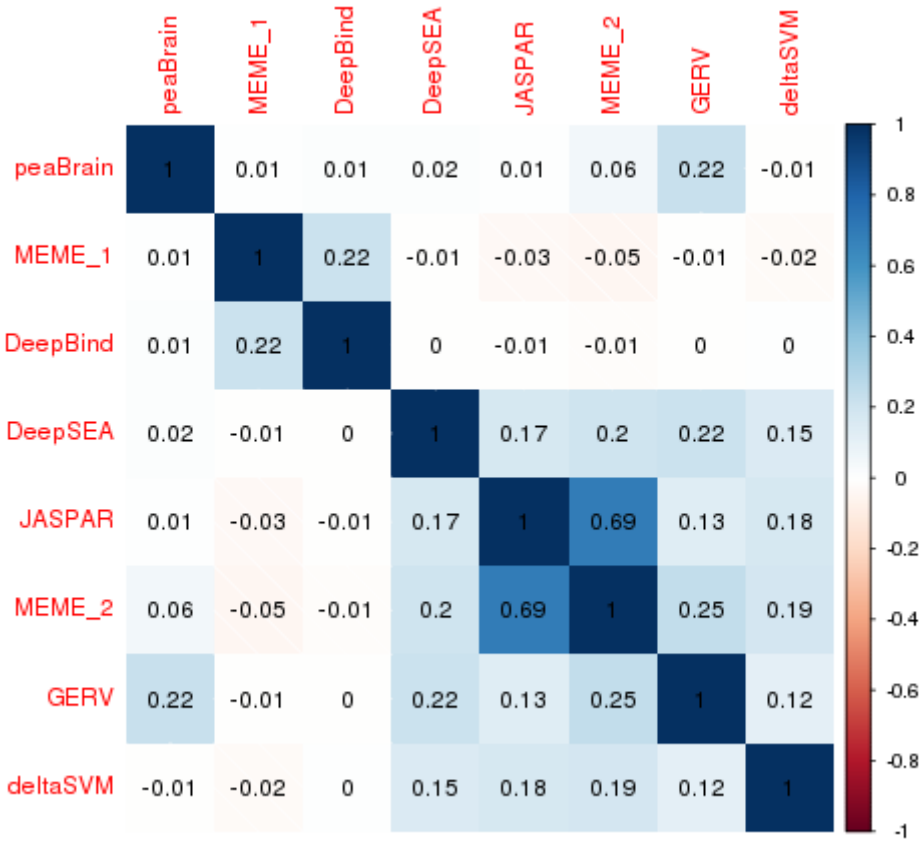


Figure 2.5 Rank correlation plot for TF-binding algorithms and the peaBrain impact score. JASPAR, MEME_1 and MEME_2 are PWM-approaches.

Application 3: Neural activations of penultimate layer of peaBrain model can be used to construct an embedding from the genes that encodes correlation information

Having demonstrated the predictive ability of the peaBrain model, I was subsequently interested in using the activations from the penultimate layer of the model as a continuous (and compressed) representation of the genes[xxxiv]. These neural activations capture both the annotated DNA (input) and its additive contributions to tissue- and phenotype-specific abundance (output) in a compressed form

[xxxiv] Embeddings are dense low-dimensional spaces into which you can map high-dimensional vectors. Embeddings are often used in natural language processing, translating words (often represented by high-dimensional one-hot encoded vectors) into a computer representation (that is, relatively, low dimensional). Ideally, the goal of any embeddings to capture some of the relationships of the high dimensional vectors (e.g. semantically similar words are close together in an embedding space). Embeddings can be learnt and subsequently re-used across models and in different applications. Here gene embeddings are derived from the peaBrain model trained to predict expression and I show that co-regulated genes have a high cosine similarity.

amenable to downstream analyses. Furthermore, as these vectors were obtained from a regression model, they readily capture only the salient portions of DNA abundance encoded in the annotated-genome (the weights of the model corresponding to the transcription factor that regulate and interact with this genome). Because of model choice, the mean abundance of each gene was encoded as a linear combination of the vector elements, *i.e.* the output of the regression model. As with my earlier analyses, I limited my analysis of the properties of the embeddings to class B models for skeletal muscle. I observed, for the skeletal muscle embeddings, that pairwise cosine similarity between these dense gene representations corresponded to the measured RNAseq correlation between the gene pair. After excluding self-correlations and weakly correlated genes (RNAseq rho < 0.5), I noted that the cosine similarity[xxxv] of the embedding was significantly correlated (Spearman's rho = 0.18; $p < 2.2 \times 10^{-16}$) to the experimentally RNAseq-derived correlation. This suggests the annotated-DNA model, without supervision, imposes a linear structure on this vector space: the angle between the vectors corresponds to the co-regulation of the gene pair.

Application 4: peaBrain DNA reporter model can predict the transcriptomic consequences of epigenetic modifications along the neural differentiation trajectory[xxxvi]

I was subsequently interested in exploring the utility of the model architectures for datasets other than GTEx. In particular, I sought to quantify the utility in understanding the consequences of epigenetic modifications that underlie the key cell fate decisions that occur during normal development/differentiation, where few samples are obtained per tissue/stage. As a case study, I focussed on neural induction of human pluripotent stem cells (WA9/H9 cell line) during four

[xxxv] 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.

[xxxvi] This is the first application of peaBrain when trained on a dataset other than GTEx. All prior tasks assessed performance of results derived from the GTEx-trained models. The task described in Chapter 2.6 highlights the utility of the architecture on datasets that are significantly smaller (i.e. with as little as 2 samples per tissue) and the utility of comparing peaBrain model predictions as an alternative (and more powerful) approach to identifying differentially expressed transcripts.

consecutive stages for which data was available (see **URLs**): human embryonic stem cells, neuroepithelial cells, early radial glial cells, and mid radial glial cells[xxxvii]. I constructed a peaBrain model specific to each stage that predicts stage-specific expression from the DNA promoter sequence annotated with H3K4me1, H3K4me3, H3K27ac, and H3K27me3 as well as DNA methylation peaks (from the corresponding stage). Each stage-specific peaBrain model (n = 4) emulates the transcriptional machinery active, which I can use to identify the subset of genes between any two consecutive stages whose expression is directly altered by epigenetic modifications in the core promoter elements. From all genes with non-zero expression in at least one of the four stages (n = 29,696)[xxxviii], between embryonic and neuroepithelial cells, I identified 1271 genes whose normalized expression changed by more than 0.5 standard deviations as a consequence of epigenetic modifications (at a q-value < 0.05). Similarly, the expression profiles of 555 genes were identified as transcriptionally altered (downstream of epigenetic alterations in the promoter) during the transition from neuroepithelial to early radial glial cells, and 851 genes for the transition between early and mid radial glial cells. Nearly all of these genes were undetected with simple differential expression analyses, which may be a consequence of limited power to detect with classical approaches (samples were collected in replicate only).

Among the genes identified in the differentiation of neuro-epithelial (from embryonic stem cells) and mid radial glial cells (from early radial glial cells), there was significant enrichment of genes implicated in schizophrenia, autism, bipolar, and depression (1.5-2.6 fold enrichment; Fisher's exact $p < 0.05$; **Tables 2.6-2.8**). This trend became more pronounced and more significant when limited to genes implicated using genetic associations alone (**Tables 2.6-2.8**). With nearly 5-10% labelled as transcription factors (1.53-3.24 fold enrichment), this cross-disease enrichment suggests a putatively causal biological mechanism early in neural development for the shared genetic correlation between these psychiatric phenotypes – an observation that is difficult to extract from GWAS data alone (due to

[xxxvii] Similar to the GTEx dataset, a model was trained for every tissue stage (n = 4 in total). Since I was interested in the differences between the stages, I performed 3 comparisons (one between each pair of the two consecutive stages).

[xxxviii] If the gene had zero expression in all 4 stages, it was excluded as I was looking to identify genes with differential expression between the 4 stages.

difficulties associated with variant interpretation) and is missed by pairwise differential peak/gene expression analyses (due to low number of samples). The enrichment of disease-implicated genes was smaller and less significant for during the transition from neuroepithelial to early radial glial cells (p-value > 0.05 for bipolar and autism), implicating the shared disease processes in sub-portions of neural development. (The analysis for all three transitions between the four stages were similarly powered and the underlying analysis was methodologically identical in all respects.)

Table 2.5 I observed significant enrichment of genes implicated in psychiatric disease among the set of genes whose expression is altered by epigenetic modifications in promoter sequences during the transition from human embryonic stem to neuroepithelial cells. The fraction column denotes the percentage of the transcriptionally-altered genes that are also disease genes, e.g. 7% of genes whose expression is altered by epigenetic modifications during this transition are implicated in autism. I assessed enrichment using Fisher's exact test. For all diseases, I included the corresponding gene sets from the Open Targets platform. For autism, I included two gene sets (one from Open Targets and one from Simon's Foundation [labelled by SFARI]).

	Odds Ratio	**p-value**	**Fraction**
Autism	2.06	6.74×10^{-9}	7.00%
Autism (genetic associations)	2.32	1.56×10^{-1}	0.24%
Autism (SFARI)	2.09	7.63×10^{-8}	5.82%
Bipolar	1.92	8.00×10^{-9}	8.50%
Bipolar (genetic associations)	2.59	6.65×10^{-8}	3.62%
Depression	1.68	2.17×10^{-9}	14.71%
Depression (genetic associations)	1.94	1.59×10^{-4}	3.38%
Schizophrenia	2.07	1.90×10^{-20}	19.35%
Schizophrenia (genetic associations)	2.47	1.61×10^{-14}	8.26%
Transcription Factors (GO: sequence-specific DNA binding)	3.24	4.26×10^{-24}	9.28%

Table 2.6 I observed reduced enrichment of genes implicated in psychiatric disease in the set of epigenetically-driven transcriptomically-altered gene expression profiles during the transition from neuroepithelial to early radial glial cells. Conservative adjustment for multiple testing eliminates all significance.

	Odds Ratio	**p-value**	**Fraction**
Autism	1.57	2.19×10^{-2}	5.59%
Autism (genetic associations)	1.70	4.53×10^{-1}	0.18%
Autism (SFARI)	1.54	4.34×10^{-2}	4.50%
Bipolar	1.47	2.68×10^{-2}	6.85%
Bipolar (genetic associations)	1.31	3.78×10^{-1}	1.98%
Depression	1.54	9.85×10^{-4}	13.87%
Depression (genetic associations)	1.70	3.79×10^{-2}	3.06%
Schizophrenia	1.64	5.50×10^{-5}	16.40%
Schizophrenia (genetic associations)	1.77	2.91×10^{-3}	6.31%
Transcription Factors (GO: sequence-specific DNA binding)	1.99	4.36×10^{-4}	6.31%

Table 2.7 As observed in the differentiation of neuroepithelial cells (from embryonic stem cells), the enrichment of genes implicated in psychiatric disease among the set of genes whose expression is altered by epigenetic modifications during the transition from early to mid radial glial cells.

	Odds Ratio	**p-value**	**Fraction**
Autism	1.85	5.91×10^{-5}	6.46%
Autism (genetic associations)	3.52	6.29×10^{-2}	0.35%
Autism (SFARI)	2.03	1.48×10^{-5}	5.76%
Bipolar	2.00	2.03×10^{-7}	8.93%
Bipolar (genetic associations)	1.66	3.17×10^{-2}	2.47%
Depression	1.82	3.64×10^{-9}	15.86%
Depression (genetic associations)	2.37	1.23×10^{-5}	4.11%
Schizophrenia	1.98	8.62×10^{-13}	18.92%
Schizophrenia (genetic associations)	2.41	5.78×10^{-10}	8.23%
Transcription Factors (GO: sequence-specific DNA binding)	1.53	1.14×10^{-2}	4.94%

2.5 Small molecule extension of the peaBrain DNA-reporter assay captures ~84% of the variance in the average molecule perturbed gene expression profiles of L1000 landmark transcripts[xxxix]

To predict the transcriptomic perturbations of small molecules, I extended the peaBrain DNA reporter assay to incorporate molecular fingerprints[xl] (boolean arrays/bitmaps that are characteristic of the molecule), in addition to the DNA core promoter sequence as input (which directly accounted for copy number and genotypic variation, see **Chapter 1, Figure 1.1b**). Given a core promoter sequence and a molecular fingerprint, the small molecule peaBrain model will predict the expression of the corresponding gene in that cell line (see **Chapter 1, Figure 1.1b**). (Unlike the previous applications of peaBrain, this required training the models specific for these cell lines:

gene promoter + molecule fingerprint → expression of gene

This is the reason that for why this application is a separate subsection of the chapter. In these set of models, I am assessing the models' ability to predict change in expression given different molecule structures.) I limited my analysis to molecular fingerprints – encoded as a standard bit vectors – generated using Morgan's algorithm[52] from the molecular-input line-entry system (SMILES) specification for each molecule. For each molecule profile (at 10uM dosage), I rank-normalized expression (averaging the replicates) and min-max scaled (per gene) prior to training. Training was conducted with expression profile data downloaded from the LINCS phase I L1000 Connectivity Map (GSE92742) dataset[xli] on NCBI's Gene Expression Omnibus (GEO; see **URLs**)[3], limited to cell lines for which genotype, copy-number, and phase I data was available. (The phase II [GSE70138] dataset was used as the external test set[xlii].) The L1000 assay measures the expression profile of 978 landmark

[xxxix] As described in a footnote in Chapter 1: L1000 is an expression-profiling assay that directly measures the abundance of 978 transcripts, which is subsequently followed by the computational inference of the remaining transcriptome (22,000 genes). In addition, 80 "invariant" genes are also directly measured to enable scaling and normalization of the imputed values. For this thesis, I only use the 978 directly measured genes and discard all imputed profiles (which tend to be of variable quality).

[xl] Molecular fingerprints are one representation/encoding of the structure of a molecule. The most common type of fingerprint is a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule.

[xli] LINCS Phase I data is in GEO Series GSE92742. This represents an earlier phase of LINCS project.

[xliixlii] LINCS Phase II data is in GEO series GSE70138. This series is updated every 6 months as more L1000 data is produced over the duration of the LINCS program (starting in 2016, through 2020). For this chapter, the Phase II represents the "external" test set.

transcripts, whose promoter sequences were used in training (imputed transcripts were discarded; see **Box 2.3 - Methods**). Genotype and copy-number data for each cell line included in the analysis, identified by PICNIC analysis of Affymetrix SNP6.0 array data, were obtained from the Sanger's Institute COSMIC cell line webpage (see **URLs** and **Box 2.3 Methods – peaBrain small molecule and shRNA screens**, *below*, for details).

Box 2.3 Methods – peaBrain small molecule and shRNA screens

peaBrain small molecule and shRNA screens were designed as the classical peaBrain DNA reporter assay, but with two additional dense layers to incorporate either small molecular fingerprints or one-hot encoded oligo-sequences. The difference between the small molecule and shRNA screens was the size of the input layer. Training and test data for both screens was obtained from the LINCS L1000 Connectivity Map phase I (GSE92742) and phase II (GSE70138) datasets, respectively; both datasets were downloaded from GEO[3] and were processed as described in text. I only included cell lines included in both phases of the LINCS L1000 and for which genotype data was available from the Sanger's Institute COSMIC cell line webpage (see **URLs**). For the drug repositioning, molecule SMILES were downloaded from the ChEMBL database[5] (v18), with drugs last updated on (17/17/01) and indications last updated on (17/15/03). Molecular fingerprints were extracted from SMILES using the GetMorganFingerprint function from RDKit[6] (an open-source cheminformatics software, see **URLs**).

I assessed the predictive performance of the cell line-specific peaBrain models using a Monte Carlo-validation scheme (**Figure 2.6,** *on next page*). In each fold, 5% of small molecules were randomly withheld as a test set (that is, neither the fingerprint nor the expression profile is ever seen by the model). Random selection of the test set was sufficient as molecular similarity was very weakly correlated with transcriptomic similarity (median Spearman's rho = 1.0%). Of the remaining 95% of small molecules, 10% was randomly selected as validation to check for early exit criteria (i.e. not used to train the model parameters). The model was then trained on the remaining drug profiles to predict the gene-level expression, 24 hours post-exposure, for any pair of molecular fingerprints and core promoter sequences (per cell line). Across all cell lines, the model had a median oos-r^2 of 84% (**Figure 2.6**). Importantly, this performance was 12% better (on average) than simply using the expression profile of the closest structurally similar molecule in the training set (e.g. using the most structurally-similar oestradiol analogue as the prediction of the oestradiol expression profile). This suggests that the model is able to extract more biologically-relevant information (likely from the promoter sequences) than that encoded in the molecular fingerprint used as input to the algorithm.

To further test the model generalizability, I assessed predictive performance on the rank-normalized LINCS phase II GSE70138 ("external") dataset. Moreover, to calibrate my assessment to the upper bound of assay reproducibility, for each test set, I sub-selected molecules assessed in both the phase I and phase II LINCS experiments. This way I am able to benchmark peaBrain predictive performance to experimental replicates (using the same assay and platform); that is, for each small molecule, I performed two tests: how well the imputed peaBrain profile compares to the phase II profile, and how well the phase I profile (i.e. the experimental replicate) compares to phase II profile. I observed almost identical performance between peaBrain predictions (median rho between peaBrain and phase II profiles = 54%) and the phase I experimental replicates in predicting phase II profiles (median rho between phase I and phase II profiles = 47%), suggesting that *in silico* peaBrain predictions of molecule-induced transcriptomic perturbations are as accurate as *in vitro* experimental replicates (see **URLs**).

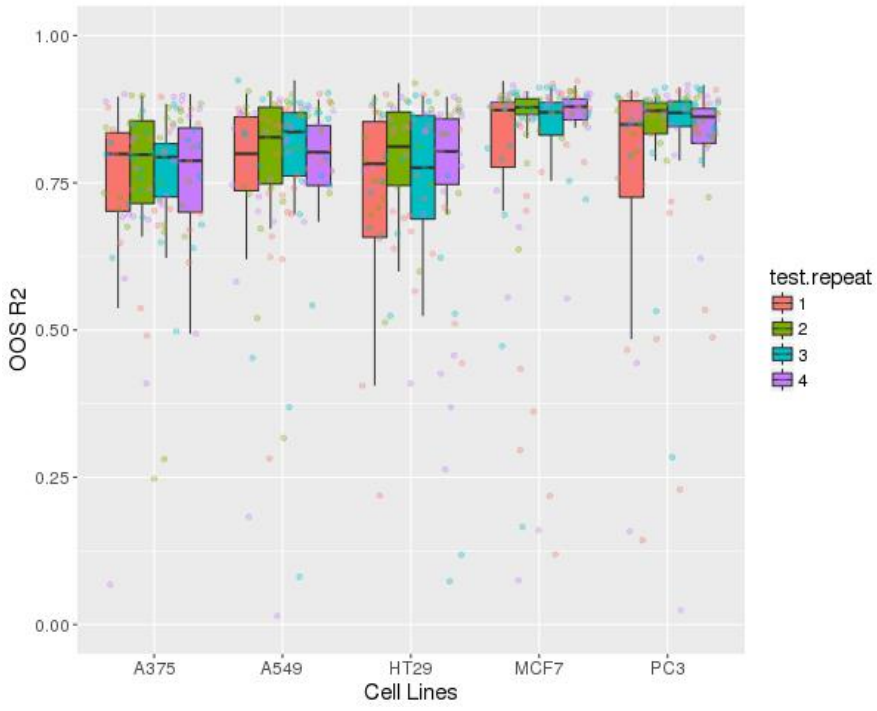


Figure 2.6 Given a core promoter sequence and a molecular fingerprint, the small molecule peaBrain model can accurately the transcriptomic perturbation of the corresponding gene across all 5 cell lines. Each point represents oos-r^2 of a single unique small molecule (included in the test set). I performed 4 folds of Monte Carlo validation per cell line (folds denoted by the colour). The y-axis was clipped at zero; only seven molecules across all cell lines and test/validation repeats had an oos-r^2 < 0. **Abbreviations:** OOS R2, out-of-sample r^2.

peaBrain imputed molecule profiles significantly outperform chemoinformatic/structure-base approaches in predicting drug approval. Having established the predictive capacity of the small molecule peaBrain models (on par with *in vitro* experimental replicates), I was interested in leveraging the peaBrain framework for drug repositioning. To this end, I imputed the expression profile for all clinically-relevant small molecules in the ChEMBL database[5] across all 5 cell lines (i.e. all molecules assigned at least to one indication). I divided all clinically-relevant molecules in two sets: molecules whose first approval occurred before the year 2000 (inclusive; n = 728 molecules), and all molecules that gained first approval after the year 2000 (n = 344 molecules). For indications relevant to each cell line (e.g. "breast neoplasms" for MC7 or "non-small cell lung carcinoma" for A549; **Table 2.8,** *on next page*), I identified the subset of molecules assigned to the corresponding indication as the positive class (background is all other clinically-relevant molecules). I then trained a random forest classifier on the pre-2000 molecule set using the imputed peaBrain molecule profiles. I sought to assess how well the imputed peaBrain expression profiles can identify clinically-relevant molecules for a given indication, from all other molecules. To establish a baseline for comparison, I also constructed an equivalent classifier using only structural information (i.e. the molecule fingerprint used to impute expression). I assessed performance using five different metrics[xliii] (useful metrics for classification with class imbalance):

(a) F1 score (harmonic mean of precision and recall);

(b) average precision (average of precisions achieved at each threshold weighted by the increase in recall from the previous threshold; corresponds roughly to the area under the precision-recall curve but is more conservative);

(c) precision score (positive predictive value; the fraction of true positives among all molecules predicted as positive by the classifier); and

(d) recall score (sensitivity; fraction of true positives over the number of all molecules relevant for the given indication).

[xliii] These five scores cover virtually all common/established metrics for assessing performance in unbalanced datasets.

All scores range from 0 to 1, with larger values indicating (elements of) better performance.

Across all five cell lines and the corresponding five indications, the imputed peaBrain expression profiles significantly outperformed chemoinformatic/structure-based approaches in predicting molecule clinical relevance for any given indication (**Table 2.8**, *below*), with 64-86% increase in F1-scores, 22-63% increase in precision, 75-94% increase in recall, and 13-19% increase in average precision. peaBrain models that use molecule expression profiles from all cell lines have markedly increased performance compared to peaBrain models that use expression profiles from a single cell line, which in turn outperform structure-based approaches.

Table 2.8 Tabulated classifier summary statistics on the post-2000 clinically-relevant molecule set (after training on the pre-2000 molecule set) for all five indications. For each indication, I constructed two peaBrain models: one using the expression profiles for the indication-relevant cell line and one using the expression profiles for all cell lines (denoted by "all"). The % increase columns highlight the increase in performance of the peaBrain classifiers compared to the baseline structure-based approach. The "all" peaBrain models significantly outperform all other approaches; the single cell line peaBrain models also have markedly increased performance compared to the baseline structure-based approach. This performance is consistent even with stricter definitions for class indications (after removing molecules generally indicated for the MeSH term "neoplasms").

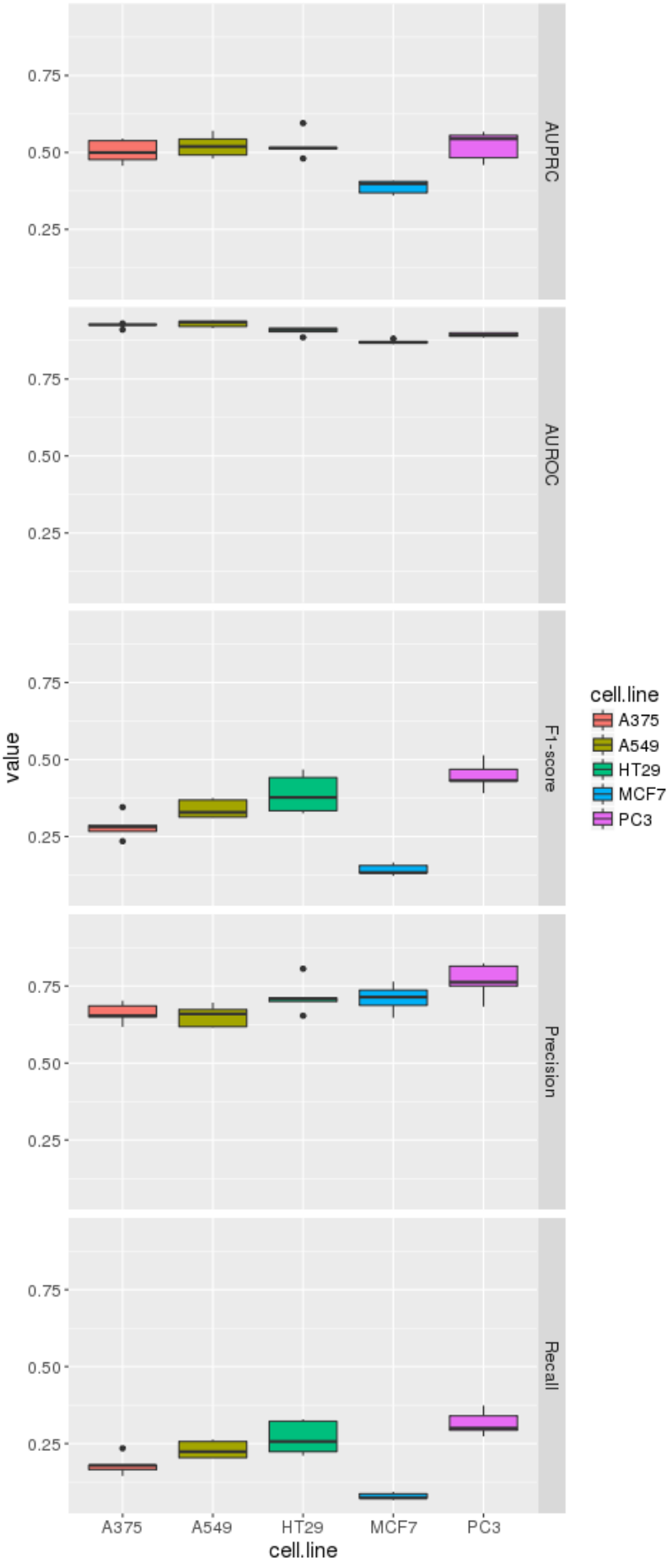
Melanoma + Neoplasms (A375)	structure-based approach	peaBrain (A375)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.12	0.14	**20.37**	0.27	**78.97**
Average Precision	0.37	0.42	**12.05**	0.41	**10.84**
Precision Score	0.40	0.70	**54.55**	0.52	**25.35**
Recall Score	0.07	0.08	**15.38**	0.18	**90.91**
Non-small cell lung carcinoma + Neoplasms (A549)	structure-based approach	peaBrain (A549)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.14	0.18	**26.93**	0.29	**71.79**
Average Precision	0.36	0.40	**9.99**	0.43	**15.52**
Precision Score	0.44	0.64	**38.02**	0.55	**22.49**
Recall Score	0.08	0.10	**25.00**	0.20	**83.33**
Colorectal neoplasms + Neoplasms (HT29)	structure-based approach	peaBrain (HT29)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.17	0.18	**6.90**	0.33	**65.19**
Average Precision	0.39	0.39	-0.99	0.46	**15.82**
Precision Score	0.50	0.45	**9.52**	0.63	**22.22**
Recall Score	0.10	0.11	**10.53**	0.22	**75.86**
Breast neoplasms + Neoplasms (MCF7)	structure-based approach	peaBrain (MCF7)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.14	0.33	**83.74**	0.35	**89.18**
Average Precision	0.38	0.37	-3.26	0.46	**18.87**
Precision Score	0.32	0.50	**43.90**	0.62	**64.06**
Recall Score	0.09	0.25	**96.77**	0.25	**96.77**
Prostatic neoplasms + Neoplasms (PC3)	structure-based approach	peaBrain (PC3)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.18	0.25	**30.73**	0.30	**51.15**
Average Precision	0.43	0.39	-9.22	0.44	**2.22**
Precision Score	0.50	0.61	**19.61**	0.56	**11.11**
Recall Score	0.11	0.15	**33.33**	0.21	**62.07**

shRNA extension of the peaBrain DNA-reporter model captures ~85% of the variance in perturbed gene expression, allowing us to identify 292 new transcription factor-target interactions. Using a nearly identical model and assessment procedure, I can also predict the transcriptomic consequences of shRNA. Instead of a molecular fingerprint, I one-hot encoded[xliv] the shRNA oligo-sequence as input, alongside the promoter sequence (**Chapter 1, Figure 1.1c**). To highlight the utility of the shRNA model, I sought to predict the transcriptomic consequences of 330,617 unique shRNA oligo-sequences (a subset of which targeted 19,992 human genes) to enable inference of the human regulatory network; the extensive number of conditions provides (multiple) snapshots of the transcriptome with nearly every gene perturbed. I used a meta-inference approach (aggregation of multiple inference approaches) for network reconstruction, limited to genes directly measured on the L1000 platform (978 genes; see **Box 2.3**).

Using the HTRIdb[53] (an open-access database for experimentally verified human transcriptional regulation interactions) as the gold standard, I assessed my network inference for all 5 cell lines using a 5-fold Monte Carlo validation scheme with 10% of edges as test set (in every repeat). I noted the regulatory networks for each of the 5 cell lines captured experimentally-verified interactions at a precision ranging from 66-77%, with AUROC ranging from 87-93%, AUPRC from 39-52%, recall from 26-37% and F1-scores from 14-45% (**Figure 2.7**). For comparison, the "Dialogue on Reverse Engineering Assessment and Methods" (DREAM) project constructed high-confidence networks for *Eschierica coli* and *Saccharomyces cerevisiae* (simpler organisms) at a precision of ~50%[54]. The inferred networks, for all 5 cells, have been made available (see **URLs**).

Figure 2.7 (*on next page*). Plots depicting the performance of network inference using the large-scale shRNA-perturbed expression matrix, across all 5 cell lines. Boxplot for each metric was generated using a 5-fold Monte Carlo validation scheme (with 10% of edges randomly selected as the test set). *Abbreviations:* AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve.

[xliv] From footnote in Chapter 1: One-hot encoding is a technique commonly used in machine learning to encode categorical integer features with each channel indicating the presence (1) or absence (0) of the corresponding DNA letter.


AUPRC
AUROC
F1-score
Precision
Recall
value
0.75
0.50
0.25
A375
A549
HT29
MCF7
PC3
cell.line
cell.line
A375
A549
HT29
MCF7
PC3

I subsequently identified all edges inferred as interactions – but absent from the experimentally-verified HTRIdb curation – as new/candidate regulatory interactions (n = 212 total). Using ChIP-seq data from ENCODE[10], I verified that 83% of have experiment support; that is, the predicted transcription factor binds within 50kb of the transcription start site of the corresponding target. Of all 212 interactions, 40% bind within 2kb of the transcription start site. All new inferred interactions have been made publicly available (see **URLs**).

Chapter 3: Individual Variation peaBrain

3.1 peaBrain can predict transcriptomic consequences of individual variation

I subsequently extended the peaBrain model to incorporate the transcriptomic consequences of individual genotype variation (**Chapter 1, Figure 1.1d**). Given whole genome sequencing data[xlv] of a group of individuals (such as GTEx participants), I 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 given sequence.

To do this, I 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 these set of analyses, I did not make use of individual level epigenomic and regulatory annotations as these were not available.) More concretely, unlike aformentioned models, for a single gene, this peaBrain extension predicts the difference between the expression of two individuals as a function of the difference in the sequences between the two individuals[xlvi] (for the given gene; see **Box 3.1 – Methods – Pre-processing for heritability and variant-sensitive regression**, *on next page*). By jointly modelling the input "difference" sequence in a non-linear manner, I hypothesized that I 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.

[xlv] Whole genome sequencing is required because I reconstruct the individual's sequence. Genotype data may (or may not) work depending on the quality and depth of imputation.

[xlvi] Why the difference? The difference in expression between any two individuals can be attributed to the difference in sequences between the two individuals. Thus, looking at the "difference in sequence" effectively removes all sequence features that are not relevant to the difference in expression (i.e. this is one approach to "de-noising the input"). Furthermore, by training the model on differences between individuals, I can increase the number of training points.

Box 4.1 Methods – Pre-processing for heritability and variant-sensitive regression

Pre-processing for heritability and variant-sensitive regression for the individual variation peaBrain model was performed as recommended by the authors of QTLTools[2]. For each tissue, I 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 peaBrain DNA reporter assay analyses, I further limited my analysis to protein-coding genes (number of genes differed between tissues) and to autosomal chromosomes for simplicity.

I defined the input sequence, for each protein-encoding gene on an autosomal chromosome, 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, 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, my 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[1]. Incidentally, the 1Mb interval has also been used by other approaches in imputing RNA expression from genotype (namely, TWAS)[8]. Using the unphased whole genome sequencing GTEx data, I reconstructed the individual's sequence from the quality controlled VCF. In other words, I 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."[1] For each individual, I generated two copies of the gene 1Mbps input sequence; phasing did not matter as the sequences were combined prior to modelling.

Consistent with evidence from eQTL studies[55], I noted the 4kbps core promoter used in the earlier models did not capture enough cis-heritability as estimated by constrained GCTA[56] and thus was not sufficiently informative for this predictive task[xlvii]. 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. I selected a 1Mbps input length, centred on the annotated TSS (0.5Mbps upstream and 0.5Mbps

[xlvii] Constrained GCTA heritability analyses: I converted the GTEx whole genome sequencing VCF to PLINK binary bed file (using Plink v1.9). Using GCTA v1.24.4, I 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 and TWAS[8].. For each gene, I 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.

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, the 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[1]. 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[xlviii]. Limited to the genes with significant non-zero heritability in LCLs ($n = 816$), on 600 cores, the complete analysis took approximately a month[xlix]. (The earlier peaBrain models only required several hours of training[l].) Prior to training, for each individual, I re-constructed the 1Mbps input sequence from the variants called from whole genome sequence (WGS) data (see **Box 4.1 Methods – Pre-processing for heritability and variant-sensitive regression**, *above*). Exploring the peaBrain architecture, fine-tuning the model parameters, and deploying the models was conducted on NVidia's P100 GPUs; the bulk of the training, however, was run on CPUs.

The peaBrain model used here shares the same core architecture to the peaBrain models described in earlier chapters (see **Chapter 1, Figure 1.1d**). Three concatenated peaBrain architectures were connected by a dense fully-connected layer prior to the output neuron (**Appendix I – Supplementary Table 2**[li]). The input 1Mbps sequence is split into three inputs: 0.48Mbps upstream, 4kbps core promoter, and 0.48Mbps downstream sequences. The 4kbp core promoter is the input to a CNN with identical structure and hyperparameters as described for peaBrain DNA reporter assay (described above

[xlviii] Model size scaled with the input of the DNA sequence (i.e. the larger the input DNA sequence, the larger the memory footprint).
[xlix] The longer training time can be attributed to the larger input sequence (i.e. takes more time to manipulate computationally) and the number of models needed (as each gene required a separated model).
[l] peaBrain models in earlier chapters relied only on the 4kbps promoter sequence, which is substantially shorter than the 1Mbps required for the tasks detailed in this chapter.
[li] Appendix I – Supplementary Table 2 details the model architecture for this task.

in detail). The upstream and downstream sequences are input to networks with identical architecture, but different pooling hyperparameters: a pool size of 100 for the first pooling layer, 50 for the second, and 10 for the last. Number of filters was consistent between all networks (n = 11) and was determined empirically (see Stage 1 model descriptions above). The fully connected output from each sequence is concatenated, before one penultimate fully-connected layer and a single output node. These models were trained to predict the differences between individuals (rather than direct prediction of expression). As humans are diploid, for each individual, the input sequence was the sum of the one-hot encoding of each of the 1Mbps sequences corresponding to the "maternal" and "paternal" sequences; phasing did not matter because addition is commutative (i.e. the sum was consistent). A separate model was constructed for each gene with significant non-zero heritability. For any pair of individuals, A and B, the input sequence was defined as the difference between the one-hot encoded sequences, with the corresponding output as the difference between the two individuals. I included both differences, (A – B) and (B – A), during training[lii]. After removing individuals with cryptic relatedness (see GCTA analysis below), the GTEx dataset was randomly split into train and test individuals (95% of subjects for training and 5% for testing), with the model trained on all the pairwise differences between train individuals and tested on all pairwise differences between test individuals. The training set was further sub-divided into training and validation sets, with the latter used to exit early after a minimum 100-epoch training. As described below, overall model performance was assessed using the oos-r^2 on five to ten random repetitions of 95/5 train/test splits; the number of repetitions was dependent on how quickly each model reached exit criteria.

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[26], DeepSEA[28],

[lii] This was done only in training and not in testing.

MPRA[liii,15], BiT-STARR-seq[liv,17], and HiDRA[lv,16]), I designed four tasks (**tasks E-H**; described below). For the comparison with elastic net, in line with other recent studies in the field[8], I 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[56] (limited to the 1Mbps input sequence; $p<0.01$). By limiting analysis to genes with detectable *cis*-heritability, I can make more meaningful conclusions about the comparative performance of the different methodologies[lvi]. I restricted analysis to LCLs[lvii] to enable comparisons with empirical data (**tasks F-H**) and to reduce the compute burden. (I do not perform transcriptome-wide case-control analyses due to the compute resources required as highlighted above.)

3.2 peaBrain identifies functional architecture that is inaccessible with current high-throughput experimental assays.

First (**task E**)**,** I compared the predictive performance of peaBrain to that of a regularized linear model (an implementation of elastic net identical to that used in PrediXcan)[lviii]. RNA-seq samples from the GTEx dataset (n = 94 individuals after filtering) were pre-processed, residualised to account for cryptic relatedness, biological confounders, and technical variance, and rank transformed to normality (see **Box**

[liii] Massively parallel reporter assays (MPRAs) are assays that can assess the transcriptomic consequences of tens of thousands of short DNA sequences. The DNA is first synthesized on an array, and the oligos can subsequently be integrated into a plasmid, which can subsequently be inserted into a cell.
[liv] BiT-STARR-seq is a variation of self-transcribing active regulatory region sequencing (STARR-seq), in which a source genome is fragmented and recombined into a reporter vector before transfection into cells.
[lv] HiDRA (High-Definition Reporter Assay) is an assay that combines components of Sharpr-MPRA and STARR-Seq with genome-wide selection of accessible regions derived from ATAC-Seq.
[lvi] If the expression of the gene was not significantly heritable, it made no sense to predict the expression from the sequencing data.
[lvii] LCLs stand for lymphoblastoid cell lines, also known as EBV-transformed lymphocytes. It is one of two cell lines in the GTEx set of 45 tissues. Empirical data (from experimental *in vitro* high throughput assays) was only available for LCLs.
[lviii] I used an additive genetic model (i.e. an elastic net regularized linear model) as my baseline comparison as described elsewhere[22]. Briefly, for each gene, an elastic net model was used to model expression (alpha = 0.5; selected to match PrediXcan[26]. 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[52].

3.1 Methods – Pre-processing for heritability and variant-sensitive regression, *above*) before modelling. Model performance for both linear models and peaBrain was assessed by generating oos-r^2 for 5% of individuals randomly withheld from training and unrelated to individuals in the training set (repeated 3-10 times, depending on the how quickly the model reached the exit criteria and the performance of earlier repeats). For each of the 816 genes with non-zero heritability (GCTA $p < 0.01$), I calculated the 95% confidence interval for the oos-r^2, defining a gene as successfully predicted if the entire oos-r^2 confidence interval exceeded zero to ensure I only consider genes with high-confidence models. Whilst regularized linear models were able to capture *cis*-heritability for 28 of the 816 genes, the equivalent number for peaBrain was 113[lix]. For the rest of this chapter, I call these 113 genes as "captured" by the peaBrain model. *Cis*-heritability for 3 genes was captured by both models, with the oos-r^2 confidence interval largely overlapping (see **URLs**). **Supplementary Table 5 (see URLs)[lx]** tabulates the performance metrics (confidence and point estimates for oos-r^2 from both classes of models) and estimated GCTA heritability for all genes.

peaBrain predictions were significantly and positively correlated with univariate eQTL coefficients[lxi]. Having established the predictive ability of peaBrain, I was interested in whether I can use the best-performing peaBrain models to measure the impact of single variants, compared to DeepSEA logFC estimates and MPRA log skew estimates (**Task F**). For all "captured" genes (n=113)[lxii], I selected all variants identified as significant eQTLs in the GTEx v6p univariate eQTL analysis (n=16,019 variants) and replicated the analysis with the Geuvadis dataset[57] (n=17,279 variants

[lix] I termed these 113 genes as "captured" by the peaBrain model; that is, a significant portion of the variance in expression could be predicted by the peaBrain model using the sequence data alone.
[lx] Supplementary Table 5 is too large to embed here. It is available electronically (see **URLs**) and contains the results for all genes.
[lxi] 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 the peaBrain model predicts the difference between two sequences, I 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).
[lxii] I only include captured genes because I know the peaBrain model can not predict expression for any of the remaining genes (as assessed by the out-of-sample metrics). Thus, analysis of any of these genes is not meaningful (i.e. "noisy" analysis).

for the EU population and n = 1601 variants for the YRI [Yoruba from Ibadan, Nigeria] population). Unlike GTEx v7, eQTLs from GTEx v6p were derived from genotyping arrays (Illumina OMNI 5M or 2.5M) and thus did not include WGS used to train the peaBrain model. The Geuvdais dataset (both EU and YRI populations) were eQTLs derived from WGS from a set of non-overlapping subjects. For each eQTL (including indels), I created pairs of 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. peaBrain predictions significantly and positively correlated with the univariate eQTL coefficients from the GTEx analysis (Spearman's rho = 0.09; p = 3.02 $x10^{-32}$; **Figure 3.1**, *on next page*), from the EU-Geuvadis analysis (rho = 0.10; p = 9.60 $x10^{-38}$; **Figure 3.2**), and from the YRI-Geuvadis analysis (rho = 0.18; p = 8.64 $x10^{-13}$; **Figure 3.3**). Results were consistent across all datasets when I relaxed my heritability to include genes with GCTA $p < 0.05$: Spearman's rho equal to 0.08 (p = 3.02 $x10^{-25}$), 0.07 (p = 3.65 $x10^{-25}$), and 0.18 (p = 1.68 $x10^{-13}$) for the GTEx, EU-Geuvadis, and YRI-Geuvdais univariately-significant eQTLs, respectively. Alongside this positive correlation, I observed that many variants with large coefficients across all three datasets had small peaBrain predictions (**Figures 3.1-3.3**). This shrinkage in peaBrain estimates is consistent with the appreciable LD between eQTL-associated variants at any given locus, such that only a subset of significantly-associated variants are actually functional. Whilst univariate linear models are not capable of distinguishing between functional versus "hitchhiker" variants, the joint modelling of the input sequence in peaBrain allows a more direct assessment of the function of each variant. In comparison, I noted that the MPRA log skew estimates[lxiii] did not correlate with the univariate eQTL coefficients in the GTEx (rho = 0.014; p = 0.60), the EU-Geuvadis (rho = -0.018; p = 0.58), or the YRI-Geuvadis (rho = 0.011; p = 0.82) eQTL analyses. The BiT-STARR-seq log skew estimates[lxiv] also did

[lxiii] MPRA variant results were obtained from Supplementary Table 1 of Tewhey *et al.*15 Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519-1529 (2016). (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 allelic skew from the combined MPRA LCL analysis (alt/ref), was used as the MPRA log skew estimate.

[lxiv] The BiT-STARR-seq data was similarly processed (see **URLs**). As with the peaBrain and eQTL analysis, significance of the spearman (rank) correlation between each of the experimental assays allelic log skews and the univariate eQTL coefficients was assessed using the cor.test function.

not correlate with the univariate eQTL coefficients in the GTEx (rho = -0.031; p = 0.50), the EU-Geuvadis (rho = 0.035; p = 0.44), or the YRI-Geuvadis (rho = -0.112; p = 0.32) eQTL analyses. Similarly, for GM12878 (LCL) annotations, I noted that the maximum log fold changes from DeepSEA[lxv] did not correlate with the magnitude of the eQTL coefficients in the GTEx analysis (rho = -0.001; p = 0.92), the EU-Geuvadis analysis (rho = -0.010; p = 0.18), and the YRI-Geuvadis (rho = 0.016; p = 0.52). The DeepSEA results suggest that simply predicting chromatin effects and TFBS is not sufficient for predicting the transcriptomic consequences of sequence variation for this set of selected genes. This is consistent with experimental evidence suggesting that most genetic variants in DNase footprinted (and other annotated) regulatory regions are silent[58], and the fact that histone modifications (e.g. methylation) alone are frequently insufficient to transcriptionally perturb promoters[59]. However, it is important to note that DeepSEA was trained on the reference genome (i.e. not exposed to genotype variation), while peaBrain was trained to predict differences in expression as a function of differences in sequence between any pair of individuals.

Figure 3.1 *(on next page)* 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}).

[lxv] To obtain the DeepSEA logFC for GM12878 annotations, a vcf file of the corresponding eQTLs was uploaded to the DeepSEA platform (see **URLs**).

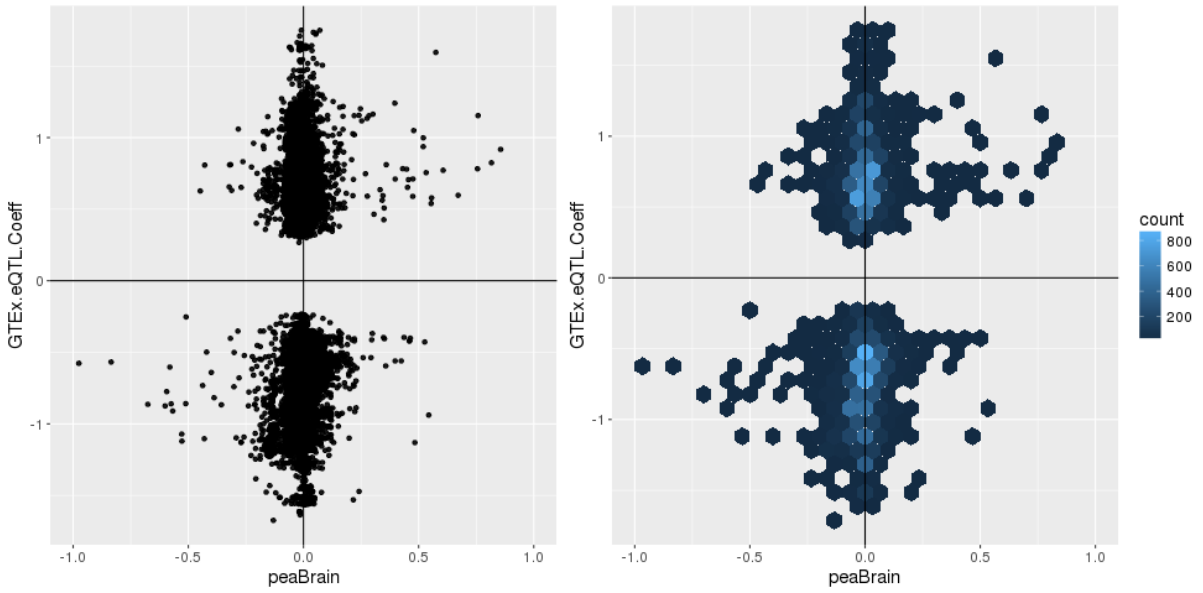


Figure 3.2 Scatter (*right*) and hexa-bin (*left*) plots of variant-expression effects as estimated in LCLs by peaBrain (limited to genes whose 95% confidence interval for the oos-r^2 is entirely above 0; n = 113 genes; **Task F**). Each point corresponds to a variant that is univariately significant in the EU-Geuvadis eQTL analysis (n = 17,279 eQTLs). The y-axis is the magnitude of the univariate EU-Geuvadis eQTL coefficient for the corresponding variant. The correlation between the EU-Geuvadis coefficient and the peaBrain prediction is positive and significant (Spearman's rho = 0.10; p = 9.60 $x10^{-38}$).

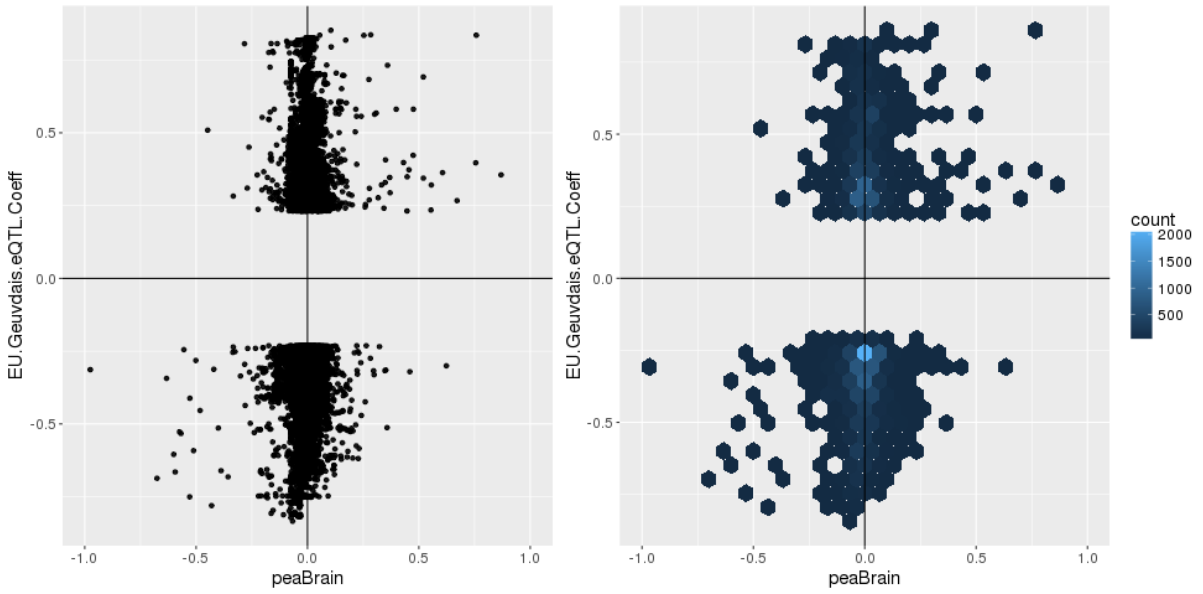
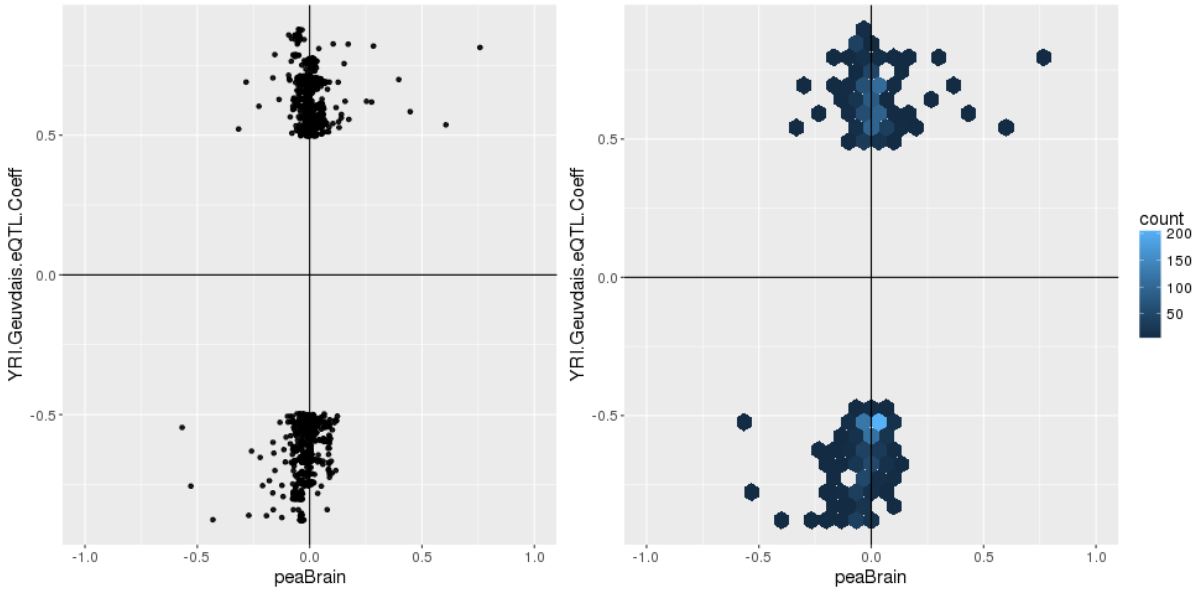


Figure 3.3 Scatter (*right*) and hexa-bin (*left*) plots of variant-expression effects as estimated in LCLs by peaBrain (limited to genes whose 95% confidence interval for the oos-r^2 is entirely above 0; n = 113 genes; **Task F**). Each point corresponds to a variant that is univariately significant in the YRI-Geuvadis eQTL analysis (n = 1601 eQTLs). The y-axis is the magnitude of the univariate YRI-Geuvadis eQTL coefficient for the corresponding variant. The correlation between the YRI-Geuvadis coefficient and the peaBrain prediction is positive and significant (Spearman's rho = 0.18; p = 8.64 $x10^{-13}$).



Next, I 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[16]; MPRAs leverage shorter synthesized DNA sequences[14]. 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. I 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 the exact cell line was not specified[17]. For each chromatin annotation, I assessed significance using a

Box 3.2 Methods – Comparison of peaBrain, MPRA, BiT-STARR-seq and HiDRA

The core 15-state model and DNAse accessibility annotations for the GM12878 EBV-transformed lymphoblastoid cell line (LCLs) were downloaded from the Roadmap project (see **URLs**). HiDRA data was downloaded from the GEO series GSE104001 (see **URLs**). HiDRA estimate was defined as the log fold change in average counts between the alternate and reference group (after normalizing for DNA count); direction did not matter as only the magnitude was used in this analysis. The MPRA and BiT-STARR-seq data was downloaded and pre-processed as described above. Notably, the chromatin states/DNA accessibility annotations, the HiDRA, and MPRA estimates were derived from the same cell line; that is, there is possibility of overestimating the performance of either method. For any given annotation, I assessed the predictive ability of the magnitude of the variant estimate (from any of the three approaches) to predict whether the variant overlapped with the annotation. The magnitude of the variant estimate for each approach (peaBrain, HiDRA, BiT-STARR-seq, and MPRA) corresponded to either the transcriptomic impact or activity of that variant. Only the absolute magnitude, after rank-transformation to normality, of each variant was used in modelling. For each approach and for each annotation, a logistic model was used (fitted using the glm function in R, family = "binomial") and the confidence interval was obtained using the confint function (derived from profiling the likelihood function). For the granular variant-level assessment, annotations were downloaded from RegulomeDB (dbSNP 141; see **URLs**).

simple logistic model after rank-transformation of all estimates to normality (to ensure coefficients were comparable; see **Box 3.2 Methods – Comparison of peaBrain, MPRA, BiT-STARR-seq, and HiDRA**, *on next page*). (It is important to note that my analyses are limited to the utility of these approaches in identifying and prioritizing putatively functional eQTLs; all of these experimental and *in silico* assays have other applications beyond functional eQTL discovery, such as functional enhancer analysis.) The coefficient of the model corresponded to the extent to which each approach was predictive of chromatin states/accessibility. More concretely, the logistic coefficients give the change in the log odds of the annotation overlap for a one-unit increase in the normalized score.

For peaBrain, as opposed to analysing the consequences of all possible variants/indels within 1Mbps input sequences for the "captured" 113 genes (which is very computationally expensive), I focussed my analysis on all 23,595 univariately-significant eQTLs (from either the GTEx or Geuvadis datasets). I noted that variants with higher peaBrain estimates were significantly enriched in DNase accessible sites and transcriptionally active regions, and significantly depleted from heterochromatin and repressed sequences (**Table 3.1**). In contrast, the magnitudes of the MPRA log skew estimates were not significantly associated with any chromatin state or accessibility annotation after Bonferroni correction (**Table 3.1**). This absence of enrichment/depletion was consistent whether I analysed all variants

assessed on the platform (n = 26,986 variants after excluding those with no match in Ensembl's VEP database) or limited my analysis to the subset of variants also present in the peaBrain analysis (n = 1589 MPRA variants; i.e. univariately-significant eQTLs for the 113 "captured" genes). It is important to note that variants assessed on the MPRA platform were already selected, in part, because their eQTL status in the Geuvadis dataset; that is, excluding negative controls and LD-based selection, all variants assessed on the MPRA assays were univariately-significant eQTLs.

Similarly, variants with high magnitudes of the BiT-STARR-seq log skew estimates were not significantly enriched in transcriptionally active chromatin (or depleted from repressed/quiescent intervals), irrespective of whether I assessed performance on all variants assessed on the platform (n = 43,494) or limited to univariately-significant eQTLs for the 113 "captured" genes (n = 621). Using nominal p-value thresholds, HiDRA performed better than either MPRA or BiT-STARR-seq when 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.

To ensure a fair comparison between all the methods, I did not select significantly active variants/fragments; that is, I did not apply any (significance) filter at the variant-level. 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 I 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. I 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[15-17]. 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), I was 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 – suggesting that framework is useful in identifying putatively functional variants.

Having established that variants with higher peaBrain estimates are enriched in transcriptionally active chromatin (irrespective of any variant-level filtering), I sought to subsequently evaluate the four aforementioned methods on a more granular level using RegulomeDB[60] (**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[60]. As RegulomeDB contains annotations from multiple tissues, I 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), I assessed significance using a simple logistic model after rank-transformation of all method estimates to normality (to ensure coefficients were comparable). 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. 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 3.2**). 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 "captured" 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.

Table 3.1 *(on next page)* 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 I 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-values are 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 I 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), I am 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 x10^{-5})**	0.01 (0.463)	0.21 (2.64 x10^{-2})	0.01 (0.352)	-0.05 (0.732)	**-0.05 (2.92 x10^{-13})**	0.09 (0.438)
TssA	**0.32 (<2 x10^{-16})**	0.03 (0.218)	0.25 (2.05 x10^{-2})	-0.02 (0.140)	0.01 (0.934)	0.00 (0.965)	0.09 (0.360)
TssAFlnk	0.10 (1.55 x10^{-2})	0.06 (4.72 x10^{-2})	0.29 (2.17 x10^{-2})	0.03 (4.01 x10^{-2})	0.61 (4.23 x10^{-3})	-0.02 (1.28 x10^{-2})	0.03 (0.788)
TxFlnk	-0.13 (7.69 x10^{-2})	-0.01 (0.81)	-0.16 (0.672)	0.08 (0.141)	1.61 (5.59 x10^{-2})	-0.02 (0.424)	-0.06 (0.787)
Tx	-0.02 (0.297)	-0.04 (4.91 x10^{-2})	-0.05 (0.471)	-0.14 (6.13 x10^{-3})	-0.80 (9.58 x10^{-2})	0.01 (0.571)	-0.07 (0.325)
TxWk	0.04 (1.42 x10^{-2})	-0.01 (0.662)	-0.01 (0.923)	0.01 (0.630)	0.48 (0.136)	**0.03 (4.66 x10^{-4})**	0.00 (0.973)
EnhG	0.04 (0.547)	-0.12 (6.97 x10^{-3})	-0.26 (0.167)	0.04 (0.529)	-0.91 (4.35 x10^{-2})	-0.05 (1.85 x10^{-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 x10^{-2})	**-0.03 (1.39 x10^{-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 x10^{-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 x10^{-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 x10^{-2})	NA	-0.05 (0.653)	-0.81 (0.440)	-0.03 (0.230)	-0.26 (0.735)
ReprPCWk	**-0.25 (<2 x10^{-16})**	-0.033 (0.133)	NA	-0.05 (3.73 x10^{-2})	-1.00 (6.80 x10^{-2})	-0.02 (2.45 x10^{-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 x10^{-4})**	-0.01 (0.896)

Table 3.2 In the 113 "captured" genes, peaBrain estimates can significantly delineate variants with established regulatory function (**Task H**). The log-skew estimates from the experimental assays, both across all variants assessed on each platform ("all") and limited to eQTLs for the 113 "captured" genes ("shared"), are uninformative. Both the peaBrain estimates and log-skew for the experimental assays were rank-transformed to normality to facilitate comparison between the methods. 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.

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

3.3 peaBrain discussion and conclusions

In this chapter, I have introduced a general computational framework for recapitulating high-throughput expression-based assays, by emulating the transcriptional machinery of human tissues/cell lines.

As a DNA reporter assay, the computational framework revealed 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, the average expression of all genes in a single tissue and the reference genome is sufficient to learn both TFBS and allele-specific binding. Taken together, this is broadly consistent with anecdotal experimental evidence[43] and suggests that non-transcriptional downstream processes play a secondary role in regulating mean expression.

The predictive ability of the peaBrain DNA reporter assay 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.

With a simple extension of the DNA reporter assay to incorporate small molecule structures (or shRNA oligo-sequences), I am able to accurately predict the perturbation of expression with performance as good as *in vitro* experimental replicates. I show that the peaBrain imputed expression is sufficiently informative for drug repositioning, and more so than classical cheminformatics/structure-based approaches. It is important to note that these results described in this chapter do not represent an upper bound to performance. There are several ways to improve the peaBrain model, including: (a) incorporating additional information on the cell line (such as whole-genome sequencing and epigenetic modifications of the DNA sequence); (b) choice of fingerprint used to encode molecular structure; and (c) designing a training set that more accurately reflects the "molecule space" in which I was interested (e.g. by intelligently designing which molecules the peaBrain algorithm is exposed to).

To incorporate the consequences of individual variation on gene abundance, I extended the peaBrain DNA reporter assay to capture a 1Mbps window, a balance between computational tractability and biological "signal". Unlike the small molecule screen, this extension of peaBrain leverages differences in the sequences between individuals to predict differences in expression (rather than prediction from the sequence directly); 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. In other words, it is not yet possible to readily extract meaningful motif sequences from the input data. Reconstruction of the individual sequences to generate the difference input required that I use a quality controlled VCF to reconstruct individual sequences, 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, the peaBrain model 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.

It is important to note that, unlike many methods with similar conceptual origins (e.g. prediXcan), the individual variation peaBrain model was not designed with the sole intent of predicting gene expression. Rather, one of the primary goals of individual variation peaBrain models is identifying putatively functional eQTLs; for a substantial increase in computational cost, I get much higher variant sensitivity (i.e. peaBrain framework is not yet suitable for cohort imputation). To a first approximation, I observed 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 were not 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. By comparing across all variants (without any significance filtering), I 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.

All together, these results suggest that peaBrain framework is useful understanding the effects of biological entities on the transcriptome – as powerful as and more inexpensive than its experimental counterparts. peaBrain's performance in predicting mean abundance and individual variation further implicates the importance of the invariant genomic context and distance to the annotated TSS for interpreting the effects of non-coding variation and small molecules in a tissue-specific manner. Importantly, by "emulating" the transcriptional machinery of human tissues/cell lines, the results also suggest that the expression effects of small molecules, shRNA oligo-sequences, and non-coding variation can be built by relying more on automated machine learning, rather than hand-designed or selected features.

Chapter 4: Differential Dependence to Functionalize GWAS Signals

Integrating molecular and clinical data that underlie complex phenotypes in a tissue and cell-type dependent manner remains an unsolved problem[1]. This lack of integration is a hurdle to subsequently translating molecular discoveries into functional and clinical understanding of disease pathophysiology[2,3]. With the recognition that genome sequence information alone will provide only limited specification of disease development and progression[4,5], it is necessary to develop and formalize a systematic approach to combine sequence variation data with molecular signatures and clinical phenotypes[6]. One possible solution to this translation hurdle is the development and implementation of machine learning (ML) algorithms that readily incorporate multi-omic and clinical data types to cluster, classify and regress against clinical outcomes[7-9]. (ML approaches that incorporate multiple data types are termed multi-view algorithms.) The notable advantage of such an approach is drawing on the extensive theoretical and practical justifications associated with these ML algorithms[9,10]. A typical ML workflow for clinical discovery can be divided into two steps:

(a) dimensionality reduction within and across molecular data types and "appropriate" feature selection/extraction; and

(b) clustering or classification/regression against clinical outcomes, adjusting for technical confounders and conflating biological covariates.

One of the most popular unsupervised ML techniques applied to molecular data is clustering[11], i.e. the unsupervised subdivision of biological entities (e.g. genes) and/or samples such that similar items fall into the same group. Clustering can follow classical dimensionality reduction steps (e.g. principal component analysis)[12] or can be upstream of more sophisticated (semi-) supervised ML approaches (e.g. as an initial seed)[13]. With the combinatorial possibilities for similarity metrics and clustering approaches[11], there is a bewildering variety of options. True clusters are recovered only when the underlying data conforms to the biases and requirements imposed by the clustering algorithm[11]. To date, there exists no (exhaustive) comparison of clustering methods with respect to performance or with respect to sensitivity to parameter selection and common normalization/preprocessing workflows [14-16].

In part, this is because there is no consensus on what a "good" clustering looks like[11,14,15]. Here, by focusing on disease-specific study (e.g. type 2 diabetes [T2D]) and select molecular data types, the problem is narrowed to how well the clustering algorithms perform in the context of clinical interpretability (e.g. the underlying T2D-relevant biology of the clusters) and of performance in more sophisticated downstream modelling (i.e. the practical utility of the clusters). Furthermore, in a disease- and tissue-specific context, it becomes possible to meaningfully grapple with the question of what "good" clustering looks like and the "appropriateness" of common preprocessing and transformation procedures. Furthermore, with multiple data types, I can begin to formalize the physiological relevance of these clustering techniques (e.g. by exploring the heritability of coexpression between genes or the interaction dependency between a pair of metabolite clusters) and associating latent information/variables (underlying clusters across multiple molecular data types) with clinical outcome. **Table 4.1** summarizes some challenges and existing unsupervised clustering solutions.

Table 4.1 A summary of some unsupervised clustering approaches commonly applied to molecular clinical datasets. The "app" column includes only 1-2 applications of the method to a complex molecular and/or clinical datasets. None of the approaches discussed here have been implemented for multi-view datasets. **NB:** The term "graph" in this chart denotes a network graph (i.e. a set of vertices/nodes connected by edges).

Challenge	Approach	Description	Ref	App
Classical Approach	Hierarchical Clustering	Most popular clustering approach[61]	62	62
Classical Approach	K-Means Clustering	Partition a set of observations into K clusters while minimizing the within-cluster sum of squares (for the distance metric)	63	64
Classical Approach	Self Organizing Maps	Transform an arbitrarily sized input matrix into a 2D discrete map using competitive learning	65	66
Hard thresholding (i.e. dichotomizing gene correlation) leads to loss of information	Weighted Gene Coexpression Networks	'Soft' thresholding (i.e. exponentiated correlation matrix) with weighted topological overlap measure for hierarchical clustering.	67	68
Need rigorous framework to assess the appropriate number of clusters and the quality of clusters obtained	Gaussian Mixture Models	Observations are a sample of unknown probability distribution that can be estimated by a finite mixture of Gaussians.	69	70
Need to simultaneously clusters genes and conditions (to find distinctive "checkerboard" patterns)	Spectral Co-Clustering (Spectral Bi-clustering)	Clusters rows and columns of a matrix to solve the relaxed normalized cut of the bipartite graph created from the matrix (edge weight corresponding to the matrix entry between the column and row)	71	72
Need for better and more efficient clustering performance than with k-means algorithm	Affinity Propagation	Message-passing based clustering algorithm	73	74

4.1 Overview of the Clustering Algorithms

As evident from earlier chapters, the most common molecular data type available to medical researchers, excluding genome-wide association studies (GWAS), is expression data[75]. The basis of this new "differential dependence" strategy required clustering of expression data. Thus, to gain a better idea of how to define the comparison framework for dimensionality reduction/clustering approaches, I sought to evaluate whether existing workflows could robustly define gene sets of co-regulated gene in transcriptomic RNA-seq data (using correlation similarity metrics). For this section of my thesis, I

elected to focus specifically on methods that use a correlation similarity for clustering. My study was sub-divided into two stages:

(a) a coarse filtering stage using an internal cluster metric for comparison; and

(b) an in-depth second stage to assess robustness to sample variability and *in silico* validation using functional annotations.

My study compared 3 "current" coexpression pipelines (WGCNA[76], ARACNe[77], and PCIT[78]) and 1 "new" method that I developed (a fine-tuned affinity propagation algorithm; a message-passing based approach, termed xMAP; see **Box 4.1: Methods – Clustering Algorithms [algorithm details are provided further below in the text]**) in three different tissues: whole blood, fat, and skin. This analysis used data from the EuroBATS Consortium (i.e. RNAseq datasets from the Twins UK cohort). The primary motivation for using this dataset was the number of patients (this represented one of the largest cohorts with [multiple] matched tissues and genotype samples of 449 patients)[66] and the inclusion of mono- and di-zygotic twins for subsequent applications and heritability analysis.

The first coarse filtering stage involved a silhouette coefficient analysis (SCA)[79] of the generated clusters (for each method). In SCA, for each gene, I calculated the scaled difference between the average intra-cluster distance and the average nearest-cluster distance (using a cosine similarity metric[67]). The silhouette coefficient from SCA has three notable distributions:

- A silhouette coefficient of 1 (maximum best score) indicates well-separated and cohesive clusters.

[66] To avoid circularity, I couldn't use the GTEx dataset because that is where the eQTL data (for subsequent analyses) was derived from.

[67] As mentioned in an earlier chapter, 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.

[67] This is the first application of peaBrain when trained on a dataset other than GTEx. All prior tasks assessed performance of results derived from the GTEx-trained models. The task described in Chapter 2.6 highlights the utility of the architecture on datasets that are significantly smaller (i.e. with as little as 2 samples per tissue) and the utility of comparing peaBrain model predictions as an alternative (and more powerful) approach to identifying differentially expressed transcripts.

- Negative values (minimum of -1) indicate wrong gene-to-cluster assignment (i.e. there is a different cluster to which the gene is "closer").
- Coefficients near 0 indicate overlapping clusters.

Averaging the silhouette coefficient for all genes (or looking at other central tendency measures) allows for a global perspective on the separation and density of the clusters. Despite the diversity of approaches, all 3 "current" coexpression pipelines generated poorly defined and incoherent clusters (as measured by SCA): WGCNA, ARACNe, and PCIT presented with silhouette coefficients in whole blood of -0.72, -0.87, and -0.40, respectively (**Figures 4.1**). A negative coefficient (a minimum of -1)

Box 4.1 Methods – Clustering Algorithms

WGCNA. Method was implemented using WGCNA_1.51 (CRAN package). All RNA-seq matrices (with FPKM values) were log2 transformed. Soft threshold was automatically selected using the pickSoftThreshold function; if NA was returned, softPower was arbitrarily set to 6 (as suggested by authors). Modules were identified, in blocks, using the blockwiseModules function, with default values: maxBlockSize = 2000, TOMType = "signed", minModuleSize = 30, reassignThreshold = 0, and mergeCutHeight = 0.25. These are the recommended parameters suggested by the authors.

xMAP. Method was implemented using sklearn 0.17.1 (installed with Anaconda 4.0.0). RNAseq (RPKM) data was log2 transformed first. Data was then centered to the mean and component wise scaled to unit variance. The number of features was then reduced with a principal component analysis; Minka's MLE was used to select the number of components. Affinity propagation was then applied using a cosine similarity matrix (i.e. the L2-normalized dot product of gene vectors). The number of iterations was set to 2000, with 150 iterations for convergence and a damping = 0.5. For the subsampling, clustering that did not converge (evident with the number of iterations = 2000 and the absurdly large number of clusters) were excluded from subsequent analyses.

ARACNE. Method was implemented using minet_3.28.0 (Bioconductor package). The mutual information matrix (MIM) for input was generated using the build.mim function (estimator = "spearman"). The eps threshold (for the data processing inequality [DPI]) was set to 0. This specified the minimum edge for each node triplet was always removed. Each network was subsequently used as the input matrix for the fast.greedy algorithm (from igraph_1.0.1 [from CRAN]). The networks did not converge, in a reasonable amount of time, using affinity propagation or correlation clustering. TOM/hierarchical (from WGCNA) couldn't be implemented as edge weights could be greater than 0.

PCIT. Similar to ARACNE, the MIM was generated using the build.mim function from minet_3.28.0. The matrix was fed into the default PCIT function from PCIT_1.5-3 (available on CRAN).

Silhouette analysis. The Silhouette Coefficient is calculated using the mean intra-cluster distance (I) and the mean nearest-cluster distance (N) for each gene. The Silhouette Coefficient for each gene is thus:

$$\frac{(\mathrm{I} - \mathrm{N})}{\max(\mathrm{I}, \mathrm{N})}$$

Coefficients and plots were generated using the silhouette_score and silhouette_samples from sklearn 0.17.1. Coefficients range between -1 and 1 (inclusive), with a score of 1 denoting well-defined (i.e. dense and well-separated) module definitions and a score of 0 indicating module overlap. A -1 denotes a gene was clustered in the wrong module (as another module has a closer or more similar expression profile).

for a given gene indicates that a different cluster is closer (i.e. has a higher correlation). Similar results were observed for fat and skin (all negative coefficients; **Figures 4.2 and 4.3**). This meant that many genes in the clusters were incorrectly assigned based on the correlation similarity metric. In other words, many genes were closer to other clusters than the one they were assigned by the clustering algorithm.

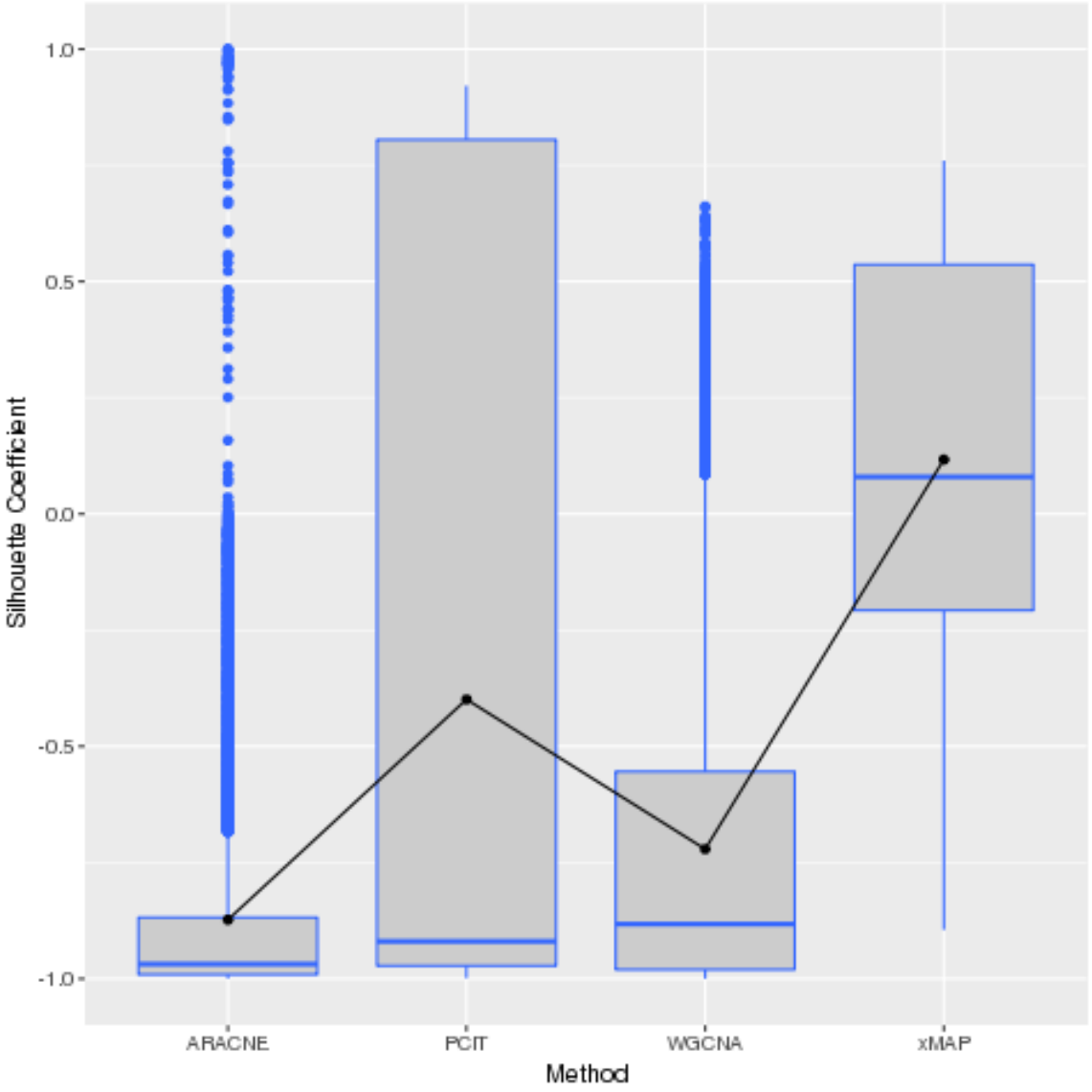


Figure 4.1 Silhouette coefficient analysis boxplots, using whole blood expression matrix, for existing coexpression pipelines, from right to left (alphabetical order): ARACNE, PCIT, WGCNA, and xMAP. A more positive silhouette coefficient is better, indicating dense and well-separated modules. A negative coefficient score (minimum of -1) indicates another module closely resembles the gene (i.e. the gene was clustered into the wrong module). A silhouette coefficient of 0 indicates overlapping modules. The average silhouette coefficient (a global metric of pipeline performance) is indicated by the solid black line connecting the boxplots. All 3 current coexpression pipelines present with average (and median) negative silhouette coefficients. xMAP outperforms all 3 methods, as depicted in the right most boxplot. Silhouette analysis for adipose and skin are depicted in **Figures 4.2 and 4.3** respectively.

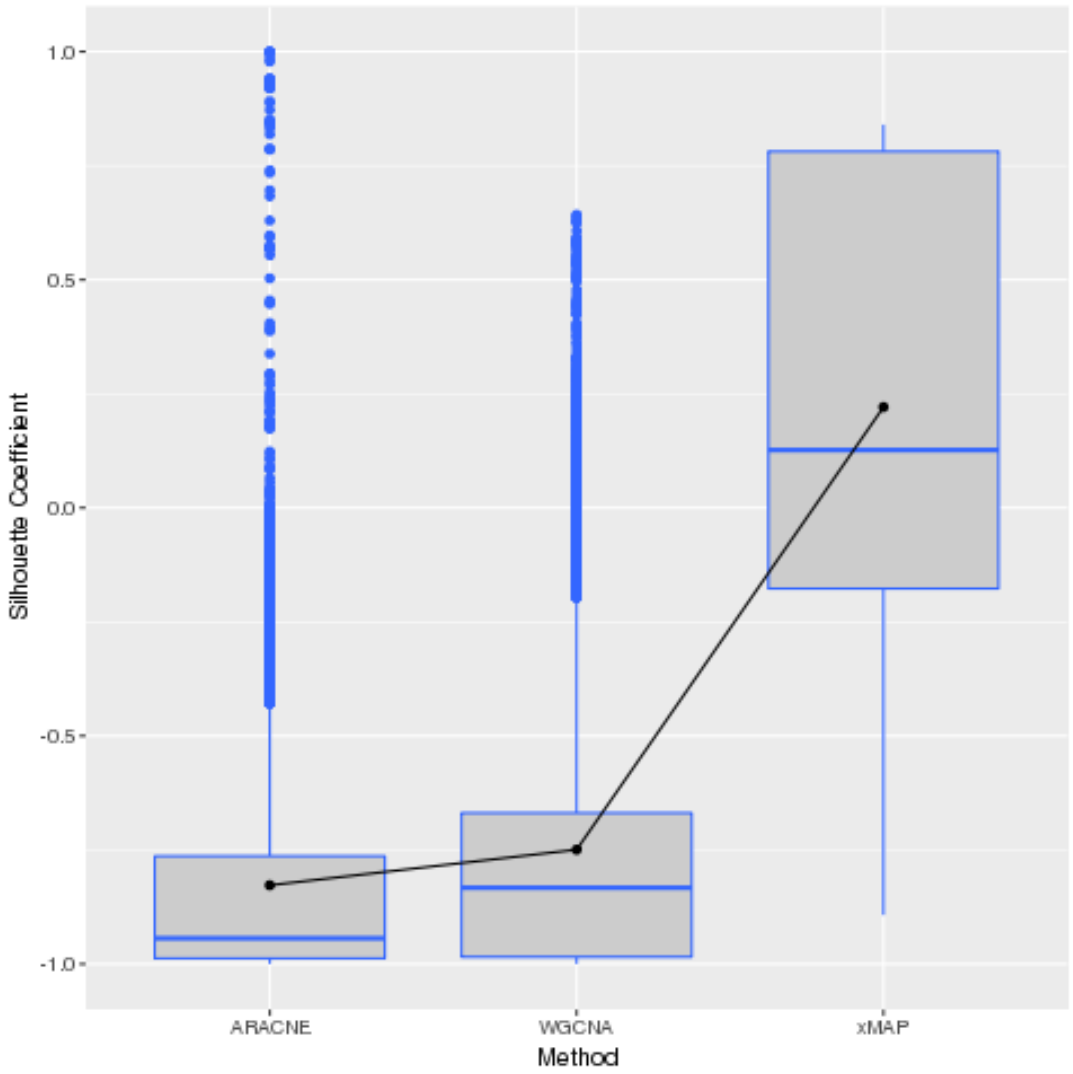


Figure 4.2 Silhouette coefficient analysis boxplots, using adipose expression matrix, for existing coexpression pipelines, from right to left (alphabetical order): ARACNE, WGCNA, and xMAP. PCIT was unfeasible to run with adipose (nearly double the number of samples compared to the whole blood matrix analysis depicted in Figure 1). Both ARACNE and WGCNA coexpression pipelines present with negative average (and median) silhouette coefficients. xMAP outperforms both methods, with positive average and median silhouette coefficient, as depicted in the right most boxplot.

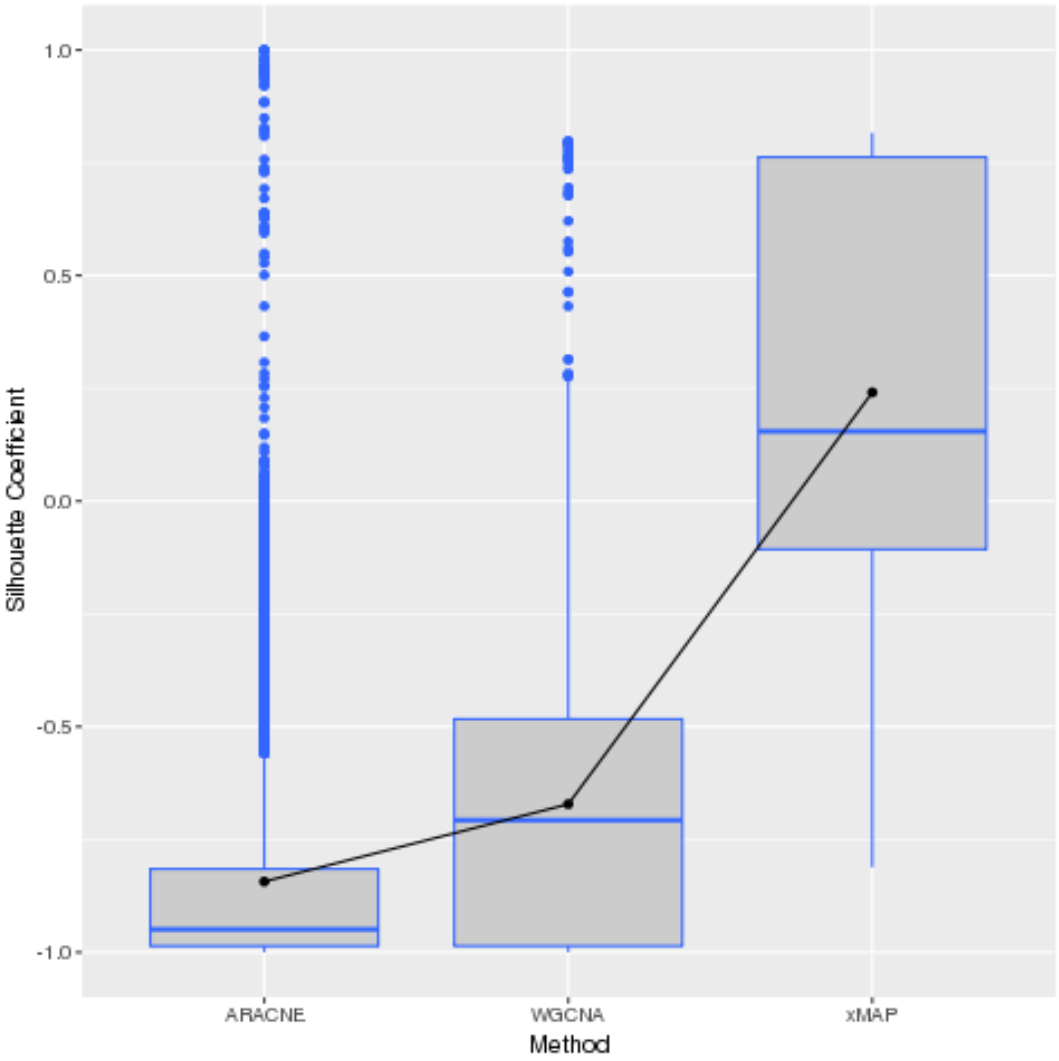


Figure 4.3 Silhouette coefficient analysis boxplots, using skin expression matrix, for existing coexpression pipelines, from right to left (alphabetical order): ARACNE, WGCNA, and xMAP. As with adipose, the memory requirements were too great to run PCIT. Both ARACNE and WGCNA coexpression pipelines present with negative average (and median) silhouette coefficients. xMAP outperforms both methods, with positive average and median silhouette coefficient, as depicted in the right most boxplot.

The poor performance of "current" methods served as the motivation to implement a new coexpression pipeline to allow for more robust cluster definitions: xMAP. By sending messages between pairs of genes, xMAP allows for a small number of "cluster center" genes to be selected as representative profiles for all other genes (i.e. archetype-based clustering). There is evidence that message-passing based algorithms outperform K-means clustering[73], which in turn outperforms hierarchical clustering (the core element of the WGCNA algorithm)[61,80,81]. This xMAP approach, which is an extension of

affinity propagation, presented with a higher silhouette coefficient than current methods: 0.12 in whole blood (**Figure 4.1**), 0.22 in adipose (**Figure 4.2**), and 0.24 in skin (**Figure 4.3**).

With xMAP and any extension of the original affinity propagation algorithm, the genes can be interpreted as nodes in a network that send messages sent to all other genes/nodes (until some convergence criteria is reached). Conceptually, these messages indicate the willingness of each gene to being an "exemplar" or "cluster center" genes (as aforementioned). The goal of xMAP is to collectively determine the "cluster center" genes for all cluster. More concretely, the messages sent between nodes/genes are stored in two matrices:

(a) a responsibility matrix, *R*, where *r(i, k)* indicates the suitability of each gene *k* is to be the "cluster center" for gene *i*; and

(b) an availability matrix, *A*, where a(i, k) indicates the appropriateness of each gene *i* to choose gene *k* as its "cluster center".

Importantly, matrices can be interpreted as log probabilities and are thus negative valued. Formally, the responsibility of gene *k* to be the "cluster center" of gene *i* is given by:

$$r(i,k) \leftarrow s(i,k) - \max[a(i,k') + s(i,k') \; \forall \; k' \neq k]$$

where *s(i, k)* is the similarity metric between genes *i* and *k*; in this context, the similarity metric is simply the correlation to facilitate comparison with all other existing clustering algorithms (which all use correlation-based metrics). The availability of gene *k* to be "cluster center" for gene *i* is given by:

$$a(i,k) \leftarrow \min[0, r(k,k) + \sum_{i' \; s.t. i' \notin \{i,k\}} r(i',k)]$$

At the very start, all matrices are initialized to zero. Each iteration the above calculations for responsibility and availability are made until convergence is reached. To avoid oscillations and ensure stability, there is a damping factor (introduced as λ by the authors of affinity propagation):

$$r_{t+1}(i,k) \leftarrow \; \lambda * r_t(i,k) + (1 - \; \lambda) * \; r_{t+1}(i,k)$$
$$a_{t+1}(i,k) \leftarrow \; \lambda * a_t(i,k) + (1 - \; \lambda) * \; a_{t+1}(i,k)$$

where *t* indicates the number of iterations. Unsurprisingly, because of these computations, xMAP has a time complexity of $O(N^2T)$, where N is the number of genes and T is the number of iterations until convergence. However, on a regular compute cluster, these computational costs are relatively trivial.

For the second stage, I sought to definitively assess the numerical stability and biological relevance of the clusters defined using the xMAP pipeline (best performing method) and WGCNA (the 3rd best performing algorithm among the four). PCIT, the second-best performing pipeline, was unfeasible to run through my test framework (requiring >36GB of memory and a >2-day run time for 1 analysis). I defined numerical stability as the "robustness" of the method with respect to sample and population variability. For example, there is an expectation that robust cluster definitions should remain stable across sets of patients/samples with similar characteristics (e.g. clustering defined using healthy whole blood tissue from a European cohort should be similar regardless of the individuals that make up the input set). To assess numerical stability of the approach, I sub-sampled (without replacement) each of the aforementioned three (tissue) transcriptomic matrices 100 times for each fraction: selecting 30, 50, 70, and 90 percent [of samples]. Each subsampled matrix was subsequently run through the xMAP and WGCNA pipelines, generating coexpression clusters. I compared each coexpression pipeline using five different metrics (see **Box 4.2**):

(i) internal stability using the adjusted rand index (ARI)[82];

(ii) internal stability using adjusted mutual information (AMI)[83];

(iii) external stability with AMI;

(iv) external stability using ARI; and

(v) comparison of cluster size/counts.

Both the ARI and AMI scores are chance-adjusted measures of the similarity between two partitions/clusterings (*i.e.*, adjusted for differences in the number of clusters). Briefly, ARI is equivalent to 0 if the gene assignments (to the clusters) are random[82]. A perfect match score (i.e. identical clustering) has a positive ARI of 1.0. AMI is conceptually similar[83]: a mutual information score normalized against chance with 0 indicating random clustering and 1.0 indicating identical clustering. Importantly, neither score makes any assumptions on the clustering structure[82, 83]. Furthermore, both

scores are also unaffected by the number of clusters and number of genes[82, 83]. Internal stability reflects how similar the clustering within each subsampling fraction is; external stability is a measure of how similar the clustering (at any particular subsampling fraction) is to the final cluster definitions generated using the complete matrix.

Box 4.2 Methods – Internal and external stability metrics

Numerical stability of xMAP and WGCNA. To assess numerical robustness, internal and external stability were quantified using the adjusted_mutual_info_score and adjusted_rand_score functions (from sklearn 0.17.1). For any two clusterings, U and V, adjusted mutual information (AMI) is calculated by:

$$\mathrm{AMI}(U, V) = \frac{[\mathrm{MI}(U,V) - \mathrm{E}(\mathrm{MI}(U,V))]}{[\max(\mathrm{H}(U),\mathrm{H}(V)) - \mathrm{E}(\mathrm{MI}(U,V))]}$$

where MI(U, V) denotes the mutual information between the two clusterings U and V; H(U) denotes the entropy associated with the partitioning U is; and E(MI(U, V)) is the expected mutual information between two random clusterings under a hypergeometic model of randomness. Thus AMI accounts for the fact that mutual information is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared. Similarly, for any two clustering, U and V, adjusted rand score is calculated by:

$$\mathrm{ARI} = \frac{(\mathrm{RI} - \mathrm{Expected\_RI})}{(\max(\mathrm{RI}) - \mathrm{Expected\_RI})}$$

where RI is the rand index, represented as follows:

$$\mathrm{RI} = \frac{(a+b)}{\binom{n}{2}}$$

where a is the number of pairs of genes that are in the same set in U and in the same set in V and b is the number of pairs of genes that are in different sets in U and in different sets V. The Expected_RI is calculated from the row and column sums of the contingency table, summarizing the overlap between the U and V clusterings. Each entry, n_{ij}, in the table is the number of genes overlapping between module (row) i from U and module (column) j from V; the row sums are denoted by a and column sums by. Thus,

$$\mathrm{Expected\_RI} = \frac{\sum_i \binom{a}{2} \sum_j \binom{b}{2}}{\binom{n}{2}}$$

Biological relevance of xMAP and WGCNA. Ensembl genes to GO term mappings (from the Homo sapiens genes [GRCh37.p13] dataset) were downloaded from the GRCh37 Ensembl biomart. CORUM core complexes were downloaded from the CORUM database, freely available here: http://mips.helmholtz-muenchen.de/genre/proj/corum/. Purity fractions were calculated using Python and visualized in R.

Initially, I was interested in the internal stability of both pipelines: how stable are the cluster definitions between all sub-sampled matrices within each sub-sampling fraction? More concretely, as an example, how similar are clusters generated using any 2 matrices containing a random 50% of all samples? I noted that:

- xMAP outperforms WGCNA using for all pairwise internal comparisons (internal denoting within subsampling fraction), across all tissues (**Figures 4.4-4.6**).
- In whole blood, even with only 30% of samples, xMAP exhibits greater internal stability than WGCNA (median AMI = 0.65; WGCNA median AMI = 0.35; **Figure 4.4**).
- With 90% of samples, xMAP remains much more stable (median AMI = 0.82), although the difference is smaller (WGNCA median AMI = 0.66; **Figure 4.4**). This difference is also apparent in the 2 other tissue comparisons (**Figures 4.5 and 4.6**).

Subsequently, I was interested in the external stability of the WGCNA and xMAP pipelines: how similar are clustering at each sub-sampling fraction to the cluster definitions using the complete matrix (i.e. 100% of samples)? I noted that:

- Across all tissue, xMAP strongly outperformed WGCNA (**Figures 4.4-4.6**).
- In whole blood, xMAP had a median AMI at 90% = 0.84 compared to WGCNA median AMI = 0.66 (**Figure 4.4**). Remarkably, for xMAP, cluster definitions are highly similar to each other at all sub-sampling fractions (median AMI at 30% = 0.67; WGNCA at 30% = 0.31; **Figure 4.4**).
- These observations were consistent for adipose and skin as well (**Figures 4.5 and 4.6** respectively). In other words, xMAP was robust with respect to gene assignment to clusters, even with relatively small numbers of samples, across all 3 tissues.

Using the complete matrices (i.e. all patients in the dataset):

- WGCNA identified 18 clusters for blood, 39 for fat, and 41 for skin. xMAP identified a larger number of clusters: 132 for blood, 113 for fat, and 143 for skin. Future work will include

exploring what happens if both methods (WGCNA and xMAP) are fixed to have the same number of clusters.

- For WGCNA, across all 3 tissues, median cluster size ranged from 107-153; for xMAP, the median ranged from 30-42 genes.

Across all three tissues, I also noted that xMAP was not sensitive to the ages of the patients that were used for clustering. Using various cut-off age ranges (across the 40-89 year span of the individuals in the EuroBATS cohort), I noted AMI and ARI scores fell within the range of expected values of the corresponding random sub-sampling fraction (**Figures 4.4-4.6**). In other words, the coxpression clusters defined using a cohort selected based on age (e.g. 40-60 years of age) does not differ from clusters defined using a random cohort of equal size. On the same datasets, the median proportion of variation in expression explained by age was reported to range between 2.2% and 5.7% depending on tissue[84].

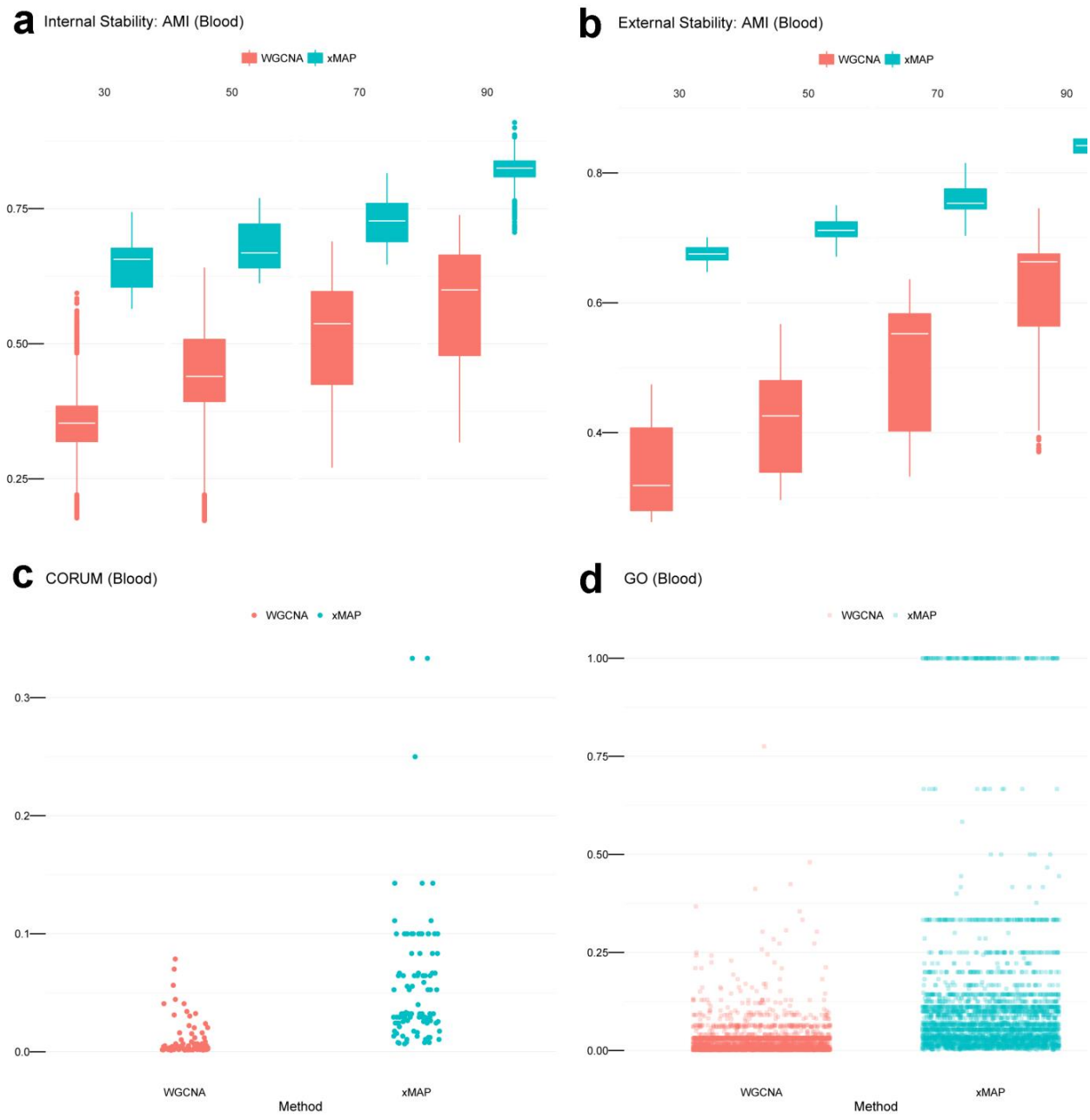


Figure 4.4 xMAP-defined modules demonstrate greater numerical robustness and biological, in *whole blood*, compared to WGCNA. xMAP module definitions exhibited greater **(a)** internal stability and **(b)** external stability than WGCNA module assignments, as defined by AMI. As highlighted in text, an AMI of 1 denotes matching, perfect clustering; an AMI of 0 indicates random gene-to-module assignment. AMI is independent of the number of modules; accounting for variation in module sizes. xMAP module definitions also highlight biologically-more relevant gene sets for both **(c)** manually curated core CORUM protein complexes and **(d)** GO terms. A higher fraction score is better, denoting a larger portion of the corresponding module containing genes from the term/complex in question.

Figure 4.5 xMAP-defined modules demonstrate greater numerical robustness and biological, in *adipose*, compared to WGCNA. xMAP module definitions exhibited greater **(a)** internal stability and **(b)** external stability than WGCNA module assignments, as defined by AMI. xMAP module definitions also highlight biologically-more relevant gene sets for both **(c)** manually curated core CORUM protein complexes and **(d)** GO terms. A higher fraction score is better, denoting a larger portion of the corresponding module containing genes from the term/complex in question.

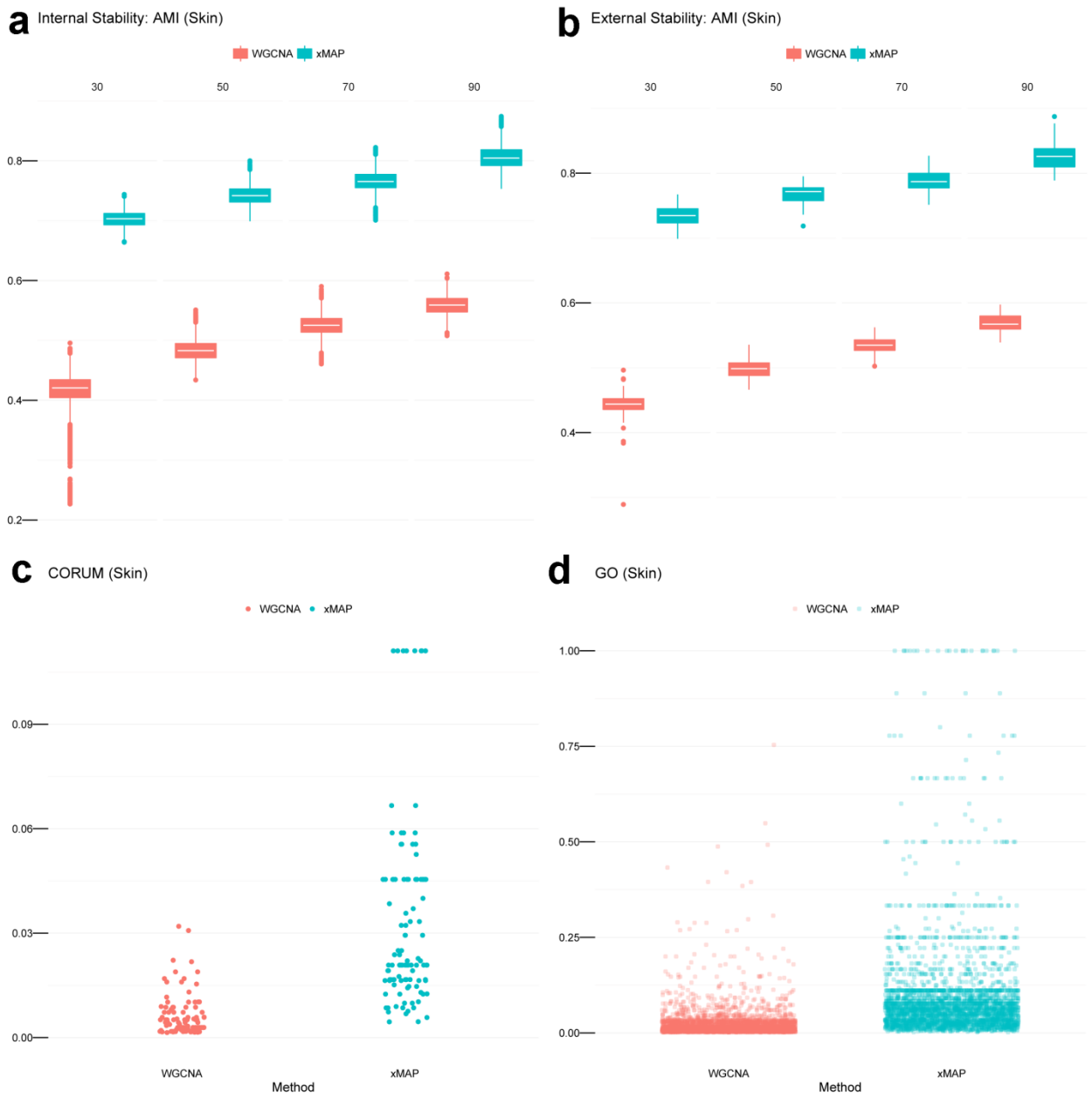


Figure 4.6 xMAP-defined modules demonstrate greater numerical robustness and biological, in *skin*, compared to WGCNA. xMAP module definitions exhibited greater **(a)** internal stability and **(b)** external stability than WGCNA module assignments, as defined by AMI. xMAP module definitions also highlight biologically-more relevant gene sets for both **(c)** manually curated core CORUM protein complexes and **(d)** GO terms. A higher fraction score is better, denoting a larger portion of the corresponding module containing genes from the term/complex in question.

Having demonstrated numeric stability, I was interested in assessing biological relevance using two metrics:

(a) How many genes in a cluster belonged to a single core protein complex (from the CORUM database)? This was the "complex purity fraction" of the genes in the cluster. CORUM is "a resource of manually annotated protein complexes from mammalian organisms… All information is obtained from individual experiments published in scientific articles, data from high-throughput experiments is excluded."

(b) How many genes in a cluster belonged to a single gene ontology term (from the GO database)? This was the "GO term purity fraction" of the genes in the cluster.

For each complex or GO term:

1. I identified the cluster containing the largest number of genes from that complex or term.
2. I then calculated the purity fraction = $\frac{\text{Number of genes in complex or GO term (in cluster)}}{\text{Total size of cluster}}$

Using both metrics (CORUM and GO terms), xMAP clusters exhibited greater cluster purity, i.e. larger fractions, than WGCNA (**Figures 4.4-4.6**). In particular:

- In whole blood, for GO terms, xMAP had a median purity fraction of 0.05 (WGCNA median purity faction = 0.01; **Figure 4.4**). There were 3082 (93%) GO Terms (total = 3304) that had a purity fraction >1% using xMAP, compared to only 1646 (50%) terms when using WGCNA.
- Genes belonging to manually-curated core protein CORUM complexes were more likely to be placed in the same cluster, than in separate clusters, with xMAP (median purity fraction = 0.03) compared to WGCNA (fraction = 0.003; **Figure 4.4**).

Across all 3 tissues, xMAP clusters tended to exhibit greater biologically consistency, with genes in the same cluster more likely to share similar function, cellular localization, and molecular function (**Figures 4.4-4.6**). After adjusting for expressed genes (RPKM > 10), this difference was consistent. Since both biological-relevant metrics favoured smaller cluster definitions, I also calculated ARI and AMI scores between the final cluster definitions generated by the coexpression pipelines and a coarse clustering

based on the GO terms (i.e. genes with the same set of GO terms were placed in the same cluster); xMAP outperformed WGCNA with this metric as well. As aforementioned, the ARI/AMI scores are adjusted for chance and account for variation in the number of clusters - thus, xMAP clusters, by definition, contain more information about biologically relevant gene sets than does WGCNA.

Why does WGCNA and other "state-of-the-art" approaches underperform? The methods do not work on "real data" because they are not robust to noise (e.g. subsampling the individuals in a dataset will change the clusters). These methods were crafted under ideal conditions, with limited samples and/or simulated data. No particular flaws stand out for any of the methods. These methods were able to get published for one single reason: If you sub-divide genes into any number of arbitrary reasonably-sized clusters and select the cluster that most correlates with a phenotype/trait of interest, you are enriching that cluster for biological signal. These methods can do that – any non-random, partially-informed algorithm can do that. However, a good clustering algorithm should be robust to noise in the data, because there is inherent biological value to the clusters themselves (independent of correlation with a phenotype/trait of interest).

4.2 xQTLs are variants that strongly associate with natural variation in cluster interactions

Having developed a robust workflow for clustering RNA-seq, I was interested in whether variants can explain natural variations between cluster interactions. In other words, can I explicitly associate genetic variants with the "statistical interaction" (e.g. correlation) between clusters in the various tissues? Identifying variants that influence the interaction between robust gene clusters would allow me to formalize many of the expression-based network analyses in a genetic framework. More practically, the association of these xQTL variants with transcriptomic patterns and perturbations would allow me to annotate GWAS-nominated variants with possible functional effects. Although xQTLs represent an association of SNPs with transcriptomic patterns, they are distinct from eQTLs (loci that contribute to variation in expression levels)[85,86] and modQTLs (loci that contribute to deviation from ideal gene

expression profiles)[1]. (At least 42% of modQTLs are directly explained by both tissue-specific and multi-tissue eQTLs[87].) Furthermore, unlike all other transcription-focused association studies, xQTLs require multiple (i.e. repeat) time point measurements from a single individual. To this end, I leveraged the presence of monozygotic twins (as a surrogate for two "time point" measurements of a genotype) from the aforementioned EuroBATS dataset to assess differential xQTL discovery among the three tissues for which data was available (blood, skin, and adipose). It is important to emphasize that there were no repeat measures for any individual; all analyses were conducted with genetically-identical (i.e. monozygotic) twin pairs. At two "time points", any pair of clusters can be visualized as 2D vectors, with the angle between the clusters proportional to the correlation/linear dependence of the two clusters (**Figure 4.7**). I was limited to linear dependence as only two "time points" were available per twin pair (with one "time point" corresponding to a single twin). With the availability of multiple time points for a subset of this cohort as part of the MultiMuther consortium, it will be possible to extend this approach. I calculated xQTLs separately for the 3 tissues using xMAP defined clusters. The angle/correlation/linear dependence measure between any pair of clusters was rank-normalized and ran through Kruskal-Wallis tests (ANOVA model) with Matrix eQTL[88]. xQTLs were discovered using a standard permutation scheme and corrected for multiple testing using Storey's FDR at 5% (lambda = 0 for conservative Benjamini-Hochberg adjustment)[68].

[68] The pipeline is analogous to the GTEX workflow eQTL calculations. Briefly, I calculated the minimum pvalue for a given module interaction across all SNPs. I scrambled expression and covariate labels n = 1000 times as part of a permutation scheme to define the empirical distribution with the minimum pvalue for every interaction. For each pair of modules (i.e. interaction), I defined the empricial pvalue as the rank of the observed in the permutation list divided by 1000. The empricial pvalues were used as input to qvalue_2.2.2 (lambda = 0 for the conservative Benjamini-Hochberg adjustement) to estimate FDR and to obtain a permutation threshold (defined as the empirical p-value of the gene that falls on the q-value = 0.05 threshold). This threshold was subsequently used to select significant xQTL-associated SNPs for each interaction.

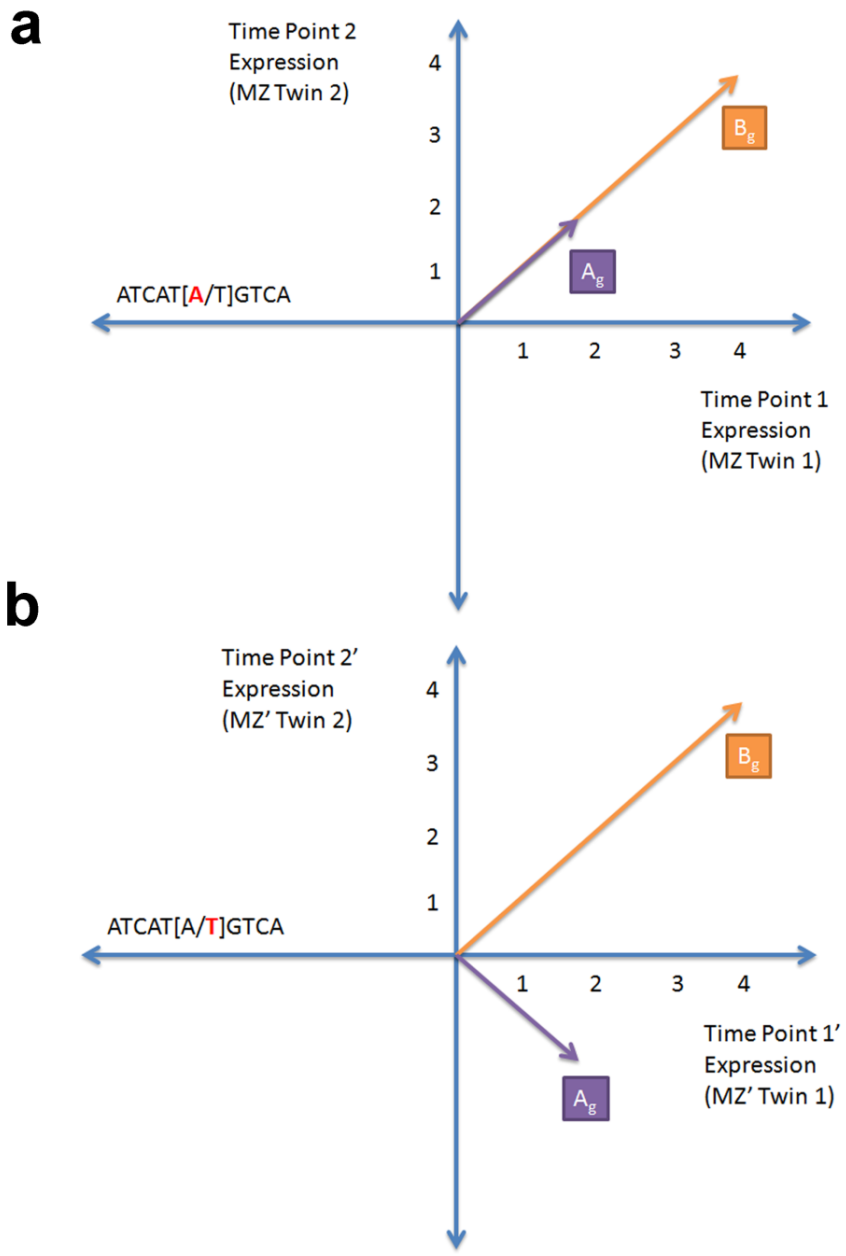


Figure 4.7 Schematic illustration of the xQTL definition. Expression axes are defined with arbitrary units. It is worth-while to note that monozygotic twins are surrogate for replicate or repeat measurements of a single individual (i.e. a pair of monozygotic twins can be "interpreted" as repeat measurements for the same genotype). **(a)** Given two clusters (e.g. clusters A and B with genes A_g and B_g as their respective centers), if I obtain "two time point" measurements (i.e. measurements from a pair of genetically-identical/monozygotic twins), I can define vectors for both genes A_g and B_g (e.g. gene A_g = [1, 1] and gene B_g = [4, 4]). The first entry of the vector for both genes denotes "time point" 1 (expression in the first twin of the pair; "x-axis") and the second entry denotes "time point" 2 (or the second twin of the pair; "y-axis"). I can subsequently calculate the angle between the two clusters and find genetic variants that associate with this natural variation in linear dependence. Here the angle between clusters A and B is 0. **(b)** With a different allele at this xQTL variant, I find that the angle between clusters A and B is 90°. Thus, in this simple and idealized example, the noted xQTL variant directly influences the correlation between modules. **Abbreviations:** Monozygotic (MZ).

In whole blood, I identified 8369 interactions with at least 1 xQTL (total number of interactions = 8646); 6149 in adipose (total = 6328); and 9768 in skin (total = 10,153). In total, 111,788 variants corresponding to significant interactions were identified for whole blood (not accounting for linkage disequilibrium [LD] between variants); 82,250 and 132,980 for adipose and skin respectively. A small portion of xQTLs identified across all 3 tissues corresponded to known eQTLs from the GTEx consortium[1] (8.4% in blood; 10% in adipose; 4.8% in skin). An even smaller portion (approximately 0.33% for all 3 tissues) overlapped with putative deleterious variants (i.e. nonsynonymous or splice variants). Thus, initially, xQTLs appear to be distinct to both eQTLs and deleterious variants (i.e. do not overlap). Furthermore, unlike eQTLs, the clusters interactions associated with the xQTLs can be readily annotated with regulatory information and pathways (derived from the gene/protein content of the cluster pairs). As part of future aims, after finalizing the xQTL pipeline, I will conduct pi-hat and various other validation/stability analyses to demonstrate the robustness of xQTLs to both module definitions and natural/technical variability. After validation, I will begin analyzing all significant xQTLs to find "xQTL-xMAP cluster interaction pairs" that have a simple and readily interpretable biology (e.g. xQTLs influencing a transcription factor that regulates the expression of genes in one module that in turn affects the other). More generally, I will conduct tissue-specific enrichment analyses of genomic features (such as histone marks and chromatin state information) among xQTLs, which when contrasted to other genomic annotations (such as chromatin states and eQTLs) would shed insight about the likely function of xQTLs as an annotation group.

xQTLs increase the power to detect variants across T2D-related phenotypes, in a tissue-dependent manner. To directly evaluate the potential application of these information-rich xQTLs in disease mapping studies, I tested the xQTLs identified in each tissue for association with disease in 15 GWAS (listed in **Table 4.2**). Here, I called an annotation (e.g. xQTL or eQTL) as "tissue specific" if enrichment was present only in one tissue, and absent in all remaining tissues. By default, as xQTLs were defined only in adipose, skin, and blood, my analyses were limited to three tissues (a notable limitation to claims of tissue specificity and exclusivity). It is also worthwhile to note that tissue specificity in other contexts depends on demonstrating heterogeneity by tissue. However, it is not

immediately obvious how to assess heterogeneity across tissues with xQTLs as both cluster definitions and interactions are highly dependent on tissue. In other words, adipose xQTLs are discovered and associated with adipose-defined cluster interactions (the clusters and interactions are different between tissues). In contrast, the definition of a "gene" does not change from tissue to tissue and thus eQTLs can be directly compared across tissues. Thus, for this discussion, "tissue specificity" is limited to enrichment signal in exclusively one tissue. To partially account for LD between variants, I increased the xQTL lists to contain variants in high LD ($r^2 > 0.8$), using the European cohort in the 1000 Genomes Project[89].

Table 4.2 Summary of data sources used in this analysis. For datasets from NCBI's Gene Expression Omnibus, the parentheses denote the corresponding academic institution.

Source Database	Dataset Accession	Brief Description	Ref
EuroBATS consortium	NA	RNAseq and genotype data from three different tissues on 700 mono- and di-zygotic twin pairs	Part of MultiMuther Access
Gene Expression Omnibus (GEO)	GSE44035 (Lund)	Islet Microarray Expression Data	90
	GSE41762 (Lund)		91
	GSE38642 (Lund)		90
	GSE25724 (Pisa)		92
	GSE18732 (Edinburgh)		93
	GSE25462 (Joslin)	Muscle Microarray Expression Data	94
	GSE19420		
Roadmap	E087	Islet 15-State Chrom-HMM Model	18,95
	E107	Muscle 15-State Chrom-HMM Model	
	E108		
GWAS	IBDG Consortium	Ulcerative colitis, Crohn's disease	96,97
	NA	Rheumatoid arthritis	98
	NA	Type 1 diabetes	99
	CARDIoGRAM	Coronary Artery Disease, HDL, LDL, Total Cholestrol, Triglycerides	100,101
	DIAGRAM & MAGIC	T2D and related traits (see text and other tables for complete list)	23,102
	GIANT	BMI	103
	ICBP	Blood Pressure Measures	104

From this pilot comparison of xQTLs and eQTLs, I noted that:

- Adipose-specific xQTLs are greatly enriched for T2D GWA signal (**Figure 4.8**). This enrichment is absent from eQTLs from all three tissues (**Figure 4.8**), and from blood- and skin-specific xQTLs.
- Adipose xQTLs are greatly enriched for GWA signal for BMI, compared to eQTLs from all three tissues and background variants (i.e. non-xQTL snps; **Figure 4.9**). Here the tissue specificity is less obvious, with signal also present in skin-specific xQTLs.
- xQTLs and eQTLs have similar enrichment for cholesterol parameters (notably, HDL and total cholesterol levels in whole blood; **Figures 4.9-4.11**).
- eQTLs, if enriched in one tissue for a particular disease, are usually also enriched in the other tissues. For example, eQTLs from all three tissues were enriched for triglyceride levels GWA signal (**Figure 4.12**), type 1 diabetes [T1D] GWA signal (**Figure 4.13**), ulcerative colitis GWA signal (**Figure 4.14**), and Crohn's disease signal (**Figure 4.15**). In fact, the largest enrichment for Crohn's disease appears to be in skin-specific eQTLs. This is unsurprising: 43% of eQTLs are shared between at least 2 of the 3 tissues[1].
- There is enrichment in blood-specific xQTLs for Crohn's disease (**Figure 4.14**), that is absent from the ulcerative colitis plots (**Figure 4.14**) and from both adipose- and skin-specific xQTLs.
- For blood pressure related traits, there is enrichment from all eQTLs across all three tissues (**Figures 4.15-4.19**).

Tentatively, based on this limited pilot information, xQTLs appear to carry slightly more explicit "tissue-specific" content than classic eQTLs (likely derived from both the tissue dependent cluster definitions and interactions). Although I do want to expand this concept of xQTLs to other tissues, this analysis is only possible with datasets that contain monozygotic twins or multiple "close" time points per individual. This precludes application of this method to GTEx and other publicly available RNAseq datasets. However, the MultiMuther consortium is collecting and analyzing data from repeat time points per individual, allowing me to assess how the method behaves in real individuals with two time points (rather in monozygotic twins). Thus, future work will aim to refine the model and improve the clusters

that form the basis of this xQTL analysis. In a future pilot, I also seek to explore xQTLs in closely-related individuals (i.e. not identical genotypes) to assess feasibility of application to other datasets.

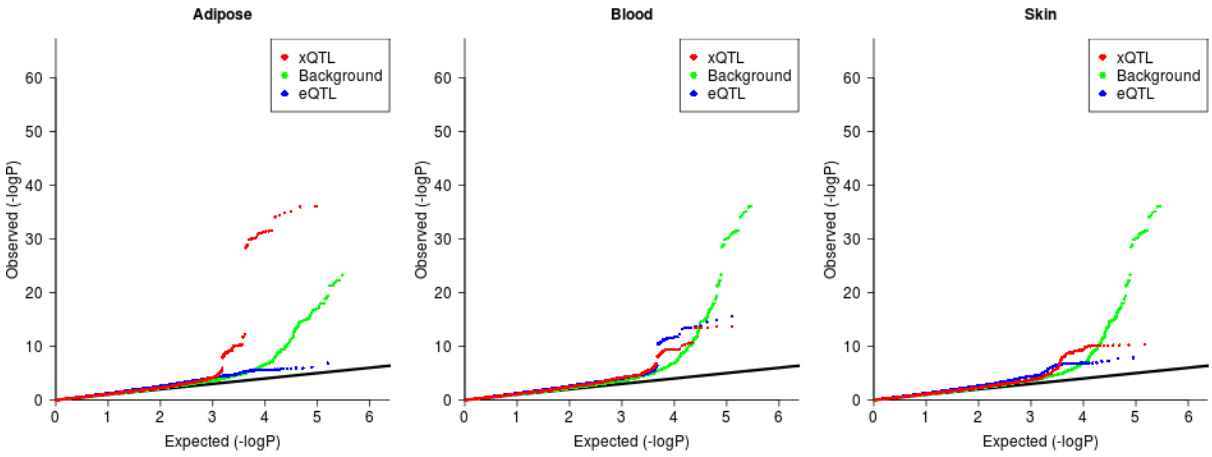


Figure 4.8 QQ-plot highlighting the enrichment of T2D GWA signal in adipose-specific xQTLs (leftmost panel) that is absent from eQTLs across all three tissues, and from blood- and skin-specific xQTLs. Enrichment is denoted by a leftward deviation from the null (black line) and background (i.e. non-xQTL variants; shown in green). As illustrated in the legend, xQTLs are coloured red, and eQTLs are coloured blue.

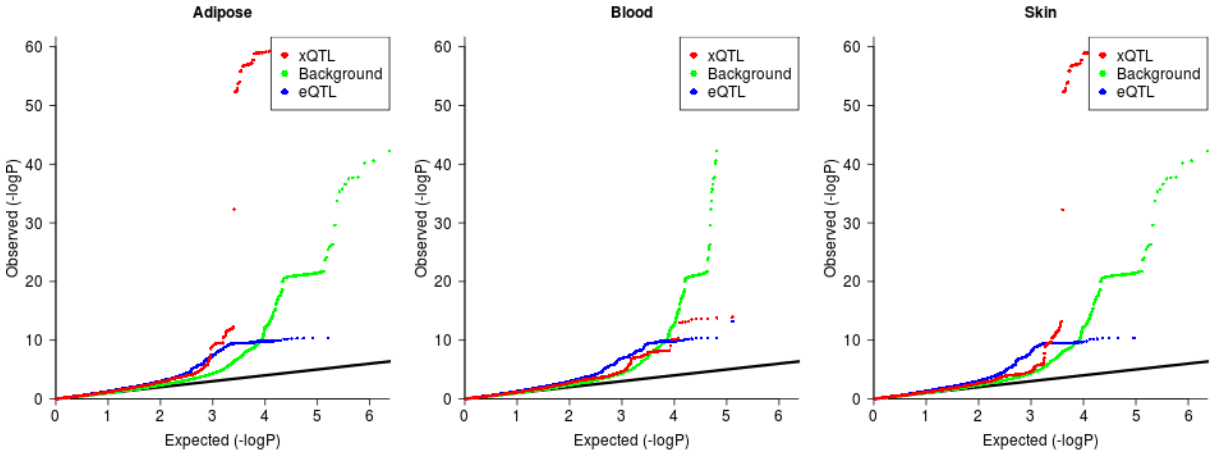


Figure 4.9 QQ-plot highlighting the enrichment of T2D GWA signal in both adipose-specific (leftmost panel) and skin-specific xQTLs (rightmost panel) that is absent from eQTLs across all three tissues and from blood-specific xQTLs. As described in text and in Figure 4.8, enrichment is denoted by a leftward deviation from the null (black line) and background (i.e. non-xQTL variants; shown in green). As illustrated in the legend, xQTLs are coloured red, and eQTLs are coloured blue.

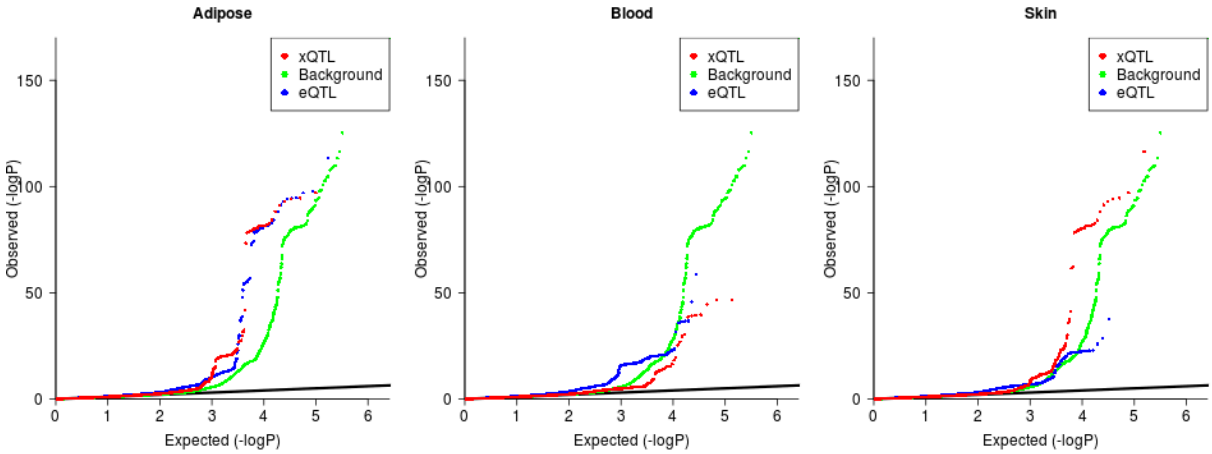


Figure 4.10 QQplot depicting the enrichment of HDL GWA signal among adipose-specific xQTLs and eQTLs.

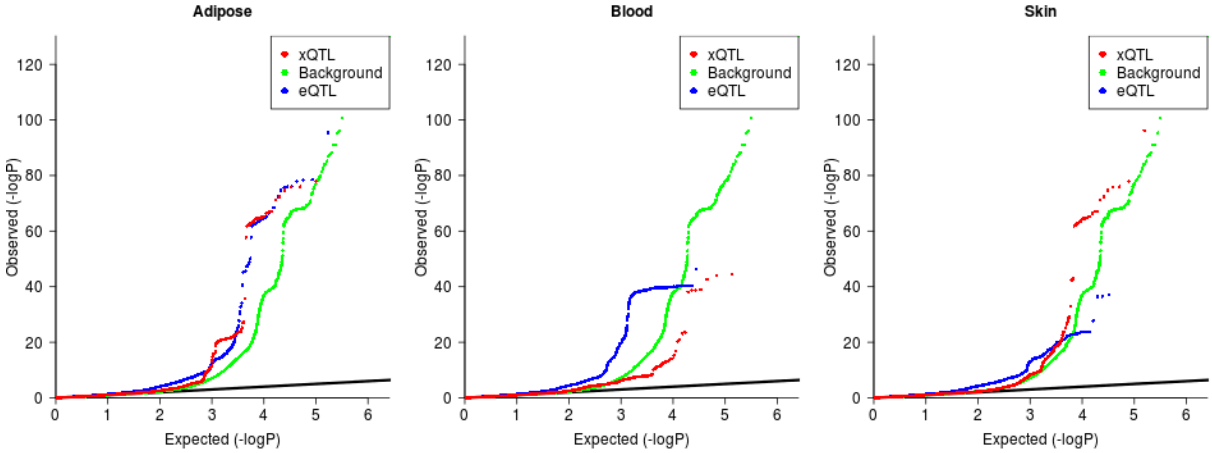


Figure 4.11 QQplot depicting the enrichment of total cholesterol GWA signal among adipose-specific xQTLs and eQTLs.

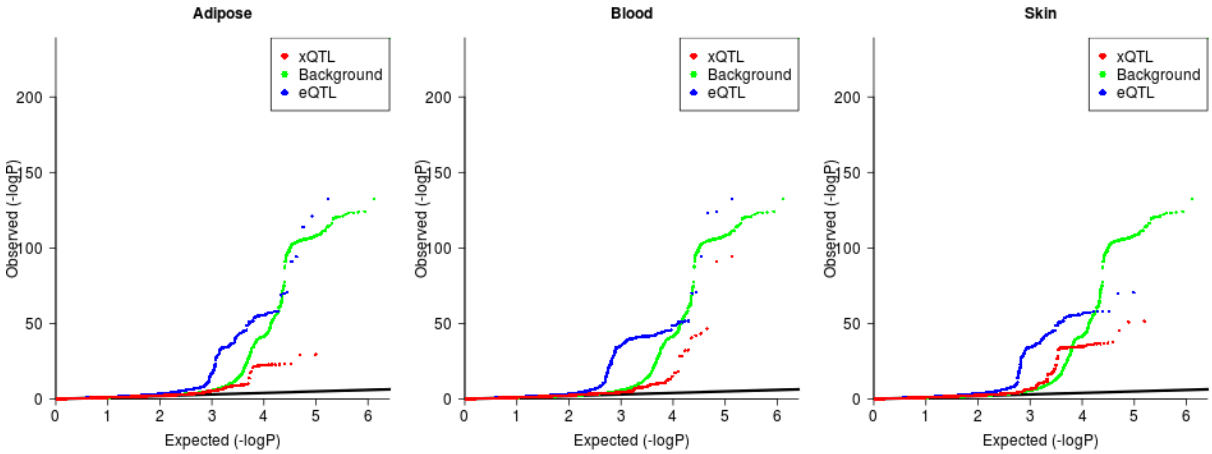


Figure 4.12 QQplot depicting the enrichment of triglycercide GWA signal in eQTLs across all three tissues (blue lines). This is one example of the lack of "tissue specificity" of eQTL annotations. With a simple interpretation of these plots, it is unclear what tissues may be involved in modulating triglycercide levels *in vivo*. There is no enrichment in xQTLs (red line).

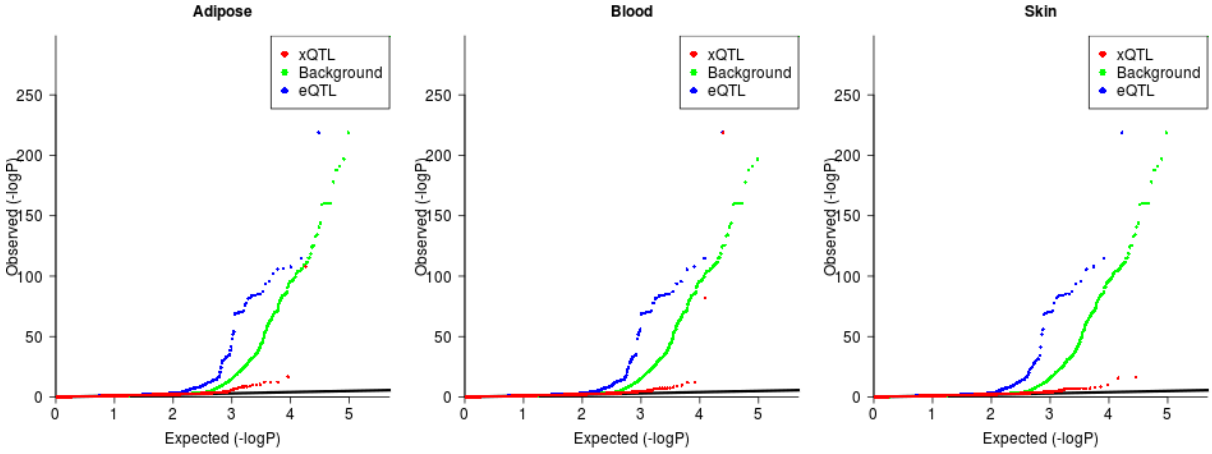


Figure 4.13 QQplot depicting the enrichment of T1D GWA signal in eQTLs across all three tissues (blue lines). This is another example of the lack of "tissue specificity" of eQTL annotations. There is no enrichment in xQTLs from any of the three tissues (red line).

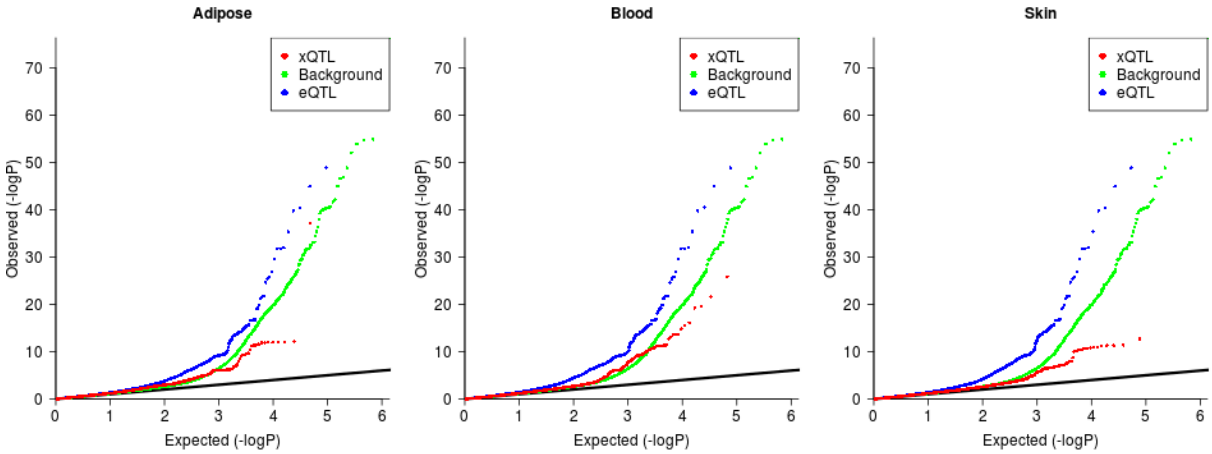


Figure 4.14 QQplot depicting the enrichment of ulcerative coilitis GWA signal in eQTLs across all three tissues (blue lines). This is one more example of the lack of "tissue specificity" of eQTL annotations. There is no enrichment in xQTLs from any of the three tissues (red line).

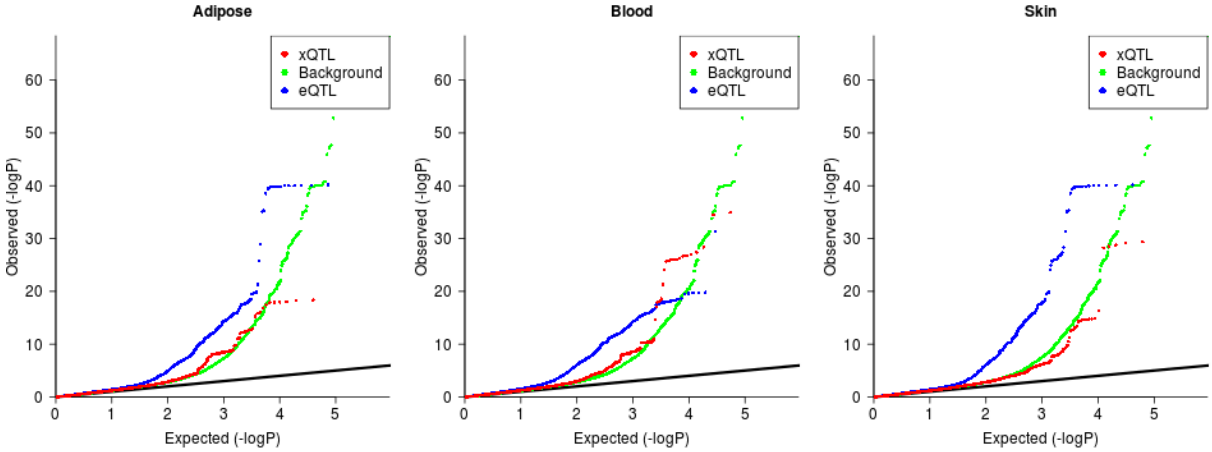


Figure 4.15 QQplot depicting the enrichment of Crohn's Disease GWA signal in eQTLs across all three tissues (blue lines). This is another example of the lack of "tissue specificity" of eQTL annotations. Notably, unlike with ulcerative coilitis, there is enrichment of signal in blood-specific xQTLs (middle panel; red line) that is absent in adipose-specific (left panel) and skin-specific (right panel) xQTLs.

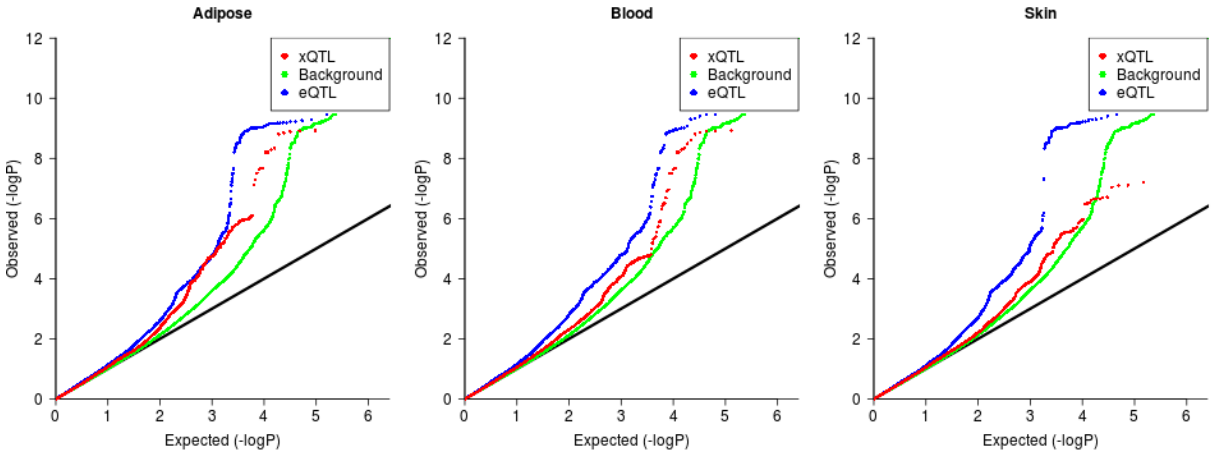


Figure 4.16 QQplot depicting the enrichment of systolic blood pressure GWA signal in eQTLs and to a lesser extent, xQTLs across all three tissues.

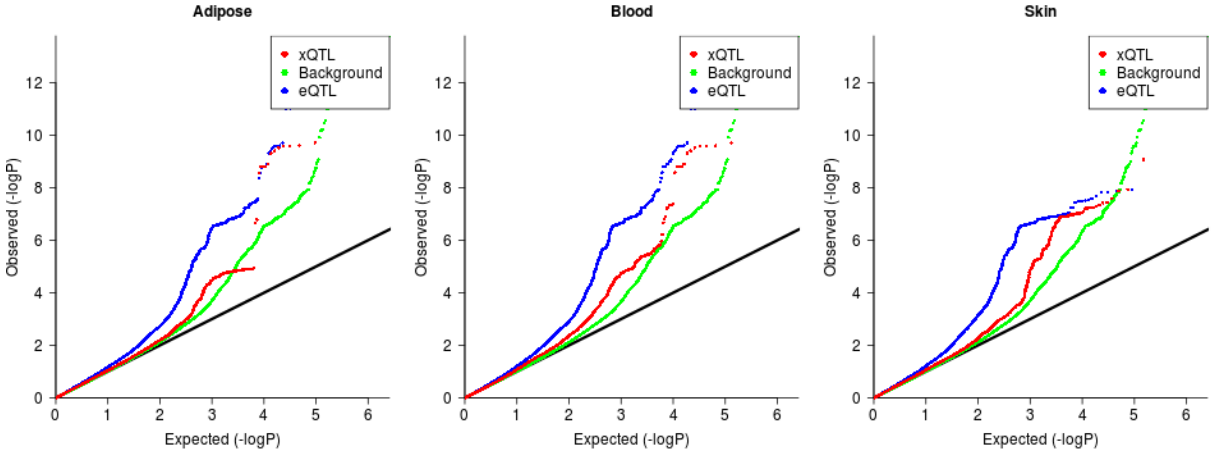


Figure 4.17 QQplot depicting the enrichment of diastolic blood pressure GWA signal in eQTLs across all three tissues.

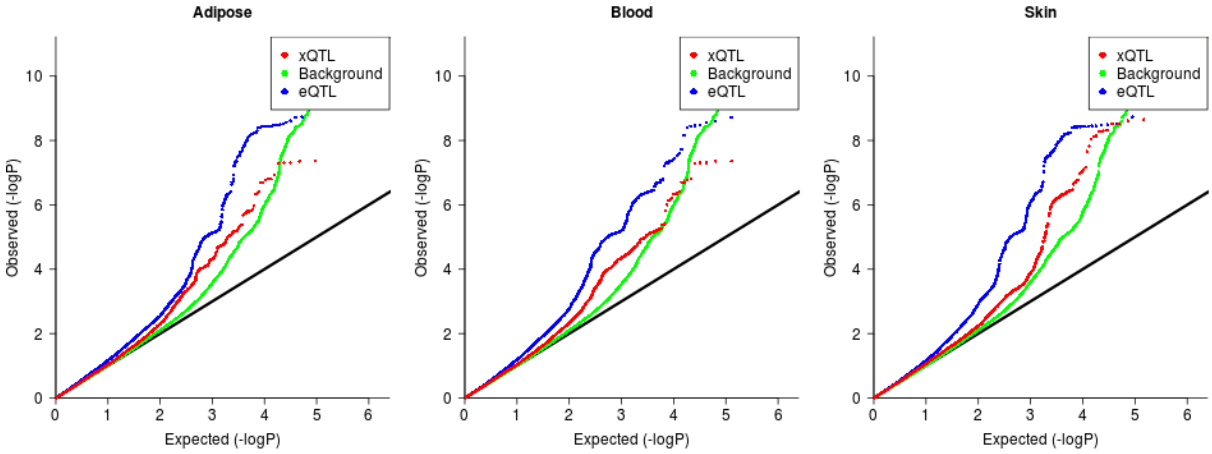


Figure 4.18 QQplot depicting the enrichment of pulse pressure GWA signal in eQTLs across all three tissues.

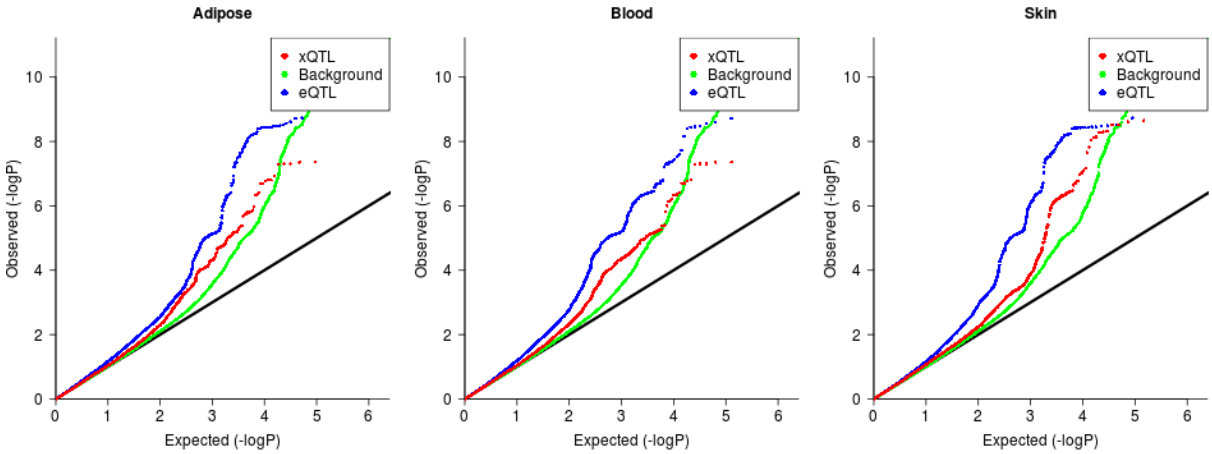


Figure 4.19 QQplot depicting the enrichment of mean arterial pressure GWA signal in eQTLs across all three tissues.

4.3 Discovery of "new" tissue-specific loci for T2D and related traits

As a final demonstration of the utility of the xMAP clustering algorithm, I integrated xMAP with a differential dependency strategy to discover "new" tissue specific loci relating to T2D and related traits. I implemented a python version of a non-parametric Bayesian approach developed in the Holmes group[105] to identify pairs of clusters that are differentially dependent between any two states (e.g. healthy versus diabetic tissue; see Fillipi *et al.* for details[105]) – this can be thought of as a non-parametric equivalent of classical differential coexpression. For readability, I defined the subset of all differentially dependent modules as xDDNs (acronym for coexpression differential dependency networks). Conceptually, xDDNs can be thought of as identifying T2D risk variants based on the secondary effect of T2D on transcriptomics. More work, however, is necessary to properly develop the biological intuition into why this approach works (e.g. which variants primarily drive the differential dependence that is at the heart to xDDNs? can other approaches capture a similar signal? what relationship do these xDDNs have with xQTLs?).

My analytic strategy using xDDNs to discovering "new" loci can be broadly divided into five stages. Given an expression dataset and GWAS summarized data for a T2D-related trait, I:

(a) define coexpression modules through modified affinity propagation (xMAP);

(b) identify module interactions that are differentially dependent between the control and disease states (can be thought of as differential coexpression);

(c) calculate enrichment of GWAS signal in variants delimited by tissue-specific xDDN annotations and sub-divisions thereof;

(d) jointly model multiple annotations, re-weighting the GWAS using select functional annotations; and finally,

(e) comparison to other tissue xDDNs and T2D-related traits (partially as controls and partially as validation).

I applied this analytic strategy to T2D (DIAGRAM[106]) and 6 T2D-related quantitative traits (MAGIC[22]): fasting glucose, glucose tolerance (2-hmy OGTT), glycated haemoglobin (HbA1C), fasting proinsulin, insulin secretion, and beta cell function (HOMA-B). Fasting insulin and insulin-resistance (HOMA-IR) were excluded, as they did not contain any GWAS hits at 5×10^{-8} threshold in the MAGIC dataset (there are other larger datasets that I hope to incorporate in future work). In this context, "new" refers to loci that did not reach genome-wide significance in any of the original MAGIC datasets used in this analysis. For expression data, I performed a systematic search for five diabetes-related tissues across Gene Expression Omnibus (GEO) and applied this strategy to two tissues for which data was available: pancreatic islets[90-92] and skeletal muscle[93,94] (data sources are highlighted in **Table 4.2**)[69]. No other large enough microarray data was found for other diabetes-relevant tissues (such as adipose). In the next iteration of this algorithm, I will replace the microarray dataset with the RNAseq datasets available from GTEx and the McCarthy group.

Variant annotations derived from tissue-specific xDDN genes are enriched for GWAS hits from T2D and related traits in a tissue-dependent manner. Using Ensembl's variant effect predictor (VEP)[107], I mapped variants to both genes and functional consequence coding annotations (e.g. missense), and subsequently, tested enrichment for variants associated with xDDNs and functional subdivisions within (n = 24 categories; complete list in **Table 4.3**). xDDNs defined using islet

[69] I couldn't use GTEx expression because that is where the eQTL data was derived from.

expression data were labelled islet-xDDNs; similarly xDDNs defined using muscle expression data were labelled muscle-xDDNs. Thus, "tissue specificity" for xDDNs is derived from differential dependence using the corresponding tissue-derived expression data. It is also important to note that cluster definitions themselves were also "tissue-specific" (i.e. derived exclusively from the tissue in which the xDDN is generated; see xMAP discussions above).

Table 4.3 Description of all 15 chromatin states from Roadmap used to annotate the 1 islet and 2 muscle epigenomes that are used in this transfer report. In text, I refer to the mnemonic descriptions.

State Number	Mnemonic	Description
1	TssA	Active TSS
2	TssAFlnk	Flanking Active TSS
3	TxFlnk	Transcr. at gene 5' and 3'
4	Tx	Strong transcription
5	TxWk	Weak transcription
6	EnhG	Genic enhancers
7	Enh	Enhancers
8	ZNF/Rpts	ZNF genes & repeats
9	Het	Heterochromatin
10	TssBiv	Bivalent/Poised TSS
11	BivFlnk	Flanking Bivalent TSS/Enh
12	EnhBiv	Bivalent Enhancer
13	ReprPC	Repressed PolyComb
14	ReprPCWk	Weak Repressed PolyComb
15	Quies	Quiescent/Low

Using a hierarchical model approach (fgwas)[108], I observed:

- Enrichment of HOMA-B (beta cell function) and insulin secretion GWA signal in intronic variants for islet-specific xDDN genes (enrichment ratio and confidence intervals are listed in **Table 4.4**). Insulin secretion GWA signal was also enriched in islet xDDN NMD transcript variants.
- Enrichment of fasting proinsulin GWA signal in islet-xDDN-specific missense, non-coding transcript exon, 5'-UTR, and upstream genetic variants (**Table 4.4**).
- None of the aforementioned annotations (e.g. intronic variants) were enriched for GWA signal independently; enrichment was observed only when limited to those influencing genes in the islet-xDDN.

- When looking at variants influencing genes in the muscle xDDN, I detected no enrichment for beta cell function or insulin secretion, across all categories (**Table 4.5**).
- I noted enrichment of fasting proinsulin among the muscle xDDN 5'UTR annotation. As noted above, the islet xDDN 5'UTRs were also enriched in fasting proinsulin signal. The presence of enrichment (of similar magnitude) in the same functional annotation suggested that this enrichment may be driven by the same set of variants. Indeed, after conditioning on the islet xDDN 5'UTR annotation, I could not detect signal enrichment in the muscle xDDN 5'UTR annotation.

For two traits (fasting glucose and glycated haemoglobin), I noted significant enrichment among variants separately annotated to both islet and muscle-xDDNs. The functional annotations responsible for this enrichment, however, differed between the two tissues:

- Fasting glucose associations were enriched in intronic and upstream genetic variants in the islet-xDDN (**Table 4.4**), but localized to missense variants in muscle-xDDN (**Table 4.5**).
- For glycated haemoglobin, islet-xDDN presented with general enrichment derived from non-coding transcript, non-coding exon, missense, and splice acceptor variants (**Table 4.4**). In contrast, signal for muscle-xDDN was enriched in the 3'-UTR, downstream genetic, missense, and nonsense mediated decay transcript variants (**Table 4.5**).
- For both these traits, this enrichment was a consequence of variants influencing genes in the islet and muscle xDDNs; the functional categories themselves were not enriched (both before and/or after conditioning).

Importantly, this enrichment demonstrates that both tissue and disease (i.e. differential dependence) context is critical to delineating variant effects, i.e. a combined or tissue-ambiguous analysis is not sufficient. The functional subdivisions of the islet- and muscle- xDDNs suggest that this genetic-driven causality is a consequence of variant perturbation at the transcriptional level (e.g. upstream genetic variants), post-transcriptionally (e.g. NMD transcript variants) and at the translational level (e.g. missense), and that the class of perturbation appears tissue and trait specific. No enrichment was

observed for T2D or glucose tolerance; in part, due to the low number of tissues analyzed (n = 2) and to the limited coverage of the xDDNs generated using microarray data (limited to n = 9596 and 8598 total genes for the pancreatic islets and skeletal muscle platforms respectively).

Table 4.4 (*next page*) Summary of independent enrichment results for functional subdivisions of variants influencing islet-xDDNs. The ln-fold enrichment is rounded to 2 significant digits, with the 95% confidence interval displayed in parentheses. Red-coloured entries denote significant enrichment. NA entries denote failure of the fgwas model to converge.

Table 4.5 (*page after next*) Summary of independent enrichment results for functional subdivisions of variants influencing muscle-xDDNs. The ln-fold enrichment is rounded to 2 significant digits, with the 95% confidence interval displayed in parentheses. Red-coloured entries denote significant enrichment of GWA signal exclusively in muscle xDDN. Brown entries denote significant enrichment that disappears after conditioning on the corresponding islet xDDN term (i.e. enrichment is not specific to islet xDDN).

Table 4.4

Islet-xDDN Variants	Beta Cell Function	Insulin Secretion	Fasting Proinsulin	Fasting Glucose	Glycated Haemoglobin	Glucose Tolerance	T2D
3' UTR	-30 (<-50, 5.4)	-26 (<-46, 6.4)	-26 (<-46, 4.6)	-29 (<-49, 4.7)	1.7 (<-20, 5.0)	-26 (<-46, >20)	-0.5 (<-20, 4.4)
5' UTR	-30 (<-50, 8.2)	5.0 (<-20, 9.3)	**6.1 (3.7, 8.0)**	5.3 (<-20, 7.2)	-27 (<-47, 5.6)	-25 (<-45, >20)	4.2 (<-20, 5.8)
Coding sequence	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8(<-20, >20)	4.0 (<-20, >20)	-2.4(<-20, >20)
Downstream gene	1.6 (<-20, 4.3)	3.6 (<-20, 6.6)	1.5 (<-20, 3.5)	1.8 (<-20, 3.6)	1.6 (<-20, 3.4)	-26 (<-46, >20)	1.4 (-2.7, 2.7)
Frameshift	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Incomplete terminal codon	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Inframe deletion	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Inframe insertion	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Intron	**3.3 (1.1, 6.2)**	**4.3 (0.99, >20)**	1.5 (-0.6, 3.2)	**2.0 (0.41, 3.5)**	1.3 (<-20, 2.9)	-29 (<-49, >20)	0.7 (-0.8, 1.8)
Mature miRNA variant	-18 (<-20, >20)	-17 (<-20, >20)	-21 (<-41, >20)	-16 (<-20, >20)	-15 (<-20, >20)	-23 (<-43, >20)	-18 (<-20, >20)
Missense variant	-23 (<-43, 7.1)	-28 (<-48, 7.8)	**4.8 (0.39, 6.7)**	-25.5918(<-45, 5.8)	**5.2 (2.7, 6.8)**	0.3(<-20, >20)	3.4(<-20, 5.5)
NMD transcript variant	2.5 (-0.6, 5.0)	**4.1 (0.9, 6.9)**	-1.4(<-20, 2.8)	1.7(-1.2, 3.4)	1.5(<-20, 3.6)	-27.6754(<-47, >20)	0.2(-3.3, 2.0)
Non-coding transcript exon	-30 (<-50, 5.3)	-28 (<-48, 6.2)	**4.4 (2.3, 6.0)**	-35 (<-55, 3.9)	**4.0 (1.9, 5.5)**	-25 (<-45, >20)	3.1 (-2.7, 4.5)
Non-coding transcript	1.0 (<-20, 3.5)	-26 (<-46, 3.9)	1.9 (-2.0, 3.6)	0.6 (<-20, 2.5)	**2.2 (0.09, 3.6)**	-28 (<-48, >20)	0.9 (-1.6, 2.2)
Protein altering variant	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Splice acceptor	-21 (<-41, >20)	-23 (<-43, >20)	-20 (<-40, >20)	7.3(<-20, >20)	**8.3 (4.3, 11.7)**	-19 (<-20, >20)	-27 (<-47, 8.9)
Splice donor	-23 (<-43, 10)	-18 (<-20, >20)	33 (<-20, >53)	-18 (<-20, 9.1)	-19 (<-20, >20)	8.7 (<-20, >20)	-22 (<-42, >20)
Splice region	-24 (<-44, 8.2)	-23 (<-43, 8.8)	-24 (<-44, 6.9)	-28 (<-48, 7.0)	-25 (<-45, 6.3)	-23 (<-43, >20)	-29 (<-49, 5.8)
Start lost	-18 (<-20, >20)	34 (<-20, >54)	-18 (<-20, >20)	-17 (<-20, >20)	-19 (<-20, >20)	-15 (<-20, >20)	-17 (<-20, >20)
Stop gained	-21 (<-41, >20)	-23 (<-43, >20)	-26 (<-46, 9.2)	-21 (<-41, >20)	-18 (<-20, >20)	-17 (<-20, >20)	9.2 (<-20, >20)
Stop lost	-16 (<-20, >20)	33 (<-20, >53)	-19 (<-20, >20)	-18 (<-20, >20)	-27 (<-47, >20)	-17 (<-20, >20)	-22 (<-42, >20)
Stop retained	-26 (<-46, >20)	-21 (<-41, >20)	-19 (<-20, >20)	-17 (<-20, >20)	-18 (<-20, >20)	34 (<-20, >54)	-24 (<-44, >20)
Synonymous	-25 (<-45, 7.6)	-27 (<-47, 8.0)	-27 (<-47, 5.8)	-24 (<-44, 5.7)	5.0(<-20, 6.7)	-27 (<-47, >20)	-24 (<-44, 4.7)
Upstream gene	2.6 (-0.8, 4.8)	NA	**2.9 (0.73, 4.5)**	**2.6 (0.42, 4.0)**	2.2 (-0.1, 3.7)	-30 (<-50, >20)	1.5 (<-20, 2.9)

Table 4.5

Islet-xDDN Variants	Beta Cell Function	Insulin Secretion	Fasting Proinsulin	Fasting Glucose	Glycated Haemoglobin	Glucose Tolerance	T2D
3' UTR	-25 (<-45, 5.5)	-29 (<-49, 6.1)	-27 (<-47, 4.2)	-30 (<-50, 4.6)	**4.3 (2.0, 5.7)**	-29 (<-49, >20)	-24 (<-44, 3.8)
5' UTR	-25 (<-45, 7.2)	-24 (<-44, 9.1)	4.8 (1.5, 6.6)	-24 (<-44, 5.5)	-26 (<-46, 5.3)	-25 (<-45, >20)	3.8 (<-20, 5.5)
Coding sequence	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Downstream gene	-28 (<-48, 3.0)	3.2 (<-20, 6.0)	1.7 (-2.0, 3.3)	1.3 (<-20, 3.2)	**2.6 (0.88, 3.9)**	-27 (<-46, >20)	0.7 (<-20, 2.5)
Frameshift	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Incomplete terminal codon	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Inframe deletion	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Inframe insertion	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Intron	1.0 (-2.3, 3.3)	1.6 (-1.6, 4.3)	1.1 (-0.8, 2.6)	0.1 (<-20, 1.9)	0.6 (<-20, 2.3)	-30 (<-50, >20)	0.4 (-1.3, 1.5)
Mature miRNA variant	-22 (<-42, 10.4)	-17 (<-20, >20)	-15 (<-20, >20)	-24 (<-44,10.4)	36 (<-20, >56)	-13 (<-20, >20)	-21 (<-40, 10)
Missense variant	-25 (<-45, 7.7)	5.5 (<-20, 8.9)	4.4 (<-20, 6.3)	**4.6 (1.3, 6.5)**	**5.9 (4.1, 7.4)**	6.4 (<-20, >20)	-24 (<-44, 4.5)
NMD transcript variant	-27 (<-47, 3.6)	2.0(<-20, 6.7)	-26 (<-46, 2.4)	-29 (<-49, 1.5)	**2.3 (0.03, 3.8)**	-30 (<-50, >20)	-27 (<-47, 1.1)
Non-coding transcript exon	-28 (<-48, 5.1)	-25 (<-45, 6.4)	4.0 (2.0, 5.5)	2.8 (<-20, 4.6)	**3.6 (1.2, 5.1)**	0.6 (<-20, >20)	2.4 (<-20, 4.1)
Non-coding transcript	-28 (<-48, 2.3)	-27 (<-47, 3.9)	1.4(<-20, 3.1)	0.2 (<-20, 2.5)	1.7 (-0.4, 3.1)	-31 (<-51, >20)	0.8 (-1.7, 2.1)
Protein altering variant	-1.8 (<-20, >20)	1.9 (<-20, >20)	-1.2 (<-20, >20)	-0.3 (<-20, >20)	-1.8 (<-20, >20)	4.0 (<-20, >20)	-2.4 (<-20, >20)
Splice acceptor	14 (<-20, >20)	-24 (<-44, >20)	-21 (<-41, >20)	-20 (<-40, 11)	-21 (<-41, >20)	-23 (<-43, >20)	-17 (<-20, >20)
Splice donor	-22 (<-42, 9.6)	-23 (<-43, >20)	-19 (<-20, >20)	NA	-20 (<-40, >20)	-21 (<-41, >20)	-23 (<-43, 9.9)
Splice region	-25 (<-45, 8.1)	-27 (<-47, 8.1)	-26 (<-46, 6.5)	-28 (<-48, 6.8)	-22 (<-42, 6.1)	-29 (<-49, >20)	-25 (<-45, 5.8)
Start lost	-19 (<-20, >20)	34 (<-20, >54)	-19 (<-20, >20)	35 (<-20, >55)	-23 (<-43, >20)	-23 (<-43, >20)	39 (<-20, >59)
Stop gained	-20 (<-40, >20)	-19 (<-20, >20)	-18 (<-20, >20)	-17 (<-20, >20)	-21 (<-41, 11)	-18 (<-20, >20)	NA
Stop lost	-21 <-41, >20)	33 (<-20, >53)	-19 (<-20, >20)	-15 (<-20, >20)	-22 (<-42, >20)	-17 (<-20, >20)	-21 (<-41, 9.5)
Stop retained	-20 (<-40, >20)	-20 (<-40, >20)	-19 (<-20, >20)	-16 (<-20, >20)	34 (<-20, >54)	33 (<-20, >53)	35 (<-20, >55)
Synonymous	-28 (<-48, 7.2)	-25 (<-45, 8.1)	-25 (<-45, 5.2)	-27 (<-47, 5.0)	4.2 (<-20, 6.2)	-24 (<-44, >20)	-23 (<-43, 4.6)
Upstream gene	-35 (<-55, 3.1)	0.8 (<-20, 5.7)	2.4 (0.4, 4.0)	-1.5 (<-20, 2.8)	0.6 (<-20, 3.0)	-25 (<-45, >20)	1.7 (-2.2, 3.0)

Chromatin state information and eQTLs reveal complementary but no "tissue specific" information to xDDN annotations. I was interested in comparing tissue-specific chromatin state-derived annotations, with those annotations from my xDDN-based strategy. Using Roadmap's 15-state core model[18] (complete list of states in **Table 4.3**), I searched for enrichment among the chromatin states from 1 islet and the union of 2 skeletal muscle epigenomes (sources listed in **Table 4.2**). Future work will include adding chromatin-state data available internally from the McCarthy group. With my limited pilot data, I observed enrichment in a subset of states, but the enrichment was not informative about the underlying disease tissue (**Tables 4.6 and 4.7**). For example, only the Weak Repressed PolyComb (14_ReprPCWk) state from skeletal muscle was enriched for loci influencing insulin secretion. Similarly, both islet- and muscle-specific chromatin states were enriched for loci influencing beta cell function and fasting proinsulin (traits that are decidedly islet-specific; **Tables 4.6 and 4.7**). As with the xDDN analysis, I observed no enrichment for glucose tolerance among the islet- or muscle-chromatin state or eQTL annotations. Others have noted enrichment of glucose-related traits in islet chromatin states[109,110]. This could be a consequence of methodological differences: Parker *et al.*[110] implemented a 10-state Chrom-HMM model, Pasquali *et al.*[109] applied k-medians clustering to subdivide all accessible chromatin into 5-states, and I used Roadmap's 15-state Chrom-HMM model. It is not immediately obvious how well the states correspond to each other. Unlike the xDDN analysis, T2D GWA signal was present in a subset of chromatin states and eQTL data from both tissues, suggesting that these annotations provide some complementary information to annotations derived from differential dependence networks. Notably, conditioning on chromatin state annotations that are not specific to any tissue eliminated largely all tissue-specific enrichment. While it is tempting to draw conclusions about underlying (pathophysiological) differences between the traits (e.g. noting the absence of xDDN annotations in T2D), it is important to note that xDDN annotations were derived from limited microarray data and from only 2 tissues.

Table 4.6 Summary of independent enrichment results for islet-specific chromatin states and eQTLs. The ln-fold enrichment is rounded to 2 significant digits, with the 95% confidence interval displayed in parentheses. Red-coloured entries denote significant enrichment. NA denotes failure of fgwas model to converge.

Islet-xDDN Variants	Beta Cell Function	Insulin Secretion	Fasting Proinsulin	Fasting Glucose	Glycated Haemoglobin	Glucose Tolerance	T2D
1_TssA	**4.1 (1.7, 6.0)**	NA	**3.4 (1.3, 5.0)**	**3.9 (2.4, 5.1)**	2.8 (<-20, 4.3)	31 (<-20, >51)	**3.4 (2.2, 4.3)**
2_TssAFlnk	-24 (<-44, 5.9)	NA	-27 (<-47, 4.5)	-25 (<-45, 4.7)	3.9(<-20, 5.5)	-24 (<-44, >20)	-26 (<-46, 4.2)
3_TxFlnk	-23 (<-43, 10)	-18 (<-20, >20)	NA	-21 (<-41, 8.5)	-22 (<-42, 7.6)	-20 (<-40, >20)	-27 (<-47, 6.4)
4_Tx	5.2 (-0.2, 8.1)	-29 (<-49, 6.3)	3.4 (-0.4, 5.1)	**4.2 (2.2, 5.9)**	-25 (<-45, 3.8)	-30 (<-50, >20)	**3.1 (0.50, 4.5)**
5_TxWk	-25 (<-45, 2.6)	1.1 (<-20, 3.9)	1.8 (-0.4, 3.6)	0.2(<-20, 2.1)	**1.7 (0.19, 3.0)**	-28 (<-48, >20)	**1.5 (0.30, 2.4)**
6_EnhG	**34.8 (4.3, >54)**	-26 (<-46, >20)	**6.7 (3.6, 8.7)**	**6.8 (3.6, 9.4)**	5.7 (<-20, 7.9)	6.8 (<-20, >20)	5.7 (<-20, 8.0)
7_Enh	3.3(<-20, 5.9)	2.7(<-20, 5.4)	1.2(<-20, 3.4)	**3.9 (2.4, 5.2)**	2.4(<-20, 4.1)	-26 (<-46, >20)	**2.6 (0.50, 3.7)**
8_ZNF_Rpts	**6.1 (2.3, 8.2)**	-22 (<-42, 8.2)	-28 (<-48, 4.9)	-29 (<-49, 5.5)	-28 (<-48, 4.9)	-25 (<-45, >20)	-28 (<-48, 3.5)
9_Het	-27 (<-47, 2.8)	-32 (<-52, 3.2)	-26 (<-46, 2.4)	-27 (<-47, 1.9)	-30 (<-50, 1.7)	-27 (<-47, >20)	-26 (<-46, 1.4)
10_TssBiv	-31 (<-51, 7.3)	-21 (<-41, 10.5)	-27 (<-47, 7.1)	-24 (<-44, 6.4)	-25 (<-45, 5.6)	-23 (<-43, >20)	-27 (<-47, 4.8)
11_BivFlnk	-23 (<-43, 10)	-20 (<-40, >20)	-23 (<-43, >20)	-21 (<-41, 10)	-20 (<-40, >20)	-21 (<-41, >20)	-20 (<-40, 9.4)
12_EnhBiv	-26 (<-46, 11)	NA	-25 (<-45, 10)	-27 (<-47, 10)	-25 (<-45, 8.3)	33 (<-20, >53)	-22 (<-42, 8.6)
13_ReprPC	-31 (<-51,5.7)	NA	-26 (<-46, 4.4)	-28 (<-48, 4.0)	-29 (<-49, 2.9)	-26 (<-46, >20)	NA
14_ReprPCWk	-30 (<-50, 1.8)	NA	-28 (<-48, 0.5)	-26 (<-46, 1.1)	-1.0 (-4.0, 0.7)	-28 (<-48, >20)	-0.4 (-2.8, 0.9)
15_Quies	-1.2 (-4.1, 1.8)	NA	**-1.9 (-4.6, -0.12)**	-2.0 (-3.9, -0.3)	-0.4 (-1.9, 1.1)	26 (<-20, >46)	**-1.3 (-2.5, -0.34)**
eQTLs	1.(-0.3, 4.8)	-0.9 (<-20, 2.2)	0.1 (-1.5, 1.7)	1.0 (-0.3, 2.7)	1.3 (-0.1, 3.4)	-31 (<-51, 2.9)	**1.2 (0.22, 2.3)**

Table 4.7 Summary of independent enrichment results for muscle-specific chromatin states and eQTLs. The ln-fold enrichment is rounded to 2 significant digits, with the 95% confidence interval displayed in parentheses. Red-coloured entries denotes significant enrichment.

Islet-xDDN Variants	Beta Cell Function	Insulin Secretion	Fasting Proinsulin	Fasting Glucose	Glycated Haemoglobin	Glucose Tolerance	T2D
1_TssA	-27 (<-47, 4.5)	-30 (<-50, 5.0)	2.7 (<-20, 4.6)	-27 (<-47, 3.6)	-32 (<-52, 3.7)	-28 (<-48, >20)	2.1 (<-20, 3.7)
2_TssAFlnk	-27 (<-47, 4.1)	-28 (<-48, 4.5)	**2.9 (0.43, 4.6)**	-28 (<-48, 3.1)	**3.3 (1.6, 5.7)**	-30 (<-50, 20)	1.5 (<-20, 3.1)
3_TxFlnk	-27 (<-47, 4.1)	-28 (<-48, 4.5)	2.9(0.4, 4.6)	-28 (<-48, 3.1)	3.3 (1.6, 4.7)	-30 (<-50, >20)	1.5 (<-20, 3.1)
4_Tx	-26 (<-46, 6.3)	3.4 (<-20, 7.7)	-28 (<-48, 4.7)	-24 (<-44, 4.3)	-27 (<-47, 4.5)	-25 (<-45, >20)	3.5 (<-20, 5.1)
5_TxWk	1.5 (<-20, 3.9)	-28 (<-48, 2.9)	1.0 (<-20, 2.7)	1.7 (-2.1, 3.2)	1.2 (-1.6, 2.8)	-27 (<-47, >20)	0.3 (<-20, 1.9)
6_EnhG	-26 (<-46, 1.8)	0.1 (<-20, 3.2)	-0.4 (-3.4, 1.4)	-0.2 (<-20, 1.5)	0.3 (-2.4, 1.9)	27 (-0.6, >47)	0.6 (-0.8, 1.6)
7_Enh	-25 (<-45, 4.3)	-29 (<-49, 4.5)	2.5 (<-20, 4.3)	-28 (<-48, 3.4)	2.0 (<-20, 4.0)	-25 (<-45, >20)	2.0 (-1.5, 3.6)
8_ZNF_Rpts	-28 (<-48, 2.3)	-26 (<-46, 3.6)	-27 (<-47, 2.0)	-28 (<-48, 1.1)	-27 (<-47, 1.3)	33 (<-20, >53)	1.4 (-0.0, 2.4)
9_Het	-24 (<-44, 6.5)	-27 (<-47, 8.0)	-26 (<-46, 5.5)	-32 (<-52, 5.2)	-24 (<-44, 5.0)	-26 (<-46, >20)	-31 (<-51, 3.5)
10_TssBiv	-28 (<-48, 4.0)	-27 (<-47, 4.4)	-26 (<-46, 3.1)	-27 (<-47, 2.9)	-27 (<-47, 2.5)	-28 (<-48, >20)	-32 (<-52, 1.6)
11_BivFlnk	-42 (<-62, 6.3)	-22 (<-42, 9.8)	-26 (<-46, 4.6)	-25 (<-45, 5.0)	-26 (<-46, 3.9)	-24 (<-44, >20)	**3.7 (0.73, 5.1)**
12_EnhBiv	-25 (<-45, 5.7)	4.0 (<-20, 8.7)	-28 (<-48, 3.5)	-28 (<-48, 3.7)	0.2 (<-20, 3.5)	-25 (<-45, >20)	0.6 (<-20, 3.8)
13_ReprPC	3.3 (<-20, 5.8)	-28 (<-48, 5.4)	-28 (<-48, 1.5)	1.7 (<-20, 3.8)	1.0 (-1.4, 2.7)	-28 (<-48, >20)	-26 (<-46, 1.9)
14_ReprPCWk	**2.8 (0.49, 5.3)**	**3.5 (0.96, 7.0)**	**-30 (<-51, -.19)**	1.6 (-0.0, 3.3)	-1.0(-4.2, 0.7)	-28 (<-48, >20)	0.5 (-0.9, 1.7)
15_Quies	-1.6 (-3.9, 1.0)	-3.1(<-20, 0.5)	-0.1(-2.0, 1.6)	**-1.6 (-3.5, -0.03)**	-1.1 (-2.9, 0.4)	-28 (<-48, 1.7)	**-1.4 (-2.4, - 0.36)**
eQTLs	-28 (<-48, 1.9)	1.6 (<-20, 4.2)	-26 (<-46, 1.6)	1.1 (-1.0, 2.6)	**1.7 (0.21, 3.0)**	-28 (<-48, >20)	**1.4 (0.06, 2.4)**

Construction of a joint annotation model reveals novel risk loci for T2D-related traits. Having observed that many (possibly correlated) functional categories of variants are enriched for GWA signal, I elected to combine annotations to arrive at a biological interpretation of the nature of both the xDDN and chromatin-state variants that influence these traits (i.e. selecting the "best" representatives of sets among all tissue-specific annotations). For beta cell function, fasting proinsulin, and insulin secretion, I constructed cross-validated models (n = 3; 1 for each trait) with islet-specific annotations. Since xDDN annotations were already tissue-specific (in this 2-tissue comparison; i.e. no muscle xDDN annotations were enriched), this largely meant excluding enrichment from muscle-specific chromatin state data. Briefly, I re-iteratively added annotations that improved the likelihood of the model until there was no further improvement. I subsequently searched for the penalty with the best cross-validation likelihood and dropped annotations sequentially to find the final model with the maximum likelihood[108]. Similarly, for glycated hemoglobin (HbA1c) and fasting glucose, I built separate tissue-specific models (n = 2 for each trait; total = 4). No model was built for glucose tolerance (no annotations were enriched for this trait) or T2D. The best-fitting models for each trait/tissue pair are shown in **Tables 4.8-4.14** and discussed in turn below. Having constructed the joint models, I was subsequently interested in whether these models can identify new risk loci. Reweighting of association statistics in a tissue-dependent manner identified 1 "novel" risk locus influencing both beta cell function and insulin secretion (**Figure 20**), 1 locus for fasting glucose (specific to the islet model; **Figure 4.21**), and 1 locus for glycated haemoglobin (identified with both islet- and muscle- joint models; **Figure 4.22**). As aforementioned, in this context, "novel" refers to loci that did not reach genome-wide significance in the original MAGIC dataset used in this analysis. I selected loci with a posterior probability of association (PPA) greater than 0.9 (as recommended by the author of fgwas[108]), a threshold shown to reduce the number of false positives as the genome-wide significant p value threshold of 5×10^{-8}. Re-calibrating this threshold to T2D-related traits may yield greater number of risk loci. The creation of tissue-specific models readily suggests the effector tissue for some, but not all, of the newly-identified risk loci. 2 of 3 loci have been confirmed in larger cohorts in the same or closely-related traits, and all 3 loci have been partially verified experimentally (i.e. implicated in some way in diabetes). As a quick note, the approach

identifies loci ("segments") of genes with high posterior probability of association. However, the approach in some cases (as evident with one of the loci) fails to identify the index or causal snp.

xMAP and xDDN conclusions. I've developed a new analytic strategy, extending tissue-specific differential dependence networks, to identify the causal tissues and mechanisms underlying T2D and related traits. I discovered 3 new risk loci influencing 4 traits in 2 tissues: beta cell function (islet), insulin secretion (islet), fasting glucose (islet), and glycated haemoglobin (muscle). 2 of these loci have confirmed associations with the same or related traits and all 3 phenotype-loci associations have been verified experimentally. This new strategy can be applied to other phenotypes, leveraging the convergence of expression and genetic data to make sense of the underlying variation.

Table 4.8 Islet-specific model for beta cell function; annotations are ranked in descending order of model improvement (i.e. top annotation improved the model likelihood by the greatest amount). This represents the model with the greatest cross-validated likelihood; all annotations were significant when analyzed independently.

Rank	Annotation	ln-fold Model Enrichment
1	Islet-xDDN: intron_variant	3.2 (1.0, 6.1)
2	6_EnhG (islet-specific)	39.4 (5.0, >59)

Table 4.9 Islet-specific model for insulin secretion; annotations are ranked in descending order of model improvement (i.e. top annotation improved the model likelihood by the greatest amount). This represents the model with the greatest cross-validated likelihood; all annotations were significant when analyzed independently.

Rank	Annotation	ln-fold Model Enrichment
1	Islet xDDN: intron_variant	4.3 (0.99, >20)

Table 4.10 Islet-specific model for fasting proinsulin; annotations are ranked in descending order of model improvement (i.e. top annotation improved the model likelihood by the greatest amount). This represents the model with the greatest cross-validated likelihood; all annotations were significant when analyzed independently.

Rank	Annotation	ln-fold Model Enrichment
1	6_EnhG	7.6 (4.5, 9.8)
2	Islet xDDN: 5_prime_UTR_variant	6.1 (3.0, 8.3)
3	Islet xDDN: non_coding_transcript_exon_variant	3.6 (1.6, 5.2)
4	Islet xDDN: missense_variant	3.4 (0.34, 5.2)
5	Islet xDDN: upstream_gene_variant	-5.3 (<-20, -2.3)
6	1_TssA	3.4 (1.6, 4.9)

Table 4.11 Islet-specific model for fasting glucose; annotations are ranked in descending order of model improvement (i.e. top annotation improved the model likelihood by the greatest amount). This represents the model with the greatest cross-validated likelihood; all annotations were significant when analyzed independently.

Rank	Annotation	ln-fold Model Enrichment
1	1_TssA	29.6 (28.0, 30.8)
2	6_EnhG	30.2 (33.5, 36.0)
3	4_Tx	29.2 (26.9, 30.7)
4	Islet xDDN: intron_variant	1.5 (0.18, 2.7)
5	7_Enh	28.6 (26.6, 29.9)
6	Islet xDDN: upstream_gene_variant	-94.9 (<-114.9, 0.73)
6	15_Quies	23.89 (25.3, NA)

Table 4.12 Muscle-specific model for fasting glucose; annotations are ranked in descending order of model improvement (i.e. top annotation improved the model likelihood by the greatest amount). This represents the model with the greatest cross-validated likelihood; all annotations were significant when analyzed independently.

Rank	Annotation	ln-fold Model Enrichment
1	Muscle xDDN: missense_variant	4.1 (0.60, 5.9)
2	15_Quies	-1.5 (-3.3, 0.13)

Table 4.13 Islet-specific model for glycated haemoglobin; annotations are ranked in descending order of model improvement (i.e. top annotation improved the model likelihood by the greatest amount). This represents the model with the greatest cross-validated likelihood; all annotations were significant when analyzed independently.

Rank	Annotation	ln-fold Model Enrichment
1	Islet xDDN: missense_variant	4.3 (1.7, 5.9)
2	5_TxWk	1.5 (0.13, 2.8)
3	Islet xDDN: splice_acceptor_variant	7.7 (3.9, 11.3)
4	Islet xDDN: non_coding_transcript_variant	1.0 (-0.88, 2.4)

Table 4.14 Muscle-specific model for glycated haemoglobin; annotations are ranked in descending order of model improvement (i.e. top annotation improved the model likelihood by the greatest amount). This represents the model with the greatest cross-validated likelihood; all annotations were significant when analyzed independently.

Rank	Annotation	ln-fold Model Enrichment
1	Muscle xDDN: missense_variant	3.6 (1.5, 5.4)
2	2_TssAFlnk	2.8 (0.88, 4.2)
3	Muscle xDDN: 3_prime_UTR_variant	2.6 (-0.16, 4.2)
4	Muscle xDDN: NMD_transcript_variant	1.2 (-0.64, 2.6)
5	Muscle xDDN: downstream_gene_variant	0.86 (-0.91, 2.3)
6	Muscle xDDN: non_coding_transcript_exon_variant	0.94 (-1.9, 2.8)
7	Muscle eQTL	0.38 (-1.3, 1.8)

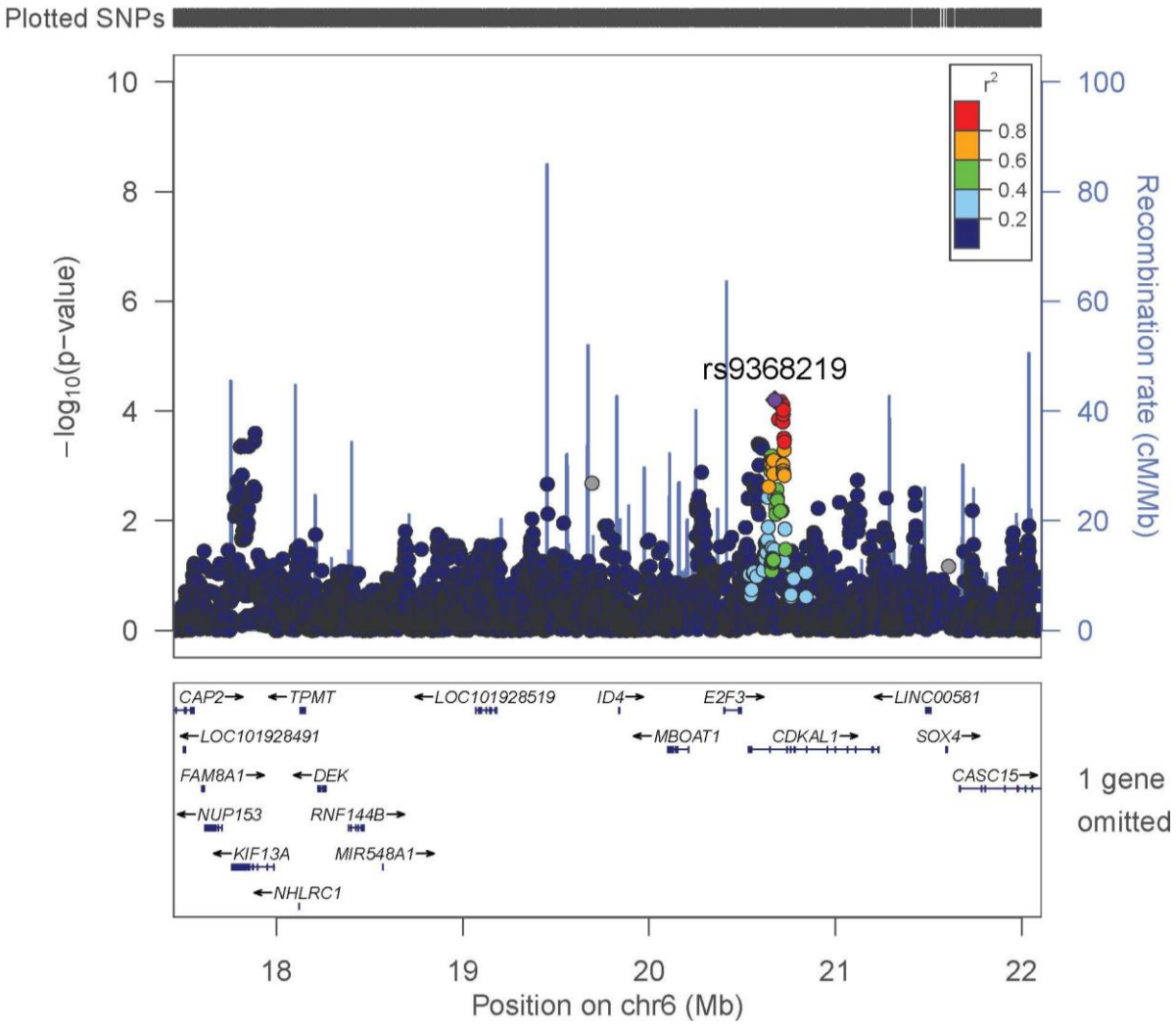


Figure 4.20 Locus zoom plot of new risk loci influencing both beta cell function and insulin secretion, discovered with the islet-specific models for both traits. The index snp for the new risk locus (chr6:17447095-22107290) associated with beta cell function (PPA of segment = 0.99941) and insulin secretion (PPA of segment = 0.99941) is rs9368219 (an intronic variant of CDKAL1). (The model did not definitively identify any index snps and thus, I elected to select the lowest pvalue variant in the segment.) CDKAL1 intronic variants have been associated with type 2 diabetes[111] (OR = 1.2; $p = 2 \times 10^{-26}$), 1-hmy glucose after an oral glucose tolerance test[112] (beta = 0.246 $mmol^{-1}$ increase; $p = 3 \times 10^{-19}$), glycated hemoglobin[113] (beta = 0.046% decrease; $p = 1 \times 10^{-11}$), fasting plasma glucose[114] (beta = 0.057 unit increase; $p = 9 \times 10^{-10}$), and acute insulin response to glucose[115] (beta = 2.05 unit decrease; $p = 1 \times 10^{-6}$). Furthermore, CKDAL1 variants have been shown experimentally to decrease pancreatic beta-cell function via decreased beta-cell glucose sensitivity that relates insulin secretion to plasma glucose concentration[116], and through genetic studies, associated with both lean and obese type 2 diabetics[117]; consistent with a role for the locus in beta cell function. Prior to constructing this model with xDDN-islet annotations, no locus on chromosome 6 was *genetically* associated with beta cell function (HOMA-B) according to NHGRI-EBI's GWAS catalogue.

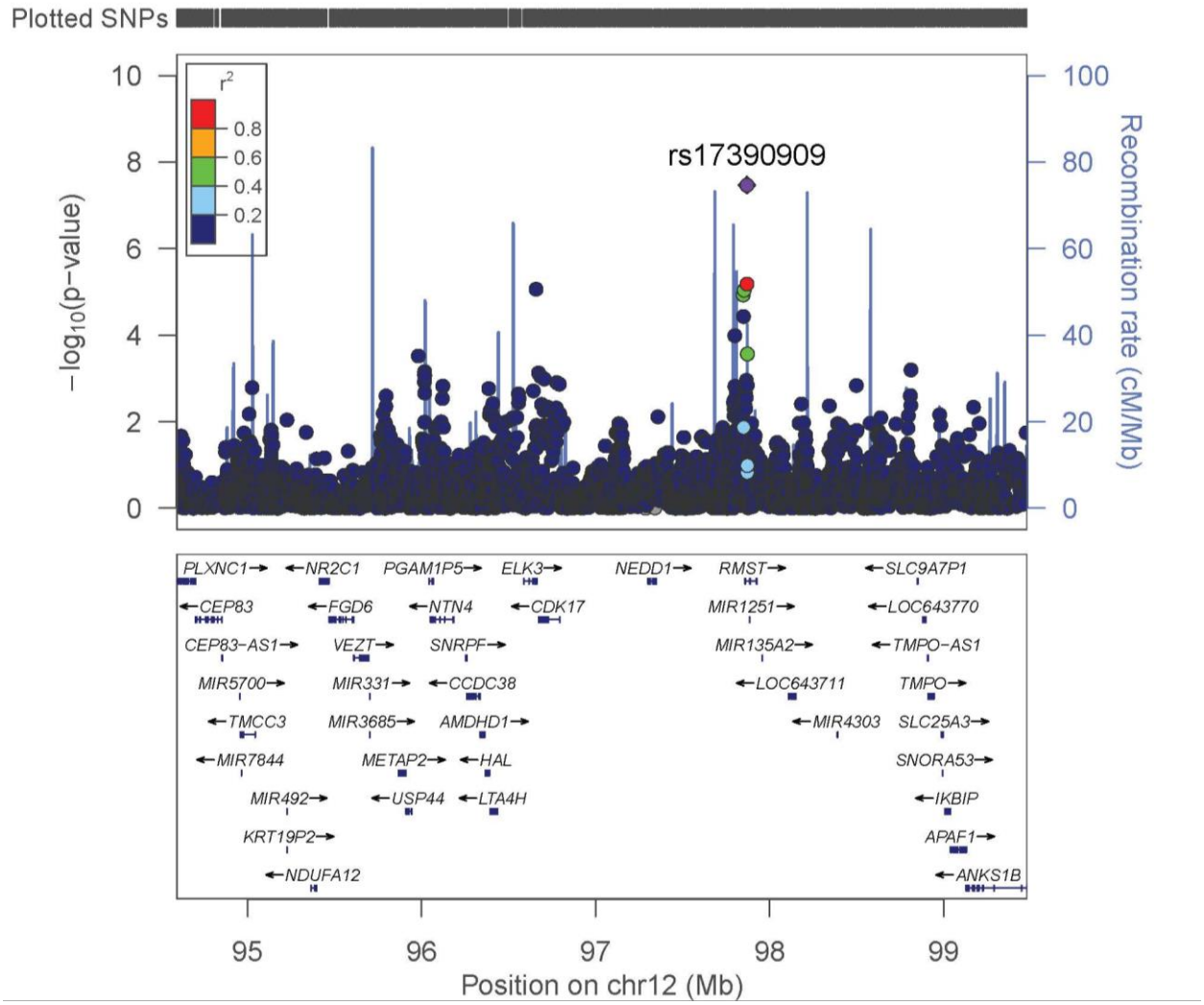


Figure 4.21 Locus zoom plot of new risk loci associated with fasting glucose, discovered only with the islet-joint model. The index snp of the new risk loci (chr6:17447095-22107290) associated with fasting glucose (PPA of segment = 0.946974) is rs17390909, an intronic variant influencing a lincRNA, RMST. This variant, represented islet active transcription start site annotations, did not reach genome wide significance in the MAGIC dataset (in up to 46,186 individuals). However, a sex-combined meta-analysis by the ENGAGE consortium[118] confirms the association of this locus with fasting glucose (index snp = rs17331697, a variant in high LD with rs17390909; $p = 1.3 \times 10^{-11}$). As this new risk locus emerges only from the islet-specific model, pancreatic islets may be the primary tissue of action.

Figure 22.

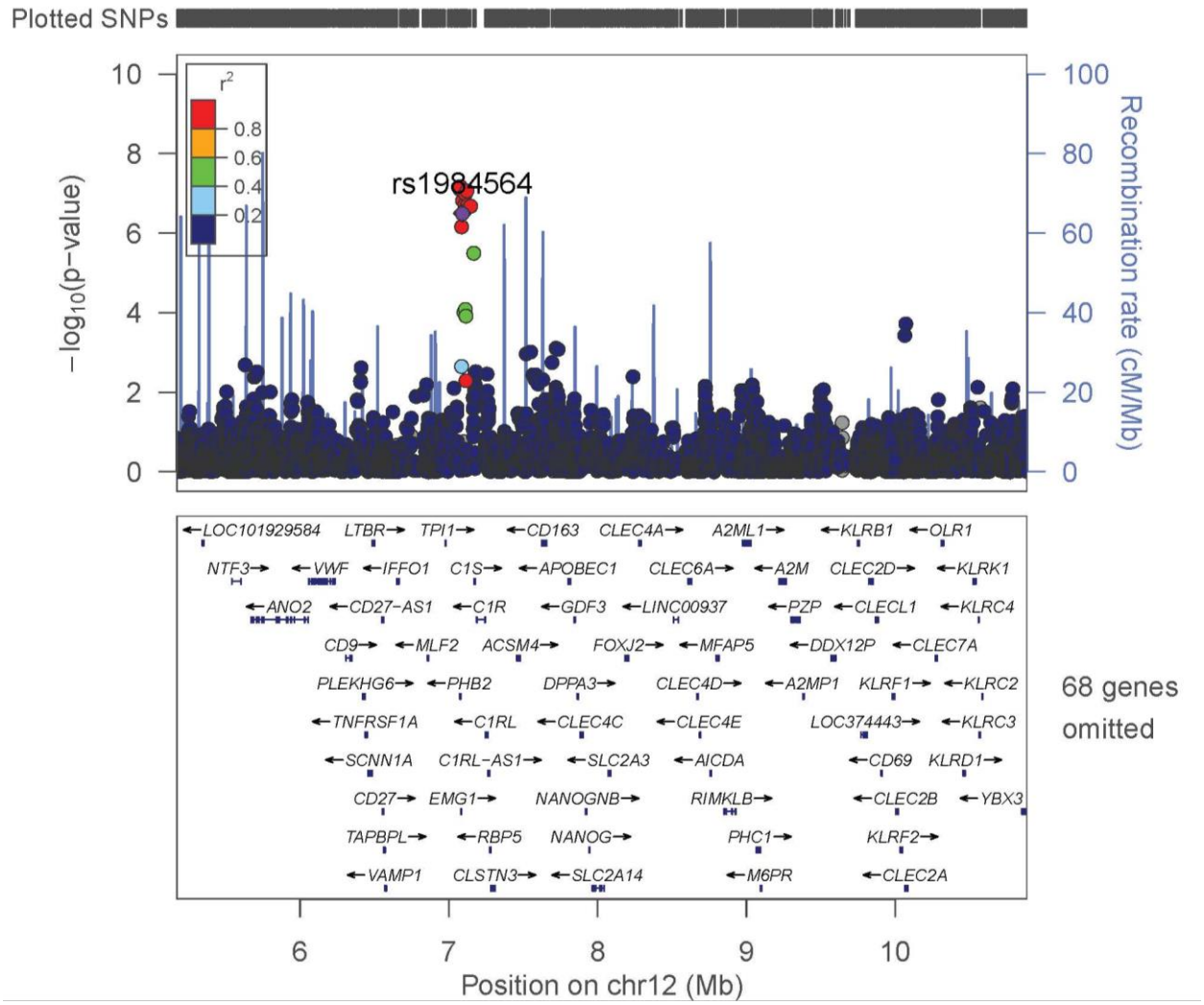


Figure 4.22 Locus zoom plot of new risk loci associated with glycated haemoglobin, discovered with both the islet- and muscle-joint models. In muscle-specific model, the index snp of the new risk of the new risk locus (chr12:5170313-10888611) associated with fasting glucose (PPA of segment = 0.991529) is rs1984564, a missense variant for LPCAT3 (conditional variant PPA = 0.766255). This does not correspond to the lowest-pvalue variant in the original dataset (rs2110073, an intronic variant for HPG2, conditional variant PPA = 0.00390974). Interestingly, only the LPCAT3 gene is in the muscle-xDDN (HPG2 is absent); for comparison, both genes are in the islet-xDDN. There is experimental evidence demonstrating that the deletion of LPCAT3 promotes enhanced insulin signalling in *human* skeletal muscle, as measured by phosphorylation of the insulin-receptor, AKT and AS160[119]. Furthermore, LPCAT3 is upregulated in primary muscle cells isolated from obese insulin-resistant patients, when compared to lean-insulin sensitive (LN) individuals[119]. This is consistent with the negative zscore (-1.64407; original data) or protective role of the allele, as rs1984564 is a tolerated missense variant for LPCAT3. This variant has never been previously implicated in *genetic* studies for T2D or related traits.

Chapter 5: Brief Summary and Conclusions

The relevant discussion and conclusions are available at the end of the respective chapters. Here, I will quickly summarize the major results and key points of the two analytic strategies introduced in this thesis.

Trained solely on promoter elements, I have developed a convolutional network model (peaBrain) that can learn to emulate the transcriptional machinery of a given cell type or tissue – allowing us to construct *in silico* DNA reporter assays. Focusing on promoter elements, I show the convolutional neural framework captures the majority of variance (>50%) in average gene expression, as assessed via a 10-fold cross-validation screen, and validate this performance on external datasets (captured on a different platform) and by predicting of expression of orthologous sequences in non-human Hominid primate tissues. By perturbing every nucleotide, I calculate non-coding impact scores using this *in silico* DNA reporter assay that correlate with evolutionary constraint and that are also predictive of disease-associated variation and allele-specific transcription factor binding. By leveraging the tissue-specificity of each assay, I further show how these tissue-specific scores can pinpoint functional tissues underlying complex traits, outperforming methods that depend on colocalization of eQTL and GWAS signals. I subsequently demonstrate the utility of this model in investigating the role of DNA and histone modifications in difficult-to-study processes, such as neural induction.

With a small extension to this *in silico* DNA reporter assay to incorporate small molecule fingerprints, this network architecture can also be used to predict the transcriptomic impact of small molecules in cell lines, with as few as 600 molecules (24 hours post-exposure). Predictive performance of this small molecule screen is on par with *in vitro* experimental replication of external test sets, allowing us to impute expression for all clinically approved molecules in the ChEMBL database. By training on the imputed expression profiles for all molecules first approved prior to the year 2000, I am able to retrospectively identify molecules that would later be assigned to the corresponding indications (post-2000), with 64-86% increase in F1-scores, 22-63% increase in precision, 75-94% increase in recall, and

13-19% increase in average precision (area under precision-recall curve), compared to current chemoinformatic/structure-based approaches.

Using the same model and with better performance, I can also predict the transcriptomic consequences of shRNA olgio-sequences (i.e. gene knockdowns). I use this shRNA peaBrain model to predict the consequences of 330,617 unique shRNA oligo-sequences (targeting 19,992 human genes) to enable the inference of regulatory networks, at a precision of 66-77% (AUROC = 87-93%, AUPRC = 39-52%, F1-score = 14-45%). I leverage these models to identify 212 new transcription factor-target interactions, of which 83% are supported by experimental ChIP-seq evidence. Finally, using an extension of the same architecture, I show how the framework outperforms linear models in predicting the consequences of genotype variation. I use the framework to identify putatively functional eQTLs that are missed by experimental high-throughput approaches that characterise variant function, such as massively parallel reporter assays (MPRAs), bi-allelic targeted STARR-seq (BiT-STARR-seq), and high-definition reporter assays (HiDRA). Chapters 2-3 describe one of the first attempts to recapitulate expression-based HTAs using a single *in silico* framework.

As with other deep learning approaches, there are limitations to peaBrain analysis; notably, that despite our best efforts for the rigorous quality control, unbiased error estimation, 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, I 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), validation with external replication test sets (LINCS) and validation using chromatin, TF-binding, and DNA accessibility annotations. 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 is encouraging evidence of model generalizability. Similarly, the correlation of the impact scores from Stage 1 analyses with evolutionary constraint and their utility in

predicting disease-associated mutations and allele-specific binding sites further underscores the true performance of peaBrain framework.

Altogether, the results from the peaBrain framework suggest that models for emulating high-throughput transcriptomic assays and ultimately understanding the effects of non-coding variation, small molecules, and shRNA on RNA abundance (and possibly more complex traits) can be built by relying more on automated machine learning, rather than hand-designed or selected features. Furthermore, the results highlight the variant sensitivity of the peaBrain model and its ability to identify putatively functional variants underlying *cis*-eQTL signals. More generally, peaBrain's performance in predicting mean abundance and individual variation further implicates the importance of the invariant genomic context and distance to the annotated TSS for interpreting the effects of non-coding variation (and the impact of small molecules) in a tissue-specific manner. By emulating the transcriptional machinery of human tissues/cells, peaBrain represents one of the first attempts to recapitulate expression-based HTAs using a single *in silico* framework. Our results are a first step towards a long-term goal of developing a general-purpose model that can learn to predict expression under any set of arbitrary constraints (e.g. for a tissue of interest, given a particular genotype, or after a fixed-time post-drug exposure).

In chapter 4, I introduce a new analytic strategy that leverages the convergence of expression and summary genetic association statistics to identify the causal tissue(s) underlying each T2D-related trait, to model the effects of genetic variation in a tissue-dependent manner, and to subsequently, discover 3 new loci that influence 4 traits in 2 different tissues. My analytic strategy, extending my work on coexpression differential dependence networks (xDDNs), outperforms tissue-specific chromatin states and expression quantitative trait loci (eQTL); neither of which alone is sufficient for the tasks. Furthermore, conditioning on functional annotations that are not specific to any cell type largely eliminates tissue-specific chromatin state enrichment. I demonstrate that the newly discovered risk loci are driven by tissue-specific xDDN annotations and as validation, that they have been confirmed in (larger) genetic studies for the same or closely-related traits and/or partially verified experimentally.

My thesis represents a first step towards my long-term goal of developing a general-purpose model that can learn to predict expression under any set of arbitrary constraints (e.g. for a tissue of interest, given a particular genotype, or after a fixed-time post-drug exposure). These general-purpose models can be used, among other things to map GWAS risk loci to physiologically-relevant tissue-specific mechanisms that underlie type 2 diabetes (T2D) and related traits.

Appendix I: Availability Statements and Supplementary Tables

Code availability statement

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

Data availability statement

Data used to train the models are available at: http://www.well.ox.ac.uk/~moustafa/peaBrain.shtml .

Supplementary Table 1. Schematic of the core 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.

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 2. Schematic of the peaBrain model used to predict the transcriptomic consequences of individual variation, which is composed of three separate networks connected by a dense layer prior to prediction. The network for the centre split is identical to the peaBrain model for the core promoter region; the networks for the upstream and downstream splits are identical to the peaBrain model for distal sequences.

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		

Appendix II: URLs

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

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

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

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

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

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

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

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

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

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

LINCS L1000 phase I (GEO), https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742

LINCS L1000 phase II (GEO), https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138

MPRA Supplementary Table,

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

peaBrain data and model, http://www.well.ox.ac.uk/~moustafa/peaBrain.shtml

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

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

RDKit, http://www.rdkit.org

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

Transcription factor binding sites (allele-specificity),
https://www.biorxiv.org/content/early/2018/02/01/253427.figures-only

REFERENCES

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

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

3 Subramanian, A. *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437-1452. e1417 (2017).

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

5 Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40**, D1100-D1107 (2011).

6 Landrum, G. (2006).

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

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

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

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

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

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

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

14 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).

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

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

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

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

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

20 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).

21 Collaboration, N. R. F. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants. *The Lancet* **387**, 1513-1530 (2016).

22 Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics* **42**, 579-589 (2010).

23 Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981 (2012).

24 Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* (2016).

25 Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology* **30**, 1095-1106, doi:10.1038/nbt.2422 (2012).

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

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

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

29 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).

30 Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *bioRxiv*, 161851 (2018).

31 Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS genetics* **13**, e1006646 (2017).

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

33 Wallace, C. Statistical testing of shared genetic control for potentially related traits. *Genetic epidemiology* **37**, 802-813 (2013).

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

35 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).

36 Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research* **24**, 14-24 (2014).

37 Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239 (2017).

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

39 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).

40 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).

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

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

43 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).

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

45 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).

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

47 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).

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

49 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).

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

51 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).

52 Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742-754 (2010).

53 Bovolenta, L. A., Acencio, M. L. & Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC genomics* **13**, 405 (2012).

54 Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nature methods* **9**, 796 (2012).

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

56 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).

57 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 https://www.nature.com/articles/ng.3979#supplementary-information (2017).

58 Moyerbrailean, G. A. *et al.* Which genetics variants in DNase-Seq footprints are more likely to alter binding? *PLoS genetics* **12**, e1005875 (2016).

59 Ford, E. E. *et al.* Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. *bioRxiv*, 170506 (2017).

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

61 D'haeseleer, P. How does gene expression clustering work? *Nature biotechnology* **23**, 1499-1502 (2005).

62 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863-14868 (1998).

63 Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**, 100-108 (1979).

64 Quackenbush, J. Computational analysis of microarray data. *Nature reviews genetics* **2**, 418-427 (2001).

65 Kohonen, T. & Somervuo, P. Self-organizing maps of symbol strings. *Neurocomputing* **21**, 19-30 (1998).

66 Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences* **96**, 2907-2912 (1999).

67 Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4**, 1128 (2005).

68 Oldham, M. C., Horvath, S. & Geschwind, D. H. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences* **103**, 17973-17978 (2006).

69 Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. & Ruzzo, W. L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977-987 (2001).

70 Rau, A. & Maugis-Rabusseau, C. Transformation and model choice for RNA-seq co-expression analysis. *bioRxiv*, 065607 (2016).

71 Dhillon, I. S. in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.* 269-274 (ACM).

72 Kluger, Y., Basri, R., Chang, J. T. & Gerstein, M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research* **13**, 703-716 (2003).

73 Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972-976, doi:10.1126/science.1136800 (2007).

74 Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature biotechnology* **27**, 199-204 (2009).

75 Baker, M. Gene data to hit milestone. *Nature* **487**, 282-283 (2012).

76 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

77 Margolin, A. A. *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7 Suppl 1**, S7, doi:10.1186/1471-2105-7-S1-S7 (2006).

78 Watson-Haigh, N. S., Kadarmideen, H. N. & Reverter, A. PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. *Bioinformatics* **26**, 411-413, doi:10.1093/bioinformatics/btp674 (2010).

79 Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53-65 (1987).

80 Gibbons, F. D. & Roth, F. P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome research* **12**, 1574-1581 (2002).

81 Datta, S. & Datta, S. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19**, 459-466 (2003).

82 Hubert, L. & Arabie, P. Comparing partitions. *Journal of classification* **2**, 193-218 (1985).

83 Xuan, N., Julien, V. & Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. (2011).

84 Vinuela, A. *et al.* Age-dependent changes in mean and variance of gene expression across tissues in a twin cohort. *bioRxiv*, 063883 (2016).

85 Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752-755 (2002).

86 Gerrits, A. *et al.* Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS genetics* **5**, e1000692 (2009).

87 Pierson, E. *et al.* Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS computational biology* **11**, e1004220, doi:10.1371/journal.pcbi.1004220 (2015).

88 Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358, doi:10.1093/bioinformatics/bts163 (2012).

89 Consortium, G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

90 Taneera, J. *et al.* Expression profiling of cell cycle genes in human pancreatic islets with and without type 2 diabetes. *Molecular and cellular endocrinology* **375**, 35-42 (2013).

91 Mahdi, T. *et al.* Secreted frizzled-related protein 4 reduces insulin secretion and is overexpressed in type 2 diabetes. *Cell metabolism* **16**, 625-633 (2012).

92 Dominguez, V. *et al.* Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic β cells. *Journal of Biological Chemistry* **286**, 4216-4225 (2011).

93 Gallagher, I. J. *et al.* Integration of microRNA changes in vivo identifies novel molecular features of muscle insulin resistance in type 2 diabetes. *Genome medicine* **2**, 1 (2010).

94 Jin, W. *et al.* Increased SRF transcriptional activity in human and mouse skeletal muscle is a signature of insulin resistance. *The Journal of clinical investigation* **121**, 918-929 (2011).

95 Roadmap, E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

96 Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature genetics* **43**, 246-252 (2011).

97 Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* **42**, 1118-1125 (2010).

98 Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature genetics* **42**, 508-514 (2010).

99 Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics* **41**, 703-707 (2009).

100 Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics* **43**, 333-338 (2011).

101 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-713 (2010).

102 Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature genetics* **44**, 991-1005 (2012).

103 Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics* **42**, 937-948 (2010).

104 Studies, I. C. f. B. P. G.-W. A. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103-109 (2011).

105 Filippi, S. & Holmes, C. A Bayesian nonparametric approach to testing for dependence between random variables. *arXiv preprint arXiv:1506.00829* (2015).

106 Morris, A., Voight, B., Teslovich, T. & Consortium, W. T. C. C. Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics* **44**, 981-990 (2012).

107 McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070 (2010).

108 Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American journal of human genetics* **94**, 559-573, doi:10.1016/j.ajhg.2014.03.004 (2014).

109 Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature genetics* **46**, 136-143, doi:10.1038/ng.2870 (2014).

110 Parker, S. C. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences* **110**, 17921-17926 (2013).

111 Consortium, D. S. D. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics* **46**, 234-244 (2014).

112 Go, M. J. *et al.* New susceptibility loci in MYL2, C12orf51 and OAS1 associated with 1-h plasma glucose as predisposing risk factors for type 2 diabetes in the Korean population. *Journal of human genetics* **58**, 362-365 (2013).

113 Ryu, J. & Lee, C. Association of glycosylated hemoglobin with the gene encoding CDKAL1 in the Korean Association Resource (KARE) study. *Human mutation* **33**, 655-659 (2012).

114 Hwang, J.-Y. *et al.* Genome-wide association meta-analysis identifies novel variants associated with fasting plasma glucose in East Asians. *Diabetes*, DB_140563 (2014).

115 Palmer, N. D. *et al.* Genetic variants associated with quantitative glucose homeostasis traits translate to type 2 diabetes in Mexican Americans: the GUARDIAN (genetics underlying diabetes in Hispanics) consortium. *Diabetes*, DB_140732 (2014).

116 Pascoe, L. *et al.* Common variants of the novel type 2 diabetes genes CDKAL1 and HHEX/IDE are associated with decreased pancreatic β-cell function. *Diabetes* **56**, 3101-3104 (2007).

117 Perry, J. R. *et al.* Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS genetics* **8**, e1002741 (2012).

118 Horikoshi, M. *et al.* Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation. *PLoS genetics* **11**, e1005230 (2015).

119 Ferrara, P. J. Deletion of LPCAT3 Promotes Enhanced Insulin Signaling in Skeletal Muscle. (2015).

Appendix III: Preprint

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).

ABSTRACT

Genome wide association studies (GWASs) for complex traits have implicated thousands of genetic loci. Most GWAS-nominated variants lie in noncoding regions, complicating the systematic translation of these findings into functional understanding. Here, we leverage convolutional neural networks to assist in this challenge. Our computational framework, peaBrain, models the transcriptional machinery of a tissue as a two-stage process: first, predicting the mean tissue specific abundance of all genes and second, incorporating the transcriptomic consequences of genotype variation to predict individual abundance on a subject-by-subject basis. We demonstrate that peaBrain accounts for the majority (>50%) of variance observed in mean transcript abundance across most tissues and outperforms regularized linear models in predicting the consequences of individual genotype variation. We highlight the validity of the peaBrain model by calculating non-coding impact scores that correlate with nucleotide evolutionary constraint that are also predictive of disease-associated variation and allele-specific transcription factor binding. We further show how these tissue-specific peaBrain scores can be leveraged to pinpoint functional tissues underlying complex traits, outperforming methods that depend on colocalization of eQTL and GWAS signals. We subsequently derive continuous dense embeddings of genes for downstream applications, and identify putatively functional eQTLs that are missed by high-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 was a single exception: in the nucleus accumbens (basal ganglia)**,** noncoding transcriptomic impact was correlated with accelerated evolution (**Supplementary Table 1**). While it is tempting to speculate about a plausible biological hypothesis to explain this exception, this result needs to be independently verified in other human samples before it is accepted as fact. 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.

peaBrain identifies functional architecture that is inaccessible with current high-throughput experimental assays.

First (**task E**)**,** we compared the predictive performance of peaBrain to that of a regularized linear model (an implementation of elastic net identical to that used in PrediXcan). RNA-seq samples from the GTEx dataset (n = 94 individuals after filtering) were pre-processed, residualised to account for cryptic relatedness, biological confounders, and technical variance, and rank transformed to normality (see **Online Methods**) before modelling. Model performance for both linear models and peaBrain was assessed by generating oos-r^2 for 5% of individuals randomly withheld from training and unrelated to individuals in the training set (repeated 3-10 times, depending on the how quickly the model reached the exit criteria and the performance of earlier repeats; see **Online Methods**). For each of the 816 genes with non-zero heritability (GCTA $p < 0.01$), we calculated the 95% confidence interval for the oos-r^2, defining a gene as successfully predicted if the entire oos-r^2 confidence interval exceeded zero to ensure we only consider genes with high-confidence models. Whilst regularized linear models were able to capture cis-heritability for 28 of the 816 genes, the equivalent number for peaBrain was 113. *Cis*-heritability for 3 genes was captured by both models, with the oos-r^2 confidence interval largely overlapping (**Supplementary Table 5**). **Supplementary Table 5** also tabulates the performance metrics (confidence and point estimates for oos-r^2 from both classes of models) and estimated GCTA heritability for all genes.

Having established the predictive ability of peaBrain, we were interested in whether we can use the best-performing peaBrain models to measure the impact of single variants, compared to DeepSEA log gold change (logFC) estimates and experimental log skew estimates from MPRA and BiT-STARR-seq (**Task F**). For all "captured" genes (n = 113), we selected all variants identified as significant eQTLs in the GTEx v6p univariate eQTL analysis (n = 16,019 variants; **see Methods**) and replicated the analysis with the Geuvadis dataset[31] (n = 17,279 variants for the EU population and n = 1601 variants for the

YRI [Yoruba from Ibadan, Nigeria] population). For each eQTL (including indels), we created pairs of 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 the exact cell line was not specified[12]. For each chromatin annotation, we assessed significance using a simple logistic model after rank-transformation of all estimates to normality (to ensure coefficients were comparable; see **Online Methods**). The coefficient of the model corresponded to the extent to which each approach was predictive of chromatin states/accessibility. More concretely, the logistic coefficients give the change in the log odds of the annotation overlap for a one-unit increase in the normalized score.

For peaBrain, as opposed to analysing the consequences of all possible variants/indels within 1Mbps input sequences for the "captured" 113 genes (which is computationally expensive), we focussed our analysis on all 23,595 univariately-significant eQTLs (from either the GTEx or Geuvadis datasets). We noted that variants with higher peaBrain estimates were significantly enriched in DNase accessible sites and transcriptionally active regions, and significantly depleted from heterochromatin and repressed sequences (**Table 2**). In contrast, the magnitudes of the MPRA log skew estimates were not significantly associated with any chromatin state or accessibility annotation after Bonferroni correction (**Table 2**). This absence of enrichment/depletion was consistent whether we analysed all variants assessed on the platform (n = 26,986 variants after excluding those with no match in Ensembl's VEP database; see **Online Methods**) or limited our analysis to the subset of variants also present in the peaBrain analysis (n = 1589 MPRA variants; i.e. univariately-significant eQTLs for the 113 "captured" genes). It is important to note that variants assessed on the MPRA platform were already selected, in part, because their eQTL status in the Geuvadis dataset; that is, excluding negative controls and LD-based selection, all variants assessed on the MPRA assays were univariately-significant eQTLs.

Similarly, variants with high magnitudes of the BiT-STARR-seq log skew estimates were not significantly enriched in transcriptionally active chromatin (or depleted from repressed/quiescent intervals), irrespective of whether we assessed performance on all variants assessed on the platform (n = 43,494) or limited to univariately-significant eQTLs for the 113 "captured" genes (n = 621). Using

nominal p-value thresholds, HiDRA performed better than either MPRA or BiT-STARR-seq when 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

is encouraging evidence of model generalizability. Similarly, the correlation of the impact scores from Stage 1 analyses with evolutionary constraint and their utility in predicting disease-associated mutations and allele-specific binding sites further underscores the true performance of peaBrain framework.

All together, the results from the Stage 1 and Stage 2 of the peaBrain framework suggest that models for understanding the effects of non-coding variation on RNA abundance (and possibly more complex traits) can be built by relying more on automated machine learning, rather than hand-designed or selected features. Furthermore, the results highlight the variant sensitivity of the Stage 2 peaBrain model and its ability to identify putatively functional variants underlying cis-eQTL signals. More generally, peaBrain's performance in predicting mean abundance and individual variation further implicates the importance of the invariant genomic context and distance to the annotated TSS for interpreting the 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 has also been 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

convolutional neural networks, connected by a dense fully-connected layer prior to the output neuron (**Supplementary Table 7**). The input 1Mbps sequence is split into three inputs: 0.48Mbps upstream, 4kbps core promoter, and 0.48Mbps downstream sequences. The 4kbp core promoter is the input to a CNN with identical structure and hyperparameters as described for Stage 1 peaBrain model (described above in detail). The upstream and downstream sequences are input to networks with identical architecture, but different pooling hyperparameters: a pool size of 100 for the first pooling layer, 50 for the second, and 10 for the last. Number of filters was consistent between all networks (n = 11). The fully connected output from each sequence is concatenated, before one penultimate fully-connected layer and a single output node. Unlike the Stage 1 peaBrain models, Stage 2 models are trained to predict the differences between individuals (rather than direct prediction of expression). As humans are diploid, for each individual, the input sequence was the sum of the one-hot encoding of each of the 1Mbps sequences corresponding to the "maternal" and "paternal" sequences ; phasing did not matter because the sum was consistent. A separate Stage 2 model was constructed for each gene with significant non-zero heritability (see **text**). For any pair of individuals, A and B, the input sequence was defined as the difference between the one-hot encoded sequences, with the corresponding output as the difference between the two individuals. We included both differences, (A – B) and (B – A), during training. After removing individuals with cryptic relatedness (see GCTA analysis below), the GTEx dataset was randomly split into train and test individuals (95% of subjects for training and 5% for testing), with the model trained on all the pairwise differences between train individuals and tested on all pairwise differences between test individuals. The training set was further sub-divided into training and validation sets, with the latter used to exit early after a minimum 100-epoch training. As described below, overall model performance was assessed using the oos-r^2 on five to ten random repetitions of 95/5 train/test splits; the number of repetitions was dependent on how quickly each model reached exit criteria.

Elastic net (regularized linear) models. We used an additive genetic model as our baseline comparison as described elsewhere[6]. Briefly, for each gene, an elastic net model was used to model 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.

Predictive 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

allelic skew from the combined MPRA LCL analysis (alt/ref), was used as the MPRA log skew estimate. The BiT-STARR-seq data was similarly processed (see **URLs**). As with the peaBrain and eQTL analysis, significance of the spearman (rank) correlation between each of the experimental assays allelic log skews and the univariate eQTL coefficients was assessed using the cor.test function. To obtain the logFC for GM12878 annotations, a vcf file of the corresponding eQTLs was uploaded to the DeepSEA platform (see **URLs**).

Comparison of peaBrain, MPRA, BiT-STARR-seq and HiDRA. The core 15-state model and DNAse accessibility annotations for the GM12878 EBV-transformed lymphoblastoid cell line (LCLs) were downloaded from the Roadmap project (see URLs). HiDRA data was downloaded from the GEO series GSE104001 (see URLs). HiDRA estimate was defined as the log fold change in average counts between the alternate and reference group (after normalizing for DNA count); direction did not matter as only the magnitude was used in this analysis. The MPRA and BiT-STARR-seq data was downloaded and pre-processed as described above. Notably, the chromatin states/DNA accessibility annotations, the HiDRA, and MPRA estimates were derived from the same cell line; that is, there is possibility of overestimating the performance of either method. For any given annotation, we assessed the predictive ability of the magnitude of the variant estimate (from any of the three approaches) to predict whether the variant overlapped with the annotation. The magnitude of the variant estimate for each approach (peaBrain, HiDRA, BiT-STARR-seq, and MPRA) corresponded to either the transcriptomic impact or activity of that variant. Only the absolute magnitude, after rank-transformation to normality, of each variant was used in modelling. For each approach and for each annotation, a logistic model was used (fitted using the glm function in R, family = "binomial") and the confidence interval was obtained using the confint function (derived from profiling the likelihood function). For the granular variant-level assessment, annotations were downloaded from RegulomeDB (dbSNP 141; see URLs).

Code availability statement

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

Data availability statement

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

URLs

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

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

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

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

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

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

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

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

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

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

MPRA Supplementary Table,

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

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

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

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

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

Transcription factor binding sites (allele-specificity),

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

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.

TABLES

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.

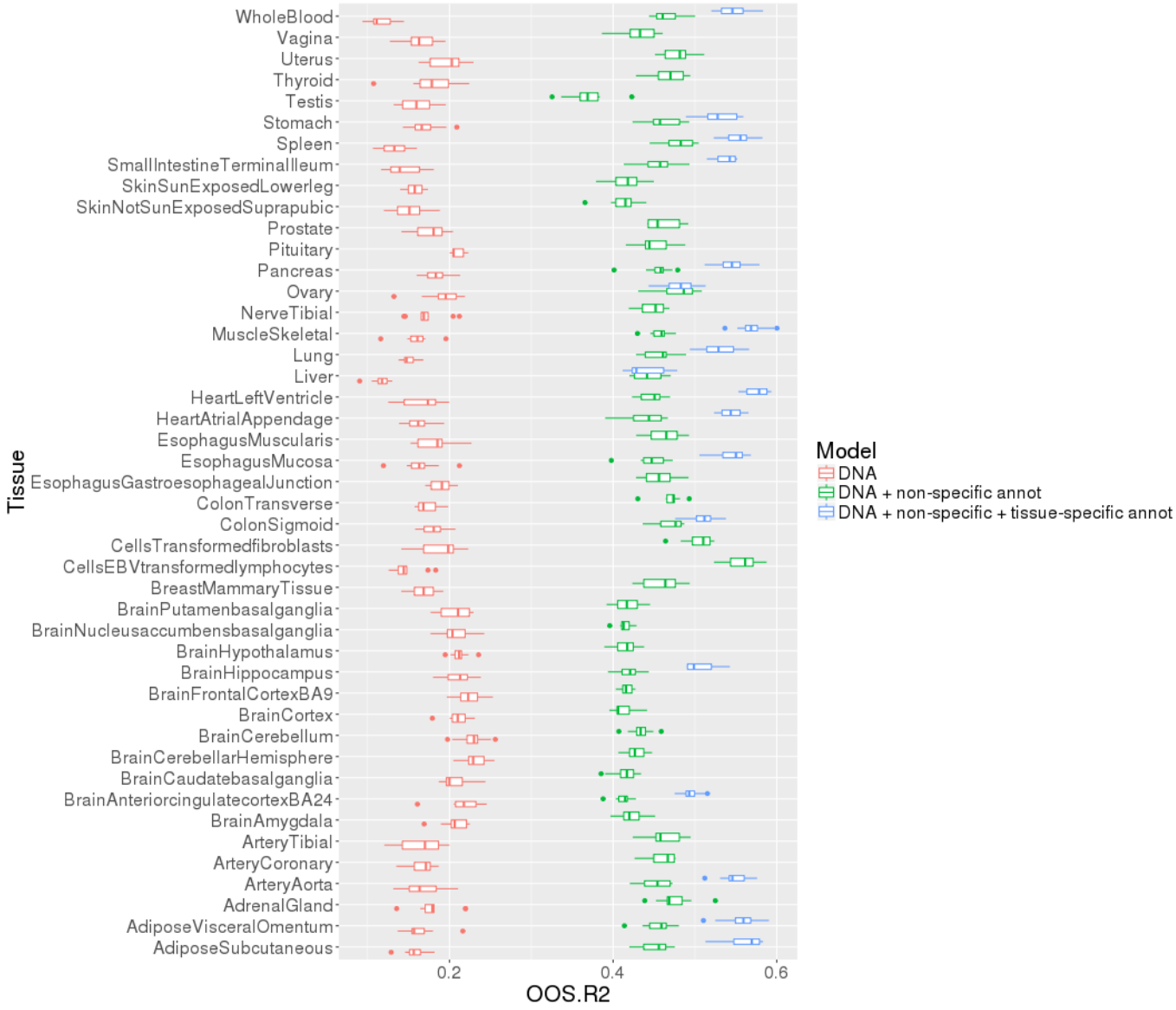
	peaBrain	MPRA log-skew		HiDRA		BiT-STARR-seq log-skew	
		all	shared	all	shared	all	shared
DNase accessibility	**0.12 (3.76 x10^{-5})**	0.01 (0.463)	0.21 (2.64 x10^{-2})	0.01 (0.352)	-0.05 (0.732)	**-0.05 (2.92 x10^{-13})**	0.09 (0.438)
TssA	**0.32 (<2 x10^{-16})**	0.03 (0.218)	0.25 (2.05 x10^{-2})	-0.02 (0.140)	0.01 (0.934)	0.00 (0.965)	0.09 (0.360)
TssAFlnk	0.10 (1.55 x10^{-2})	0.06 (4.72 x10^{-2})	0.29 (2.17 x10^{-2})	0.03 (4.01 x10^{-2})	0.61 (4.23 x10^{-3})	-0.02 (1.28 x10^{-2})	0.03 (0.788)
TxFlnk	-0.13 (7.69 x10^{-2})	-0.01 (0.81)	-0.16 (0.672)	0.08 (0.141)	1.61 (5.59 x10^{-2})	-0.02 (0.424)	-0.06 (0.787)
Tx	-0.02 (0.297)	-0.04 (4.91 x10^{-2})	-0.05 (0.471)	-0.14 (6.13 x10^{-3})	-0.80 (9.58 x10^{-2})	0.01 (0.571)	-0.07 (0.325)
TxWk	0.04 (1.42 x10^{-2})	-0.01 (0.662)	-0.01 (0.923)	0.01 (0.630)	0.48 (0.136)	**0.03 (4.66 x10^{-4})**	0.00 (0.973)
EnhG	0.04 (0.547)	-0.12 (6.97 x10^{-3})	-0.26 (0.167)	0.04 (0.529)	-0.91 (4.35 x10^{-2})	-0.05 (1.85 x10^{-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 x10^{-2})	**-0.03 (1.39 x10^{-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 x10^{-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 x10^{-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 x10^{-2})	NA	-0.05 (0.653)	-0.81 (0.440)	-0.03 (0.230)	-0.26 (0.735)
ReprPCWk	**-0.25 (<2 x10^{-16})**	-0.033 (0.133)	NA	-0.05 (3.73 x10^{-2})	-1.00 (6.80 x10^{-2})	-0.02 (2.45 x10^{-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 x10^{-4})**	-0.01 (0.896)

Table 3. In the 113 "captured" genes, peaBrain estimates can significantly delineate variants with established regulatory function (**Task H**). The log-skew estimates from the experimental assays, both across all variants assessed on each platform ("all") and limited to eQTLs for the 113 "captured" genes ("shared"), are uninformative. Both the peaBrain estimates and log-skew for the experimental assays were rank-transformed to normality to facilitate comparison between the methods. 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.

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

FIGURES

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.



SUPPLEMENTARY TABLES

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

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

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}

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

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

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

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

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).

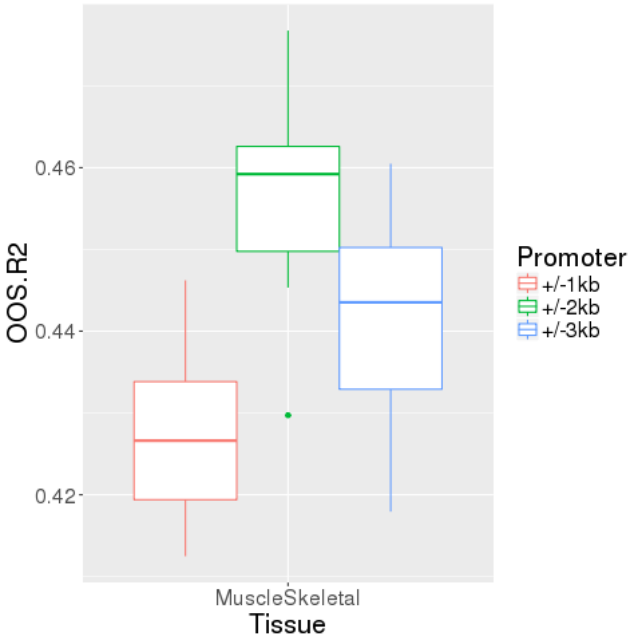
Input Sequence: 4000 x r channels
1st Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
1st Pooling Layer: Pool Size = 5, Pad = 1
2nd Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation
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.

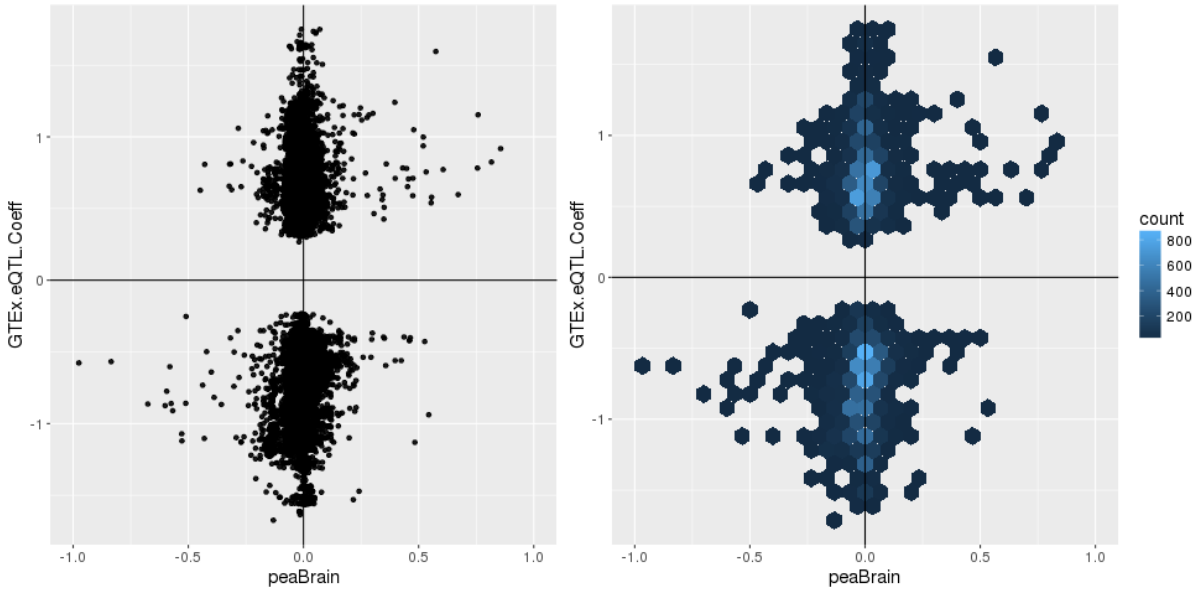
Input Sequence 1Mbps x 4 channels		
Upstream Split 0.498Mbps x 4 channels	**Centre Split** 4kbps x 4 channels	**Downstream Split** 0.498Mbps x 4 channels
1st Convolutional Layer: Number of Filters = 11 Size of Filters = 5 Stride = 1, Pad = 2 Leaky Rectify Activation	**1st Convolutional Layer:** Number 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		

SUPPLEMENTARY FIGURES

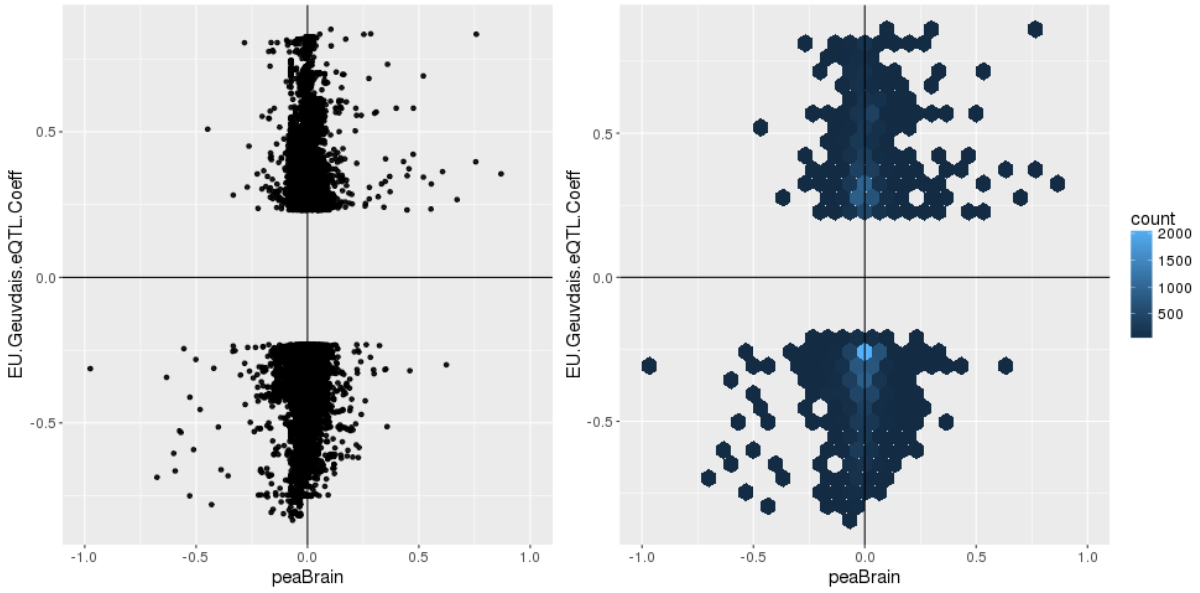
Supplementary Figure 1. Using the class-B peaBrain model for MuscleSkeletal (largest tissue by sample count in GTEx), the 4kbps promoter sequence (+/– 2kbps of annotated TSS) outperforms both 2kbps (+/– 1kbps) and 6kbps (+/– 3kbps) promoter sequences in predicting mean gene abundance.



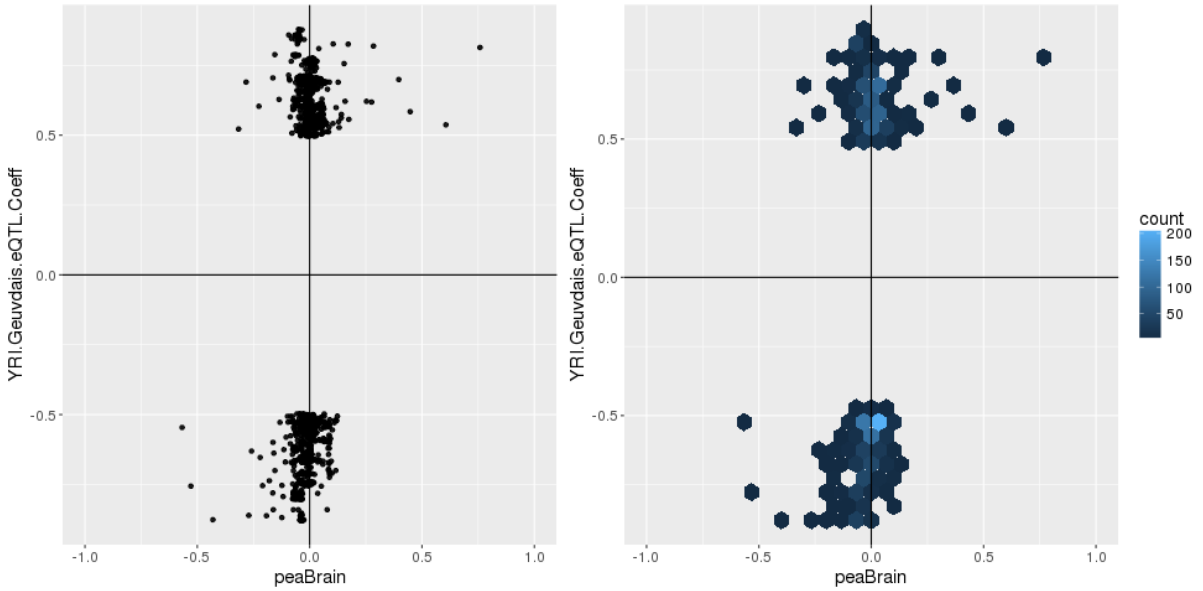
Supplementary Figure 2. Scatter (*right*) and hexa-bin (*left*) plots of variant-expression effects as estimated in LCLs by peaBrain (limited to genes whose 95% confidence interval for the oos-r^2 is entirely above 0; n = 113 genes; **Task F**). Each point corresponds to a variant that is univariately 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}).



Supplementary Figure 3. 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 EU-Geuvadis eQTL analysis (n = 17,279 eQTLs). The y-axis is the magnitude of the univariate EU-Geuvadis eQTL coefficient for the corresponding variant. The correlation between the EU-Geuvadis coefficient and the peaBrain prediction is positive and significant (Spearman's rho = 0.10; p = 9.60 x10^{-38}).



Supplementary Figure 4. 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 YRI-Geuvadis eQTL analysis (n = 1601 eQTLs). The y-axis is the magnitude of the univariate YRI-Geuvadis eQTL coefficient for the corresponding variant. The correlation between the YRI-Geuvadis coefficient and the peaBrain prediction is positive and significant (Spearman's rho = 0.18; p = 8.64 x10^{-13}).



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).

SUPPLEMENTARY NOTE 1

Regularized linear models are poor models for predicting mean abundance from core promoter sequences (10-fold cross-validated average oos-$r^2 < 0$). To assess how simple linear models compare to "deep learning" models, we fitted linear models by minimizing a regularized empirical squared loss, with a L2 (squared euclidean norm) penalty, using stochastic gradient descent from python's sci-kit learn (sklearn). The input of the class B models (non-specific annotated DNA model; 1D matrix with 32 channels) was flattened prior to modelling, i.e. the mean abundance each gene was modelled as a linear function of 128,000 variables (most were binary variables). For brevity, we limited our analysis to skeletal muscle, the tissue with the largest number of samples in GTEx. The model was fit using the partial fit function from SGDRegressor with sklearn using default parameters; the same exit criteria as described in **Online Methods** were applied. We noted that 10-fold cross-validated average oos-r^2 was consistently below zero, with performance considerably worse than the peaBrain model. The gap in performance was considerable (precludes visualization). We also repeated this analysis using a single dense neural network layer with linear activations, and noted that the 10-fold cross-validated average oos-r^2 was below 0; a single dense layer with linear activations is equivalent to a simple linear model.

Fully connected neural networks are slower and more memory-intensive, with slightly worse performance than convolutional neural networks; non-linear activations for convolutional layers improve performance over linear activations. We compared fully connected dense neural networks to the convolutional neural networks at the heart of peaBrain (**Figure 1**) for class B skeletal muscle model. We replaced each pair of convolutional-pooling layers with a single dense layer; the number of neurons in the three dense layers was limited only by GPU memory. The penultimate layer and single output neuron were kept consistent between the models. Using the skeletal muscle Class B input, we noted that fully connected performed slightly worse (oos-r^2 = 0.42) than the convolutional neural networks (oos-r^2 = 0.46). Increasing the number of layers allows fully connected neural network to reach performance parity with CNNs, at the cost of increased memory and computational cost (not feasible for Stage 2 peaBrain analyses). We subsequently wanted to assess the importance of the non-

linear activation for the convolutional layers of peaBrain. We constructed an identical model, but replacing all CNN activations with linear functions and noted that performance for this model was consistently worse (oos-r^2 = 0.43) than the classical peaBrain architecture (oos-r^2 = 0.46). It is important to note that this is not an exhaustive search of the ideal set of parameters, but an exploratory analysis to begin to understand peaBrain's performance.

DNA sequence, annotated with experimentally-derived TFBS, from core promoter sequences are insufficient to predict mean abundance with high accuracy – epigenetic/histone markers contain the bulk of the information and are not readily accessible from the DNA sequence alone. We were interested in determining the contribution of epigenetic/histone makers, alongside more general genomic annotations (such as coding sequences), in predicting the mean abundance of genes. In particular, we wanted to explore whether the DNA sequence alone was sufficient to predict expression in skeletal muscle. We noted that increasing the number of convolutional layers or the number of filters did not improve model performance (**Figure 1**). Explicitly incorporating TFBS into the model (i.e. annotating the DNA only and explicitly with TFBS) only improved performance slightly (oos-r^2 = 23%), and was still considerably worse than the full class B model with epigenetic/histone marker annotations (oos-r^2 = 46%; **Figure 1**). (Class-A DNA-only models had an average oos-r^2 of 16% for skeletal muscle; class-C models annotated with tissue-specific information had an average oos-r^2 of 57%.) The TFBS were collected from the Gene Transcription Regulation Database (GTRD) v17.4 with data on 476 human transcription factors and included peak calling with four different software (MACS, SISSRs, GEM, and PICS). In addition to including the processed peak calls, we also incorporated clusters (i.e. peaks merged for the same transcription factor but under different experimental conditions) and meta-clusters (i.e. non-redundant peaks synthesized from all four methods). This absence of improvement suggests that peaBrain model already recognizes many of the TFBS; identified by the convolutional filters inherent to the model architecture. These results indicate that experimentally-derived epigenetic and genomic annotations add information to that contained in the DNA sequence alone. As described in the main text, this is broadly consistent with the observation that other

convolutional neural networks models like DeepSEA are better at predicting TFBS (median AUC = 0.958) than at predicting histone modifications (median AUC = 0.856)[1].

peaBrain score out-performs existing measures in predicting allele-specific transcription factor binding. As with **tasks A and B** (described in the **Main Text**), we compared the performance of the non-tissue-specific peaBrain score to predictions by CADD and EIGEN in predicting allele-specific binding, after accounting for allele frequency and evolutionary conservation. We assessed performance of the three non-coding metrics across 6675 sites in core promoter regions after filtering for duplicate sites[2]; 1896 of which exhibited allele-specific binding at an unadjusted binomial $p < 0.05$ (see **Online Methods**). We noted that only peaBrain impact score was significantly predictive of allele-specific binding sites (coefficient = 35.38 [12.00, 58.67]; p = 0.003; see **Table 1** in **Main Text**); relaxing the binomial p-value threshold (i.e. increasing the number of sites considered as allele-specific) brings the other non-coding metrics to significance. peaBrain's discriminative ability to identify allele-specific binding sites is consistent with our earlier observation that explicitly adding TFBS annotations did not improve the model. Notably, peaBrain's ability indicates that average expression of all genes in a single tissue and the reference genome is sufficient to learn both TFBS and allele-specific binding.

To further investigate peaBrain's ability to identify allele-specific binding sites, we compared peaBrain impact scores to predictions by methods specifically designed to predict TFBS, including two neural-network methods (DeepBind[3] and DeepSEA[1]), two kmer-based variant scoring methods (gkmSVM[4] and GERV[5]), and three position-weighted matrices (PWM)-related methods[2]. These methods depend on modelling TF ChIP-seq data in various ways and may have multiple models for the same TF. After confirming the predictive ability of these methods to identify allele-specific binding sites, we noted that peaBrain scores positively correlated only with GERV measures, a kmer-based variant scoring algorithm (**Figure 2**). Unlike the other methods, peaBrain (and GERV) do not assume the existence of canonical motifs and learn TFBS by modelling sequences (or kmers) directly (i.e. not simply by modelling the absence or presence of a ChIP-seq peak). In contrast, for both DeepBind and DeepSEA, we noted positive correlation with at least one PWM-method. These methods generally assume the

existence of canonical TF binding sites and predictions are based on the extent of perturbation of those motifs. While this comparison is limited to variants for which data was available, the peaBrain results suggest that explicitly characterizing TF motifs is not necessary to understand the consequences of sequence variation on TF binding and transcriptional dysregulation.

Neural activations of penultimate layer of peaBrain model can be used to construct an embedding from the genes that encodes correlation information. Having demonstrated the predictive ability of the peaBrain model (see **Main text** for details), we were subsequently interested in using the activations from the penultimate layer of the model as a continuous (and compressed) representation of the genes. These neural activations capture both the annotated DNA (input) and its additive contributions to tissue- and phenotype-specific abundance (output) in a compressed form amenable to downstream analyses. Furthermore, as these vectors were obtained from a regression model, they readily capture only the salient portions of DNA abundance encoded in the annotated-genome (the weights of the model corresponding to the transcription factor that regulate and interact with this genome). Because of model choice, the mean abundance of each gene was encoded as a linear combination of the vector elements, *i.e.* the output of the regression model. As with our earlier analyses, for brevity, we limited our analysis of the properties of the embeddings to class B models for skeletal muscle. We observed, for the skeletal muscle embeddings, that pairwise cosine similarity between these dense gene representations corresponded to the measured RNAseq correlation between the gene pair. After excluding self-correlations and weakly correlated genes (RNAseq rho < 0.5), we noted that the cosine similarity of the embedding was significantly correlated (Spearman's rho = 0.18; $p < 2.2 \times 10^{-16}$) to the experimentally RNAseq-derived correlation. This suggests the annotated-DNA model, without supervision, imposes a linear structure on this vector space: the angle between the vectors corresponds to the co-regulation of the gene pair.

peaBrain-derived gene embeddings also encode membership to pathways and other curated gene sets. We were interested in further exploring the utility of these embeddings in other applications. We noted that this dense representation from the class B skeletal muscle pea Brain model encodes

membership to the MSigDB Hallmark curated gene sets (average 10-fold cross-validated for all pathways AUC = ~0.70, **Table 1**), suggesting that the representations themselves, not only encode abundance and regulatory information, but also functional relationships. (We filtered pathway sets not relevant to the tissues, such as "PANCREAS_BETA_CELLS", "COMPLEMENT", or "SPERMATOGENSIS"). Taken all together, this suggests the gene embeddings capture both the annotated DNA (input) and its additive contributions to tissue-specific abundance (output) in a compressed form amenable to downstream analyses (e.g. network-based analyses).

Supplementary Note 1 – Table 1. Tabulated 10-fold cross-validated AUC for genomewide pathway membership predictions using class B MuscleSkeletal Embeddings.

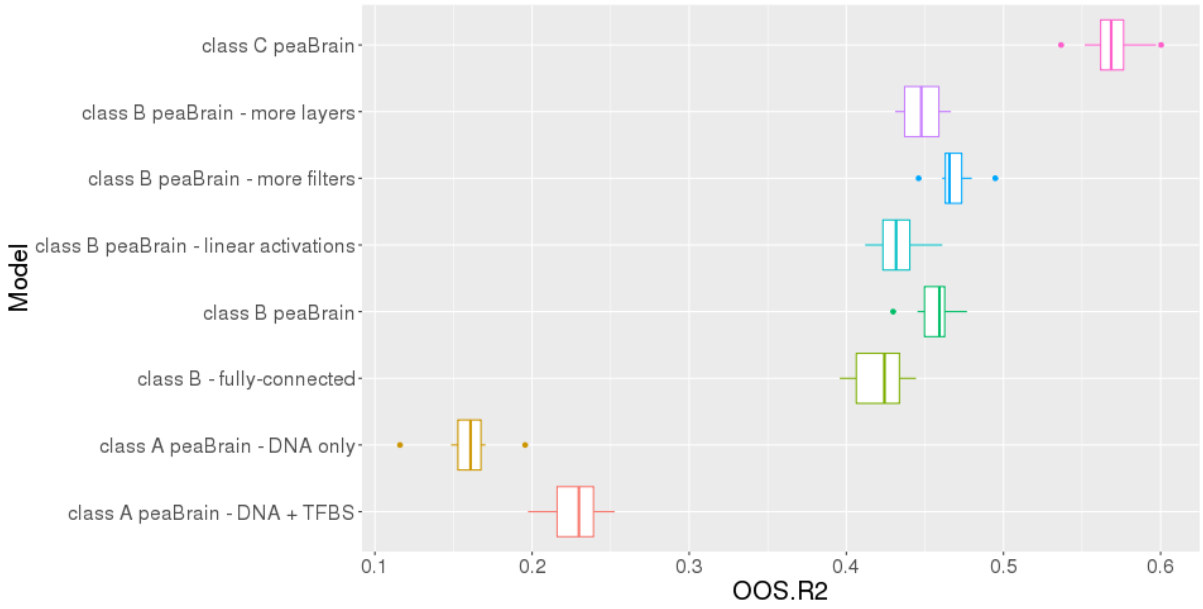
Hallmark Gene Set	10-fold cross-validated average auc
MYC_TARGETS_V1	0.80
MYC_TARGETS_V2	0.79
G2M_CHECKPOINT	0.77
UNFOLDED_PROTEIN_RESPONSE	0.76
OXIDATIVE_PHOSPHORYLATION	0.76
MTORC1_SIGNALING	0.76
EPITHELIAL_MESENCHYMAL_TRANSITION	0.74
MITOTIC_SPINDLE	0.74
E2F_TARGETS	0.73
REACTIVE_OXIGEN_SPECIES_PATHWAY	0.73
TNFA_SIGNALING_VIA_NFKB	0.72
PROTEIN_SECRETION	0.72
TGF_BETA_SIGNALING	0.71
UV_RESPONSE_DN	0.71
PI3K_AKT_MTOR_SIGNALING	0.71
DNA_REPAIR	0.71
HYPOXIA	0.70
P53_PATHWAY	0.69
APOPTOSIS	0.68
APICAL_JUNCTION	0.68
ADIPOGENESIS	0.67
MYOGENESIS	0.66
IL2_STAT5_SIGNALING	0.66
ANGIOGENESIS	0.66
GLYCOLYSIS	0.65
PANCREAS_BETA_CELLS	0.65
ANDROGEN_RESPONSE	0.65
KRAS_SIGNALING_DN	0.64
HEME_METABOLISM	0.64
CHOLESTEROL_HOMEOSTASIS	0.64
HEDGEHOG_SIGNALING	0.63
APICAL_SURFACE	0.63
UV_RESPONSE_UP	0.63
INTERFERON_GAMMA_RESPONSE	0.63
ESTROGEN_RESPONSE_EARLY	0.62
INTERFERON_ALPHA_RESPONSE	0.62
ESTROGEN_RESPONSE_LATE	0.61
NOTCH_SIGNALING	0.61
KRAS_SIGNALING_UP	0.61
INFLAMMATORY_RESPONSE	0.61
COAGULATION	0.60
SPERMATOGENESIS	0.60
WNT_BETA_CATENIN_SIGNALING	0.58

PEROXISOME	0.58
IL6_JAK_STAT3_SIGNALING	0.57
ALLOGRAFT_REJECTION	0.57
COMPLEMENT	0.57
BILE_ACID_METABOLISM	0.56
XENOBIOTIC_METABOLISM	0.56
FATTY_ACID_METABOLISM	0.56

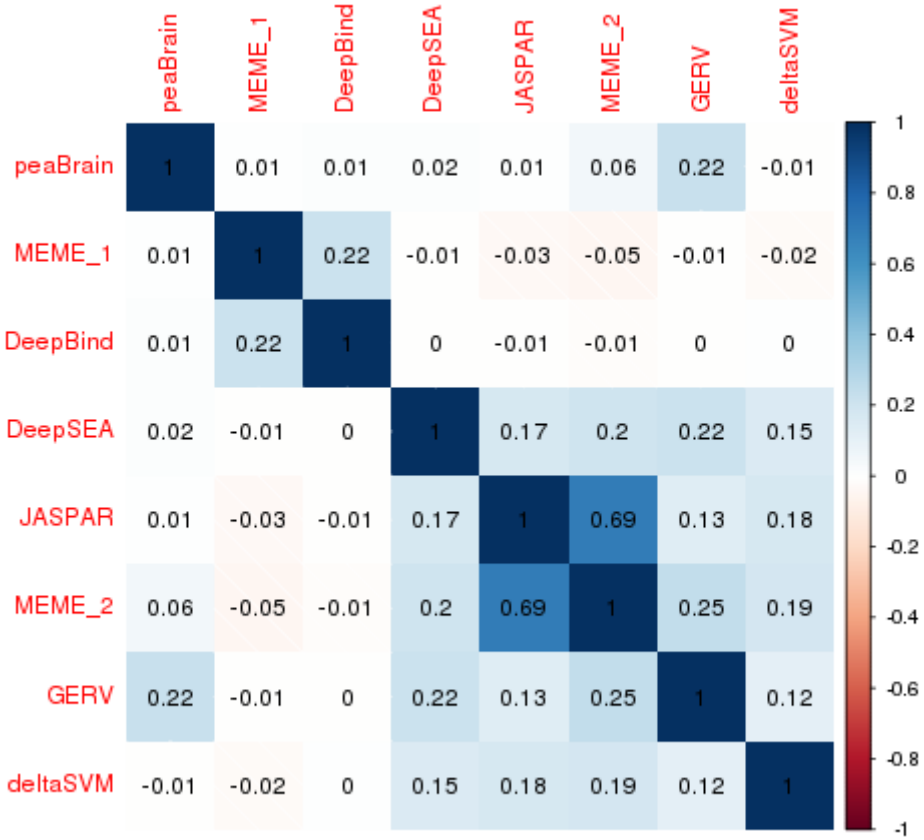
293

294

Supplementary Note 1 – Figure 1. Boxplots of 10-fold cross-validated oos-r^2, as assessed in skeletal muscle, for class A models (labelled as "class A peaBrain – DNA only"), class A with TFBS annotations (labelled as "class A peaBrain – DNA+TFBS"), class B models with tissue-agnostic annotations ("class B peaBrain – CNNs"), fully connected neural networks ("class B – fully-connected"), class B models with linear activation functions ("class B peaBrain – linear activations"), class B models with increased number of layers ("class B peaBrain – more layers"), class B models with increased number of filters ("class B peaBrain – more filters"), and class C models with tissue-specific annotations ("class C peaBrain").



Supplementary Note 1 – Figure 2. Rank correlation plot for TF-binding algorithms and the peaBrain impact score. JASPAR, MEME_1 and MEME_2 are PWM-approaches.



REFERENCES

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

2 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).

3 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).

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

5 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).

SUPPLEMENTARY NOTE 2

Stage 2 peaBrain predictions were significantly and positively correlated with univariate eQTL coefficients. Having established the predictive ability of peaBrain, we were interested in whether we can use the best-performing peaBrain models to measure the impact of single variants, compared to DeepSEA logFC estimates and MPRA log skew estimates (**Task F**). For all "captured" genes (n=113), we selected all variants identified as significant eQTLs in the GTEx v6p univariate eQTL analysis (n=16,019 variants; **see Online Methods**) and replicated the analysis with the Geuvadis dataset[1] (n=17,279 variants for the EU population and n = 1601 variants for the YRI [Yoruba from Ibadan, Nigeria] population). Unlike GTEx v7, eQTLs from GTEx v6p were derived from genotyping arrays (Illumina OMNI 5M or 2.5M) and thus did not include WGS used to train the peaBrain model. The Geuvdais dataset (both EU and YRI populations) were eQTLs derived from WGS from a set of non-overlapping subjects. For each eQTL (including indels), we created pairs of 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. peaBrain predictions 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**). Results were consistent across all datasets when we relaxed our heritability to include genes with GCTA $p < 0.05$: Spearman's rho equal to 0.08 ($p = 3.02 \times 10^{-25}$), 0.07 ($p = 3.65 \times 10^{-25}$), and 0.18 ($p = 1.68 \times 10^{-13}$) for the GTEx, EU-Geuvadis, and YRI-Geuvdais univariately-significant eQTLs, respectively. Alongside this positive correlation, we observed that many variants with large coefficients across all three datasets had small peaBrain predictions (**Supplementary Figures 2-4**). This shrinkage in peaBrain estimates is consistent with the appreciable LD between eQTL-associated variants at any given locus, such that only a subset of significantly-associated variants are actually functional. Whilst univariate linear models are not capable of distinguishing between functional versus "hitchhiker" variants, the joint modelling of the input sequence in Stage 2 peaBrain allows a more direct assessment of the function of each variant. In

comparison, we noted that the MPRA log skew estimates did not correlate with the univariate eQTL coefficients in the GTEx (rho = 0.014; p = 0.60), the EU-Geuvadis (rho = -0.018; p = 0.58), or the YRI-Geuvadis (rho = 0.011; p = 0.82) eQTL analyses. The BiT-STARR-seq log skew estimates also did not correlate with the univariate eQTL coefficients in the GTEx (rho = -0.031; p = 0.50), the EU-Geuvadis (rho = 0.035; p = 0.44), or the YRI-Geuvadis (rho = -0.112; p = 0.32) eQTL analyses. Similarly, for GM12878 (LCL) annotations, we noted that the maximum log fold changes from DeepSEA did not correlate with the magnitude of the eQTL coefficients in the GTEx analysis (rho = -0.001; p = 0.92), the EU-Geuvadis analysis (rho = -0.010; p = 0.18), and the YRI-Geuvadis (rho = 0.016; p = 0.52). The DeepSEA results suggest that simply predicting chromatin effects and TFBS is not sufficient for predicting the transcriptomic consequences of sequence variation for this set of selected genes. This is consistent with experimental evidence suggesting that most genetic variants in DNase footprinted (and other annotated) regulatory regions are silent[2], and the fact that histone modifications (e.g. methylation) alone are frequently insufficient to transcriptionally perturb promoters[3]. However, it is important to note that DeepSEA was trained on the reference genome (i.e. not exposed to genotype variation), while Stage 2 peaBrain was trained to predict differences in expression as a function of differences in sequence between any pair of individuals.

REFERENCES

1 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).

2 Moyerbrailean, G. A. *et al.* Which genetics variants in DNase-Seq footprints are more likely to alter binding? *PLoS genetics* **12**, e1005875 (2016).

3 Ford, E. E. *et al.* Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. *bioRxiv*, 170506 (2017).

SUPPLEMENTARY NOTE 3

Small molecule extension of the peaBrain DNA-reporter assay captures ~84% of the variance in the average molecule perturbed gene expression profiles of L1000 landmark transcripts.

To predict the transcriptomic perturbations of small molecules, we extended the peaBrain DNA reporter assay to incorporate molecular fingerprints (boolean arrays/bitmaps [i.e. a list of 1s and 0s] that are characteristic of the molecule structure), in addition to the DNA core promoter sequence as input (which directly accounted for copy number and genotypic variation, see **Online Methods**). As highlighted in **Table 1,** the small molecule extension can be described as:

promoter + molecular fingerprint → expression

where the fingerprint can be thought of as the "genotype" of the molecule. Given a core promoter sequence and a molecular fingerprint, the small molecule peaBrain model will predict the expression of the corresponding gene. We limited our analysis to molecular fingerprints – encoded as a standard bit vectors – generated using Morgan's algorithm[1] from the molecular-input line-entry system (SMILES) specification for each molecule. For each molecule profile (at 10uM dosage), we rank-normalized expression (averaging the replicates) and min-max scaled (per gene) prior to training (see **Online Methods** for details).

Supplementary Note 2 – Table 1. Overview of small molecule extension and its biological applications. The peaBrain small molecule model is a simple extension of the original peaBrain DNA reporter. Internal testing procedure describes the approach by which the model was trained and assessed on the training set. We filtered the phase II LINCS dataset to exclude molecules in phase I. The phase II (external set) is used only once: to assess performance once model weights are learnt using the training set.

peaBrain small molecule model: promoter + molecule fingerprint → expression
Description: For a given cell type, small molecule, and promoter element, predict expression of gene. Simple extension of the peaBrain DNA reporter model (described above).

Technical Overview
Objective: Assess the predictive performance of the peaBrain small molecule model

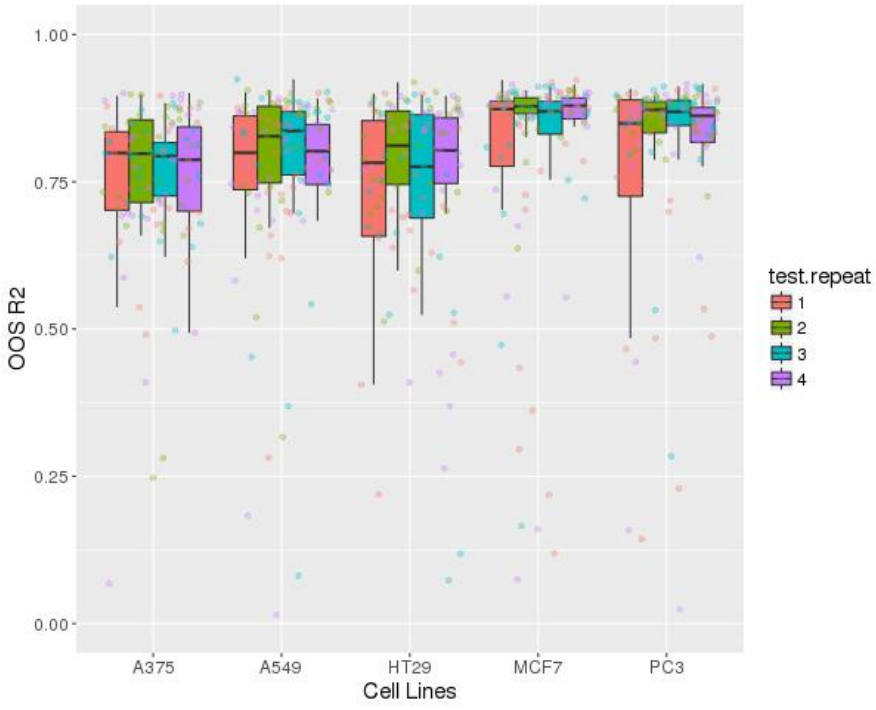
Training set	Internal Assessment	Types of Models Constructed	External Assessment	Notes
GSE92742 LINCS phase I	4-fold Monte Carlo validation (5% testing)	Single model: promoter DNA sequence + molecule fingerprint	*GSE70138* LINCS phase II	Models constructed for single dose (10uM)

Biological Applications:
1. Drug re-positioning (can retrospectively predict approval for relevant indications)
2. Drug discovery (identify compounds that selectively inhibit cancer hallmark genes)

Training was conducted with expression profile data downloaded from the LINCS phase I L1000 Connectivity Map (GSE92742) dataset on NCBI's Gene Expression Omnibus (GEO; see **URLs**)[2]. (The LINCS phase II [*GSE70138]* dataset was used as the external test set.) We limited our investigation to cell lines for which genotype, copy-number, and both phase I and phase II data was available (n = 5 cell lines: A375, A549, HT29, PC3, and MCF7). Both phase I and phase II LINCS projects use the L1000 assay, which measures the expression profile of 978 landmark transcripts (using a microarray-like technology). We discarded any imputed transcripts before training (see **Online Methods**). Genotype and copy-number data for all 5 cell lines included in the analysis, identified by PICNIC analysis of Affymetrix SNP6.0 array data, were obtained from the Sanger's Institute COSMIC cell line webpage (see **URLs** and **Online Methods** for details).

We assessed the predictive performance of the cell line-specific peaBrain models using a Monte Carlo-validation scheme (**Figure 1,** *on next page*). In each fold, 5% of small molecules were randomly withheld as a test set (that is, neither the fingerprint nor expression profile is ever seen by the model). Random selection of the test set was sufficient as molecular similarity was very weakly correlated with transcriptomic similarity (median Spearman's rho = 1.0%). Of the remaining 95% of small molecules, 10% was randomly selected as validation to check for early exit criteria (i.e. not used to train the model

parameters, see **Online Methods**). The model was then trained on the remaining drug profiles to predict the gene-level expression, 24 hours post-exposure, for any pair of molecular fingerprints and core promoter sequences (per cell line). We note that, across all cell lines, the model had a median oos-r^2 of 84% (**Figure 1**). Importantly, this performance was 12% better (on average) than simply using the expression profile of the closest structurally similar molecule in the training set (e.g. using a structurally-similar oestradiol analogue as the prediction of the oestradiol expression profile). This suggests that the model is able to extract more biologically-relevant information (likely from the promoter sequences) than that encoded in the molecular fingerprint used as input to the algorithm.



Supplementary Note 3 – Figure 1. Given a core promoter sequence and a molecular fingerprint, the small molecule peaBrain model can accurately the transcriptomic perturbation of the corresponding gene across all 5 cell lines. Each point represents oos-r^2 of a single unique small molecule (included in the test set). We performed 4 folds of Monte Carlo validation per cell line (folds denoted by the colour). The y-axis was clipped at zero; only seven molecules across all cell lines and test/validation repeats had an oos-$r^2 < 0$. **Abbreviations:** OOS R2, out-of-sample r^2.

To further test the model generalizability, we assessed predictive performance on the rank-normalized LINCS phase II *GSE70138* (external) dataset. Moreover, to calibrate our assessment to the upper bound of assay reproducibility, for each test set, we sub-selected molecules assessed in both the phase I and phase II LINCS experiments. This way we are able to benchmark peaBrain predictive performance to experimental replicates (using the same assay and platform); that is, for each small molecule, we performed two tests: how well the imputed peaBrain profile compares to the phase II profile, and how well the phase I profile (i.e. the experimental replicate) compares to phase II profile. We note almost identical performance between peaBrain predictions (median rho between peaBrain and phase II profiles = 54%) and the phase I experimental replicates in predicting phase II profiles (median rho between phase I and phase II profiles = 47%), suggesting that *in silico* peaBrain predictions of molecule-induced transcriptomic perturbations are as accurate as *in vitro* experimental replicates.

peaBrain imputed molecule profiles significantly outperform chemoinformatic/structure-based approaches in predicting drug approval.

Having established the predictive capacity of the small molecule peaBrain models (on par with *in vitro* experimental replicates), we were interested in leveraging the peaBrain framework for drug repositioning. To this end, we imputed the expression profile for all clinically-relevant small molecules in the ChEMBL database[3] across all 5 cell lines (i.e. all molecules assigned at least to one indication). We divided all clinically-relevant molecules in two sets: molecules whose first approval occurred before the year 2000 (inclusive; n = 728 molecules), and all molecules that gained first approval after the year 2000 (n = 344 molecules). For indications relevant to each cell line (e.g. "breast neoplasms" for MC7 or "non-small cell lung carcinoma" for A549; **Table 2,** *on next page*), we identified the subset of molecules assigned to the corresponding indication as the positive class (background is all other clinically-relevant molecules). We then trained a random forest classifier on the pre-2000 molecule set using the imputed peaBrain molecule profiles. We sought to assess how well the imputed peaBrain expression profiles can identify clinically-relevant molecules for a given indication, from all other

molecules. To establish a baseline for comparison, we also constructed an equivalent classifier using only structural information (i.e. the molecule fingerprint used to impute expression). We assessed performance using four different metrics (useful metrics for classification with class imbalance):

(a) F1 score (harmonic mean of precision and recall);

(b) average precision (average of precisions achieved at each threshold weighted by the increase in recall from the previous threshold; corresponds roughly to the area under the precision-recall curve but is more conservative);

(c) precision score (positive predictive value; the fraction of true positives among all molecules predicted as positive by the classifier); and

(d) recall score (sensitivity; fraction of true positives over the number of all molecules relevant for the given indication).

All scores range from 0 to 1, with larger values indicating (elements of) better performance.

Across all five cell lines and the corresponding five indications, the imputed peaBrain expression profiles significantly outperformed chemoinformatic/structure-based approaches in predicting molecule clinical relevance for any given indication (**Table 2**), with 64-86% increase in F1-scores, 22-63% increase in precision, 75-94% increase in recall, and 13-19% increase in average precision. peaBrain models that use molecule expression profiles from all cell lines have markedly increased performance compared to peaBrain models that use expression profiles from a single cell line, which in turn outperform structure-based approaches. We note that this trend for peaBrain performance is consistent even with more restrictive definitions for the positive class (i.e. when limited to molecules are that are exclusively labelled as clinically-relevant for the corresponding cell line indication and after removing molecules generally indicated for “neoplasms”).

Supplementary Note 3 – Table 2. Tabulated classifier summary statistics on the post-2000 clinically-relevant molecule set (after training on the pre-2000 molecule set) for all five indications. For each indication, we constructed two peaBrain models: one using the expression profiles for the indication-relevant cell line and one using the expression profiles for all cell lines (denoted by "all"). The % increase columns highlight the increase in performance of the peaBrain classifiers compared to the baseline structure-based approach. The "all" peaBrain models significantly outperform all other approaches; the single cell line peaBrain models also have markedly increased performance compared to the baseline structure-based approach. We observe that this performance is consistent even with more stricter definitions for class indications (after removing molecules generally indicated for the MeSH term "neoplasms").

Melanoma + Neoplasms (A375)	structure-based approach	peaBrain (A375)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.12	0.14	**20.37**	0.27	**78.97**
Average Precision	0.37	0.42	**12.05**	0.42	**10.84**
Precision Score	0.40	0.70	**54.55**	0.52	**25.35**
Recall Score	0.07	0.08	**15.38**	0.18	**90.91**
Non-small cell lung carcinoma + Neoplasms (A549)	structure-based approach	peaBrain (A549)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.14	0.18	**26.93**	0.29	**71.79**
Average Precision	0.36	0.40	**9.99**	0.43	**15.52**
Precision Score	0.44	0.64	**38.02**	0.55	**22.49**
Recall Score	0.08	0.10	**25.00**	0.20	**83.33**
Colorectal neoplasms + Neoplasms (HT29)	structure-based approach	peaBrain (HT29)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.17	0.18	**6.90**	0.33	**65.19**
Average Precision	0.39	0.39	-0.99	0.46	**15.82**
Precision Score	0.50	0.45	**9.52**	0.63	**22.22**
Recall Score	0.10	0.11	**10.53**	0.22	**75.86**
Breast neoplasms + Neoplasms (MCF7)	structure-based approach	peaBrain (MCF7)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.14	0.33	**83.74**	0.35	**89.18**
Average Precision	0.38	0.37	-3.26	0.46	**18.87**
Precision Score	0.32	0.50	**43.90**	0.62	**64.06**
Recall Score	0.09	0.25	**96.77**	0.25	**96.77**
Prostatic neoplasms + Neoplasms (PC3)	structure-based approach	peaBrain (PC3)	**% increase**	peaBrain (all)	**% increase**
F1-score	0.18	0.25	**30.73**	0.30	**51.15**
Average Precision	0.43	0.39	-9.22	0.44	**2.22**
Precision Score	0.50	0.61	**19.61**	0.56	**11.11**
Recall Score	0.11	0.15	**33.33**	0.21	**62.07**

shRNA extension of the peaBrain DNA-reporter model captures ~85% of the variance in perturbed gene expression, allowing us to identify 292 new transcription factor-target interactions.

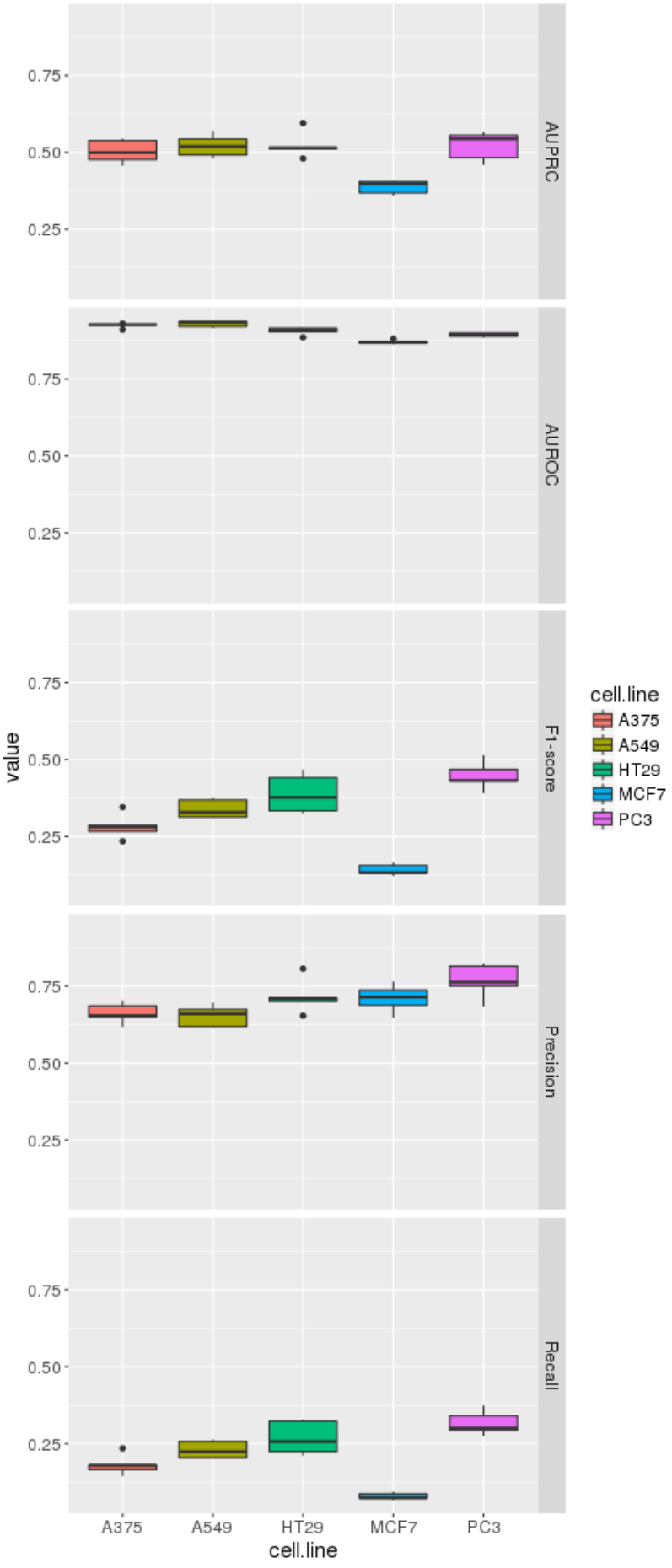
Using a nearly identical model and assessment procedure, we can also predict the transcriptomic consequences of shRNA. Instead of a molecular fingerprint, we one-hot encoded the shRNA oligo-sequence as input, alongside the promoter sequence:

promoter + shRNA → expression

To highlight the utility of the shRNA model, we sought to predict the transcriptomic consequences of 330,617 unique shRNA oligo-sequences (a subset of which targeted 19,992 human genes) to enable inference of the human regulatory network; the extensive number of conditions provides (multiple) snapshots of the transcriptome with nearly every gene perturbed. We used a meta-inference approach (aggregation of multiple inference approaches) for network reconstruction, limited to genes directly measured on the L1000 platform (978 genes; see **Online Methods**).

Using the HTRIdb[4] (an open-access database for experimentally verified human transcriptional regulation interactions) as the gold standard, we assessed our network inference for all 5 cell lines using a 5-fold Monte Carlo validation scheme with 10% of edges as test set (in every repeat). We noted the regulatory networks for each of the 5 cell lines captured experimentally-verified interactions at a precision ranging from 66-77%, with AUROC ranging from 87-93%, AUPRC from 39-52%, recall from 26-37% and F1-scores from 14-45% (**Figure 2**). For comparison, the "Dialogue on Reverse Engineering Assessment and Methods" (DREAM) project constructed high-confidence networks for *Eschierica coli* and *Saccharomyces cerevisiae* (simpler organisms) at a precision of ~50%[5]. The inferred networks, for all 5 cells, have been made available (see **URLs**).

Supplementary Note 3 – Figure 2 (*on next page*). Plots depicting the performance of network inference using the large-scale shRNA-perturbed expression matrix, across all 5 cell lines. Boxplot for each metric was generated using a 5-fold Monte Carlo validation scheme (with 10% of edges randomly selected as the test set). *Abbreviations:* AUROC, area under the receiver operating curve; AUPRC, area under the precision-recall curve.


AUPRC
AUROC
F1-score
Precision
Recall
value
0.75
0.50
0.25
A375
A549
HT29
MCF7
PC3
cell.line
cell.line
A375
A549
HT29
MCF7
PC3

We subsequently identified all edges inferred as interactions – but absent from the experimentally-verified HTRIdb curation – as new/candidate regulatory interactions (n = 212 total). Using ChIP-seq data from ENCODE[6], we verified that 83% of have experiment support; that is, the predicted transcription factor binds within 50kb of the transcription start site of the corresponding target. Of all 212 interactions, 40% bind within 2kb of the transcription start site. All new inferred interactions have been made publicly available (see **URLs**).

REFERENCES

1 Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742-754 (2010).

2 Subramanian, A. *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437-1452. e1417 (2017).

3 Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **40**, D1100-D1107 (2011).

4 Bovolenta, L. A., Acencio, M. L. & Lemke, N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC genomics* **13**, 405 (2012).

5 Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nature methods* **9**, 796 (2012).

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

SUPPLEMENTARY NOTE 4

Stage 2 peaBrain predictions were significantly and positively correlated with univariate eQTL coefficients. Having established the predictive ability of peaBrain, we were interested in whether we can use the best-performing peaBrain models to measure the impact of single variants, compared to DeepSEA logFC estimates and MPRA log skew estimates (**Task F**). For all "captured" genes (n=113), we selected all variants identified as significant eQTLs in the GTEx v6p univariate eQTL analysis (n=16,019 variants; **see Online Methods**) and replicated the analysis with the Geuvadis dataset[1] (n=17,279 variants for the EU population and n = 1601 variants for the YRI [Yoruba from Ibadan, Nigeria] population). Unlike GTEx v7, eQTLs from GTEx v6p were derived from genotyping arrays (Illumina OMNI 5M or 2.5M) and thus did not include WGS used to train the peaBrain model. The Geuvdais dataset (both EU and YRI populations) were eQTLs derived from WGS from a set of non-overlapping subjects. For each eQTL (including indels), we created pairs of 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. peaBrain predictions 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**). Results were consistent across all datasets when we relaxed our heritability to include genes with GCTA $p < 0.05$: Spearman's rho equal to 0.08 ($p = 3.02 \times 10^{-25}$), 0.07 ($p = 3.65 \times 10^{-25}$), and 0.18 ($p = 1.68 \times 10^{-13}$) for the GTEx, EU-Geuvadis, and YRI-Geuvdais univariately-significant eQTLs, respectively. Alongside this positive correlation, we observed that many variants with large coefficients across all three datasets had small peaBrain predictions (**Supplementary Figures 2-4**). This shrinkage in peaBrain estimates is consistent with the appreciable LD between eQTL-associated variants at any given locus, such that only a subset of significantly-associated variants are actually functional. Whilst univariate linear models are not capable of distinguishing between functional versus "hitchhiker" variants, the joint modelling of the input sequence in Stage 2 peaBrain allows a more direct assessment of the function of each variant. In

comparison, we noted that the MPRA log skew estimates did not correlate with the univariate eQTL coefficients in the GTEx (rho = 0.014; p = 0.60), the EU-Geuvadis (rho = -0.018; p = 0.58), or the YRI-Geuvadis (rho = 0.011; p = 0.82) eQTL analyses. The BiT-STARR-seq log skew estimates also did not correlate with the univariate eQTL coefficients in the GTEx (rho = -0.031; p = 0.50), the EU-Geuvadis (rho = 0.035; p = 0.44), or the YRI-Geuvadis (rho = -0.112; p = 0.32) eQTL analyses. Similarly, for GM12878 (LCL) annotations, we noted that the maximum log fold changes from DeepSEA did not correlate with the magnitude of the eQTL coefficients in the GTEx analysis (rho = -0.001; p = 0.92), the EU-Geuvadis analysis (rho = -0.010; p = 0.18), and the YRI-Geuvadis (rho = 0.016; p = 0.52). The DeepSEA results suggest that simply predicting chromatin effects and TFBS is not sufficient for predicting the transcriptomic consequences of sequence variation for this set of selected genes. This is consistent with experimental evidence suggesting that most genetic variants in DNase footprinted (and other annotated) regulatory regions are silent[2], and the fact that histone modifications (e.g. methylation) alone are frequently insufficient to transcriptionally perturb promoters[3]. However, it is important to note that DeepSEA was trained on the reference genome (i.e. not exposed to genotype variation), while Stage 2 peaBrain was trained to predict differences in expression as a function of differences in sequence between any pair of individuals.

REFERENCES

1 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).

2 Moyerbrailean, G. A. *et al.* Which genetics variants in DNase-Seq footprints are more likely to alter binding? *PLoS genetics* **12**, e1005875 (2016).

3 Ford, E. E. *et al.* Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. *bioRxiv*, 170506 (2017).