ARTICLES
https://doi.org/10.1038/s42256-020-0201-6
nature machine intelligence



# Deep learning decodes the principles of differential gene expression

Shinya Tasaki[1]✉, Chris Gaiteri[1], Sara Mostafavi[2] and Yanling Wang[1]

**Identifying the molecular mechanisms that control differential gene expression (DE) is a major goal of basic and disease biology. Here, we develop a systems biology model to predict DE and mine the biological basis of the factors that influence predicted gene expression to understand how it may be generated. This model, called DEcode, utilizes deep learning to predict DE based on genome-wide binding sites on RNAs and promoters. Ranking predictive factors from DEcode indicates that clinically relevant expression changes between thousands of individuals can be predicted mainly through the joint action of post-transcriptional RNA-binding factors. We also show the broad potential applications of DEcode to generate biological insights, by predicting DE between tissues, differential transcript usage, and drivers of ageing throughout the human lifespan, of gene co-expression relationships on a genome-wide scale, and of frequently differentially expressed genes across diverse conditions. DEcode is freely available to researchers to identify influential molecular mechanisms for any human expression data.**

Although all human cells share DNA sequences, gene regulation differs among cell types and developmental stages, and also in response to environmental cues and stimuli. Accordingly, when gene expression is not properly regulated, cellular homeostasis can be perturbed, often affecting cell function and leading to disease[1]. These distinctions between cell states are observed as differential expression (DE) of gene transcripts. DE has been catalogued for tens of thousands of gene expression datasets, in the context of distinctions between species, organs and conditions. Despite the important and pervasive nature of DE, it has been challenging to shift from these observations towards a coherent understanding of the underlying generative processes that would essentially decode DE—a transition that is essential for progress in basic and disease biology. We address this gap by exploiting novel computational and systems biology approaches to develop a predictive model of DE based on genome-wide regulatory interaction data.

Diverse molecular interactions have been shown to generate DE and jointly regulate gene expression at the transcriptional and post-transcriptional levels. Major classes of gene regulatory interactions have been catalogued at the genomic scale, including transcription factor (TF)–promoter interactions[2], protein–RNA interactions[3], RNA–RNA interactions[4], chromatin interactions[5] and epigenetic modifications on DNAs[6], histones and RNAs[7]. Statistical models of gene expression can help fulfil the purpose of these resources in describing the origins of gene regulation and DE[1]. However, such raw data resources have outpaced model development, probably because of the challenge of uniting diverse molecular data into a single accurate model.

Among the many possible statistical approaches to predicting DE, deep learning (DL) blends diverse data sources in a way that approximates the convergence of regulatory interactions. Indeed, DL has been applied to genomic research[8,9], including RNA splicing[10], genomic variant functions[11] and RNA/DNA binding[12]. However, accurate prediction is only one component of understanding DE, and additional genomic and systems biology analyses are helpful in understanding how predictions are fuelled by existing molecular concepts, mechanisms and classes.

To decode the basis of DE in terms of molecular regulatory interactions, we first learn to predict it with a high degree of accuracy, using a DL model we call 'DEcode' (www.differentialexpression.org). This model combines several types of gene regulatory interaction and allows us to prioritize the main systems and molecules that influence DE on a tissue-specific basis. We further establish likely molecular mechanisms for this gene regulation and validate the influence of the predicted strongest regulators. In parallel, we predict the origin of person-to-person DE, which is the major component of experimental and clinical studies. These particularly challenging predictions are validated on a genome-wide scale, as we identify key drivers of co-expression and also drivers for phenotype-associated DE. Utilizing diverse genomic datasets, we identify a complex, yet strikingly consistent set of principles that control DE. DEcode can be applied to the majority of current and future gene expression data to accelerate basic and disease biology by identifying the origins of DE in each experiment.

## Results

**Predicting DE across human tissues.** The overarching goal of this study is to understand the major regulatory principles of DE signature by predicting gene expression as a function of molecular interactions. These results should be tissue-specific and ideally have sufficient accuracy to predict the relatively small expression changes observed between individual humans. To accomplish this, we utilized deep convolutional neural networks in a system called DEcode that can predict inter-tissue variations and inter-person variations in gene expression from promoter and mRNA features (Fig. 1). The promoter features include the genomic locations of experimentally determined binding sites of 762 TFs[13] and the mRNA features encompass the locations of experimentally determined binding sites of 171 RNA-binding proteins (RBPs)[14] and the predicted binding sites of 213 miRNAs[15] (Supplementary Table 1). We first applied the DEcode framework to tissue-specific human transcriptomes of 27,428 genes and 79,647 transcripts measured in the GTEx consortium[16] to predict log-fold-changes across 53 tissues against the median expression of all tissues, as well as the median expression

[1]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. [2]University of British Columbia, Vancouver, BC, Canada.
✉e-mail: stasaki@gmail.com

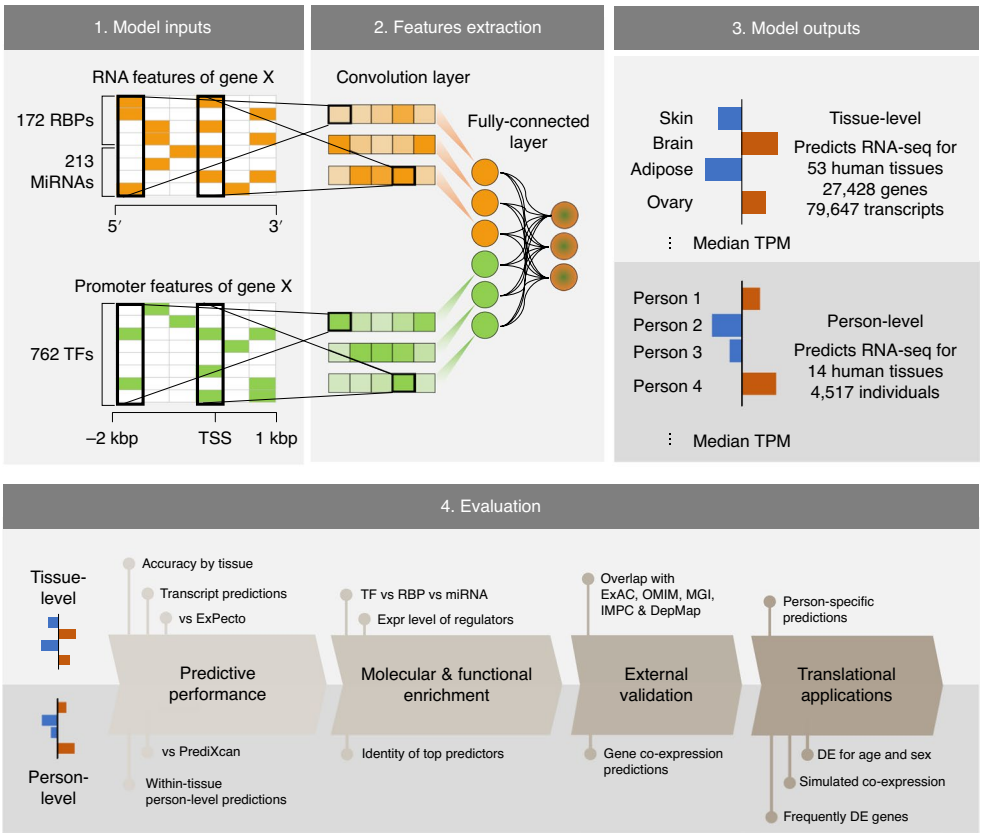



**Fig. 1 | Overview of building and evaluating the DEcode transcriptome prediction model.** DEcode takes the promoter features and the mRNA features for each gene/transcript as inputs, then outputs its expression levels under various conditions. We conducted a series of evaluations to show that DEcode defines major principles in gene regulation in arbitrary gene expression data. This capacity is strongly supported on a comparative basis to alternative methods, and on an absolute basis across diverse applications, including through predictions of transcript usage, person-specific gene expression and frequently DE genes of multiple external disease-related gene sets.

of all tissues with a multi-task learning architecture. The predicted median expression levels show high consistency with the actual observations at both gene level (Spearman's $\rho = 0.81$) and transcript level (Spearman's $\rho = 0.62$) (Fig. 2a). Moreover, the model predicted differential transcript usage within the same gene (Spearman's $\rho = 0.44$). The DEcode models predicted the DE profiles across 53 tissues for gene (mean Spearman's $\rho = 0.34$), transcript (mean Spearman's $\rho = 0.32$) and transcript usage levels (mean Spearman's $\rho = 0.16$) (Fig. 2b). The predicted gene expression for the testing genes was indeed tissue-specific, showing less correspondence with the expression profiles from other tissues (Fig. 2c).

To provide context for the statistical benefit of using systems biology data in our method for predicting DE, we contrast it to a raw-sequence-based method called ExPecto[11], as it was designed to predict GTEx gene expression from epigenetic states, estimated from promoter sequences via DL. In this comparison, DEcode showed an average of 6.1% improvement in root-mean-square error (r.m.s.e.) over ExPecto (Fig. 2d), which translates into an average correlation coefficient with actual gene expression of 0.42, a 50% increase over the 0.28 from ExPecto (Supplementary Fig. 1).

**Regulators for DE across tissues.** Beyond the predictive performance of DEcode, we utilize the model to help define the biological processes regulating DE. Many studies have demonstrated that TF–promoter interactions are critical determinants of the transcriptional activity of promoters and thereby define gene expression levels[2]. However, it is unclear to what extent RNA features (which we define as each RNA's binding sites for RBPs and miRNAs) contribute to gene expression levels. To answer this question, we retrained the DEcode model, randomizing either RNA features, promoter features or both. We found that RNA features alone explained the gene-level log-fold-changes less accurately than the model with all features, but as well as the model trained with promoter features (Fig. 3a). Interestingly, the RNA-based model performed better than the promoter-based model for the prediction of transcript expression and differential transcript usage.

To further quantify the importance of each regulator weighted in the DEcode models, we calculated DeepLIFT scores, which are a measure of the additive contribution of its binding site to each prediction[17,18], for each task of predicting fold-change (Supplementary Table 2). Based on the DeepLIFT scores, TFs and RBPs showed comparable effects on the gene-level prediction (Fig. 3b). However, at the transcript and transcript usage level, RBPs tended to influence the prediction more than TFs. We also found biological processes enriched for the influential predictors ranked by the DeepLIFT scores that are coherent with known gene regulation (Fig. 3c and Supplementary Table 3). For example, the predictors for gene and transcript expression were associated with histone modification, whereas splicing-related factors were only enriched for predictors of transcript and transcript usage.

We also examined critical predictors for each of 53 tissues (Fig. 3d). The DeepLIFT scores across tissues recapitulated the contribution of binding sites of known master regulators in each tissue such as REST for brain tissues[19], SPI1 and RUNX1 for

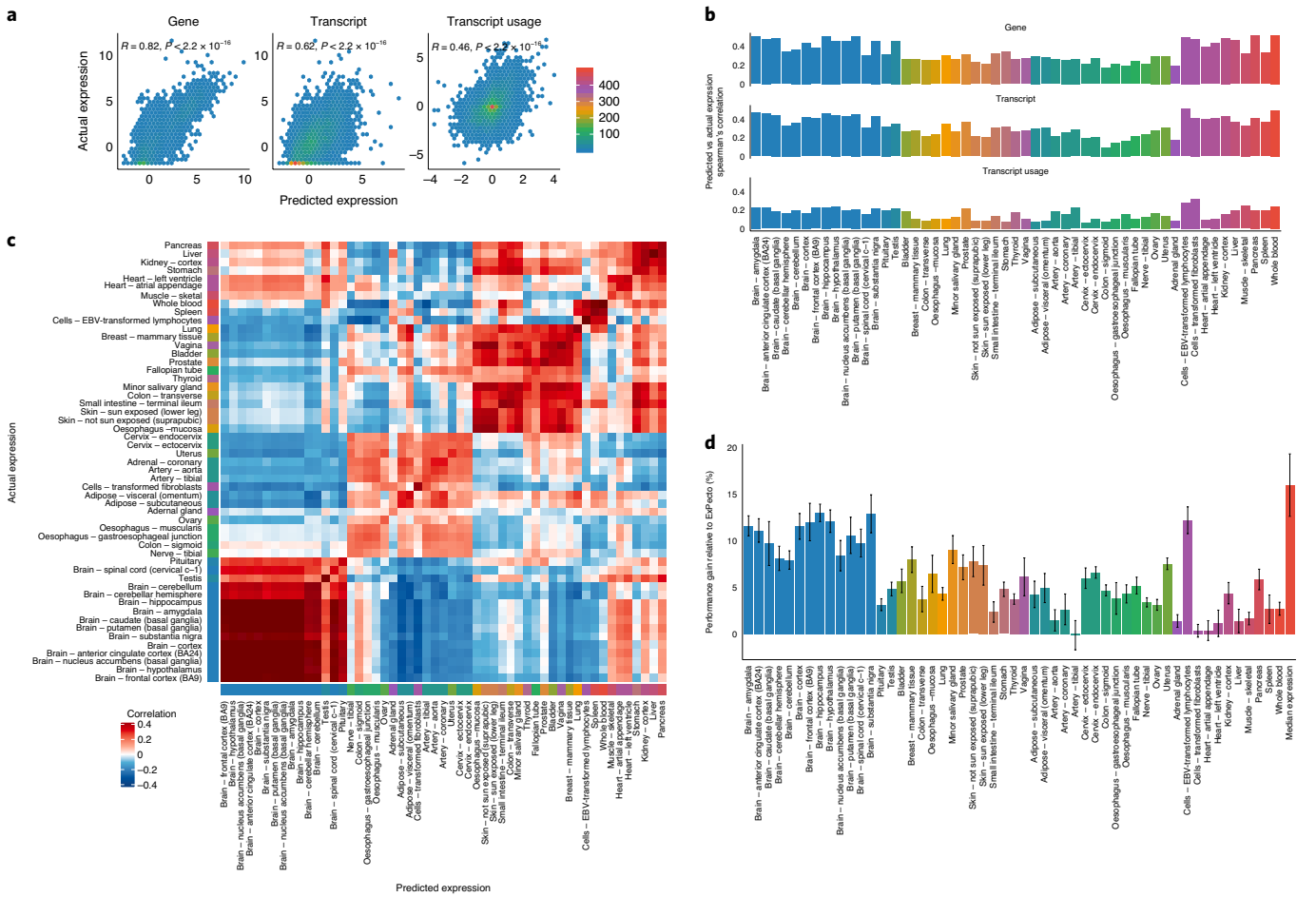



**Fig. 2 | Performance of DEcode for predicting DE across 53 tissues. a**, Predictive performances on the median absolute expression levels across tissues. The predicted $\log_2$(TPM) (TPM, transcripts per million) values for 2,705 genes and 7,631 transcripts, and the transcript usage for 1,485 genes that had multiple transcripts, were compared with the actual expression using Spearman's rank correlation. **b**, Predictive performances on the tissue-specific expression profiles. The predicted fold-changes of 2,705 genes and 7,631 transcripts were compared with the actual data using Spearman's rank correlation. Predictive accuracies of the differences in the transcript usage across tissues were also computed for 1,485 genes. The colours of the bars indicate the tissue groups based on the similarity of gene expression profiles. **c**, Heatmap showing pairwise correlations between the predicted and actual tissue-specific expression profiles of 53 tissues for the testing genes. **d**, Comparison of the performances of DEcode and ExPecto. The root-mean-square errors (r.m.s.e.) of DEcode models for the expression levels of 714 genes coded on chromosome 8 were compared with those of ExPecto. The relative improvement of DEcode over ExPecto in the median r.m.s.e. of the 10 runs is displayed as a bar plot. Error bars represent median absolute deviation.

immune-related tissues[20], TP63 and KLF4 for skin[21], HNF4A for liver[22] and PPARG for adipose-related tissues[23], which suggested that the differences in predictive contributions of binding sites of a given regulator reflect the differential activities of regulators across tissues. We hypothesized that the differential activities of a regulator could, in part, be explained by the relative abundance of a regulator across tissues. Based on this hypothesis, we contrasted DeepLIFT scores for the binding sites of each regulatory factor and its expression levels across tissues. We indeed found that 99 RBPs and 410 TFs showed significant correlations between the DeepLIFT scores of their binding sites and their expression levels (FDR < 5%) (Supplementary Fig. 2). These relationships were not based on differences in expression profiles between brain and non-brain tissues, as the relationships remained the same without brain tissues (Supplementary Fig. 3). The sign of the correlation possibly reflects whether the binding of a regulator to RNA increased or decreased the abundance of the RNA. For example, the model suggested that PPARG and PTBP1 are positive regulators of gene expression as DeepLIFT scores of PPARG or PTBP1 binding sites were higher in the tissues expressing PPARG or PTBP1 at higher levels (Fig. 3e), which are consistent with their known functions[23,24]. Conversely, the expression levels of REST, a transcriptional repressor[19], and METTL14, an RNA methyltransferase destabilizing RNAs[25], showed inverse correlations with their DeepLIFT scores, as expected. These results indicate that the DEcode model is interpretable and reflects biological mechanisms for controlling RNA abundance.

We hypothesized that if key predictors in DEcode models are truly impactful transcriptome regulators, then defects in such regulators would have significant impacts on cellular phenotypes and thereby lead to disease. To examine this hypothesis, we first obtained genes whose loss-of-function (LoF) mutations are depleted through the process of natural selection (from the Exome Aggregation Consortium, ExAC)[26] and thus that are considered to play important roles in individual fitness. We found that these LoF mutation-intolerant regulators had greater DeepLIFT score magnitudes for the prediction of the fold-change across 53 tissues (Fig. 3f and Supplementary Table 4). In particular, these associations are based on genes that are intolerant to LoF mutations only in a single allele (Supplementary Fig. 4). We also confirmed that mutations in the predicted key regulators tend to cause genetic disorders based on assessment with the Online Mendelian Inheritance in Man (OMIM) compendium. Interestingly, their roles on fitness are

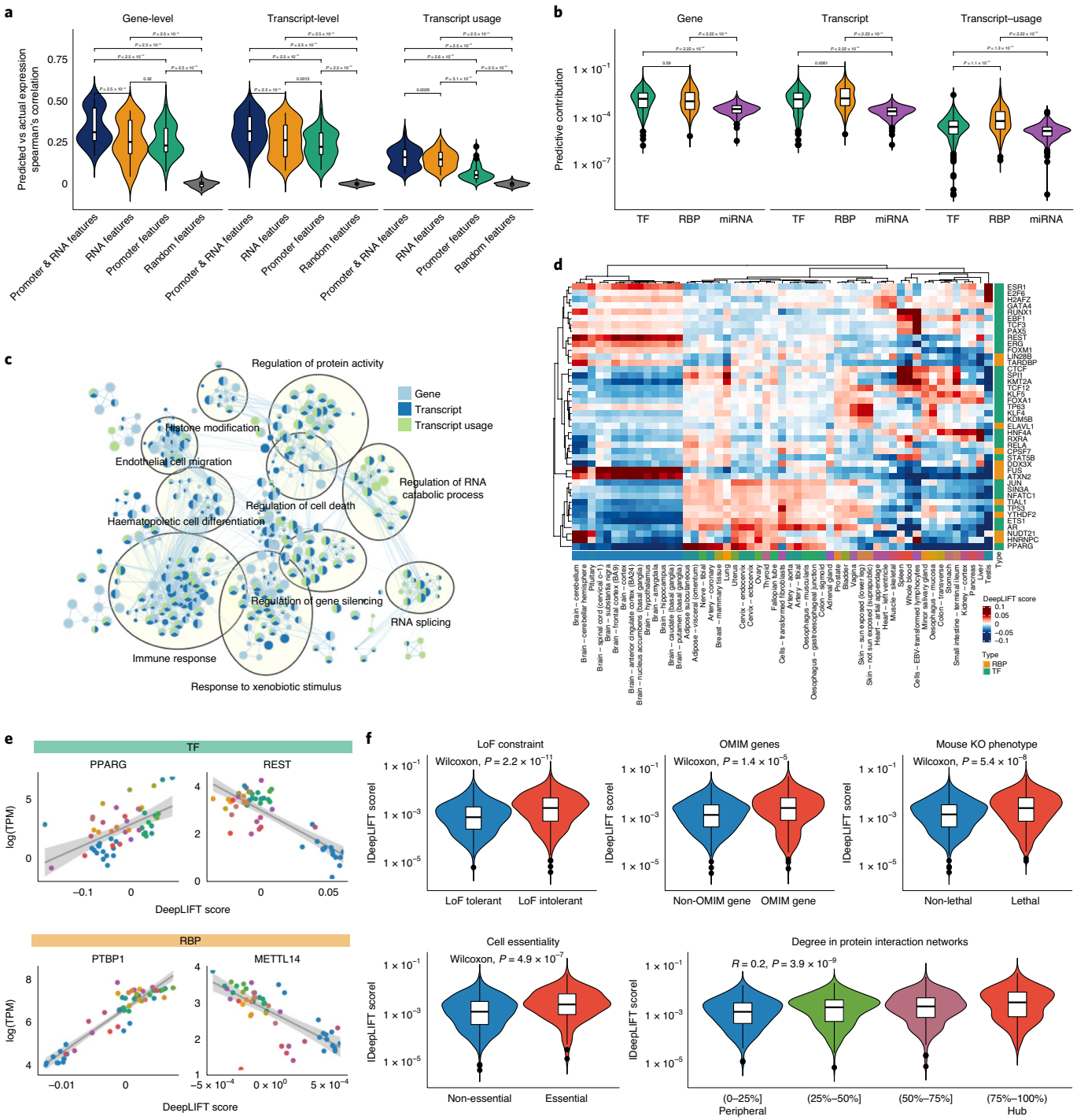



**Fig. 3 | Identification and characterization of key predictors. a**, The predictive performances of the models trained with a different set of features. The performances for the fold-change across 53 tissues were compared with the paired-sample *t*-test. **b**, Comparison of predictive contribution among classes of regulators. The regulator's predictive contribution to the fold-change across 53 tissues was estimated based on the DeepLIFT score. The two-sample Wilcoxon test was used to assess the statistical significance of the differential contribution. **c**, Gene ontology (GO) enrichment map for key predictors. GO enrichment analysis for key predictors for the fold-change across 53 tissues was conducted using the pre-ranked gene set enrichment analysis (GSEA). The significant associations (false discovery rate (FDR) < 0.05) are visualized with green/blue node colours representing certain classes of regulators. **d**, Key predictors for the tissue-specific transcriptomes. DeepLIFT scores of the top five key predictors in each tissue are displayed as a heatmap. **e**, Example relationships between the predictive contributions of a regulator and its expression levels across tissues. A line of best fit based on linear regression is shown, with 95% confidence intervals (shading). **f**, The overlap between the key predictors for the fold-change across 53 tissues and loss-of-function (LoF) intolerant genes, Online Mendelian Inheritance in Man (OMIM) genes, lethal genes in mice, essential genes in cell lines and hub proteins in protein–protein interaction networks.

probably preserved across species, as dysfunctions of the predicted key regulators also lead to pre-weaning lethality in mice. Moreover, we found that the LoF of the predicted key regulators of the transcriptome could also impair cellular viability. Intriguingly, the key predictors are more likely to be hubs in protein–protein interaction networks, which is the same property observed with

disease-causing genes[27]. These results were robust, as they were also supported by DeepLIFT scores for the transcript and transcript usage prediction (Supplementary Fig. 5). Together, the results indicate that the key predictors of the transcriptome indeed play critical roles in maintaining vital cellular and body functions. Thus the DEcode model could be used to prioritize disease-causing genes, and this capability points toward the broader validity of predicted key regulators.

**Predicting DE across individuals.** Next, we asked whether the same input of promoter and RNA features could also predict relative expression differences across individuals within the same tissue. We extended the DEcode framework to model DE across individuals for 14 representative tissues with a sample size greater than 100 in GTEx. This was challenging, as the average variance in gene expression within tissues was less than 25% of that between tissues (Supplementary Fig. 6).

The person-specific models successfully predicted fold-changes across individuals with a mean Spearman's correlation of ~0.28 (Fig. 4a). The performance was further increased to 0.34 when we filtered out the models that worked poorly for the validation data (Supplementary Fig. 7). The models were indeed person-specific, as they did not predict the gene expression profiles of unrelated individuals. To examine if the model captured the person-specific expression shared across tissues[28], we compared gene expression between tissues within the same individuals and between different individuals. The predicted expression showed better concordance between tissues from the same individuals, as is the case with actual expression data, indicating that the model has captured the person-specific regulatory mechanisms, even though we did not use any direct information that could identify individuals (Fig. 4b).

Next, to gauge the contributions of RNA and promoter features to the person-specific expression profiles, we retrained the models with randomized RNA features, promoter features or both. The RNA-feature-based model performed, on average, 85% as well as the model trained with all features. This corresponds to an average 173% performance gain compared to the promoter-feature-based model, which suggested that the post-transcriptional controls are the major determinants of the DE across individuals (Fig. 4c). The model also allowed us to investigate the person-specific activities of regulators by calculating DeepLIFT scores (Fig. 4d). At least 100 among 933 regulators in each tissue showed a significant correlation between their DeepLIFT scores and expression levels across individuals (Extended Data Fig. 1a). The signs of these correlations were consistent between tissues and consistent with those of the cross-tissue model (Extended Data Fig. 1b,c). This suggested that DE between individuals and between tissues can be modelled by the universal relationships between regulators and their targets.

To examine whether specific genes contributed to the per-person accuracy of the predicted gene expression, we also assessed its accuracy on a per-gene basis. The predicted expression of a majority of the testing genes (78% on average) showed significant positive correlations with the actual gene expression (FDR < 5%). To assess whether systems biology data improved per-gene performance over a raw-sequence-based method, we compared DEcode with PrediXcan[29], which predicts person-specific gene expression from genetic variations in *cis*-regulatory regions of genes. The PrediXcan model predicted the gene expression levels of only ~11% of the testing genes at FDR < 5%, which was far less than that of DEcode (Fig. 4e). This suggested that the differential activity of transcriptional and post-transcriptional regulators has a larger effect on gene expression than genetic variations in *cis*-regulatory regions.

The genes that DEcode could predict with the greatest accuracy were similar across tissues (Extended Data Fig. 1d). This suggested that the predictability of gene expression is defined by gene characteristics rather than a target tissue. We therefore explored gene characteristics that were associated with the per-gene accuracy of the predicted expression. We found that the models showed higher performance for the genes that are registered in multiple gene annotation databases than those found only in the GENCODE database (Extended Data Fig. 1e). Because the GENCODE-specific genes are novel or putative and thus their annotations might not be well established, it is reasonable that the performance of the model for those genes was lower than other well-established genes. Beyond the annotation reliability, we found that the number of known binding features for each gene had a larger effect on the predictability (Extended Data Fig. 1f). This suggested that the prediction becomes more accurate as more information on RNA and promoter interactions becomes available. Interestingly, the number of binding features in RNAs was a stronger determinant of the predictive accuracy than that in promoter regions (Fig. 4f). RNA–protein interactions are largely missing, as global RNA-binding profiles are available for only ~10% of known RBPs[30]. Thus, the incompleteness of RNA features is likely to be an origin of the lower accuracy for a portion of genes.

**Generative process of trait-related expression changes.** Next, we asked whether the person-specific expression profiles predicted by the DEcode models also retained trait-associated DE changes. For this, we conducted DE analysis against the donor's age and sex using the predicted gene expression data. Notably, the test statistics of the predicted data showed significant positive correlations with those of the actual data in all tissues for both traits (Fig. 5a). In particular, age- and sex-specific expression changes were well preserved in the predicted data in lung (Spearman's $\rho = 0.59$, $P < 2.2 \times 10^{-16}$) and hippocampus (Spearman's $\rho = 0.47$, $P < 2.2 \times 10^{-16}$), respectively. The predicted associations were tissue-specific, as they were the closest to those of corresponding tissues in 9 and 11 out of 14 tissues for age and sex, respectively (Fig. 5b). We also explored the regulators for the age- and sex-related gene expression changes by associating regulators' DeepLIFT scores with age and sex. We found that many regulators, for example, 717 in the tibial artery and 904 in the breast mammary tissue, showed age- and sex-dependent changes at FDR 5%, respectively (Fig. 5c and Supplementary Table 5), which showed the capability of DEcode to associate transcriptional regulators with phenotypes. Although there were more TFs associated with phenotypes than RBPs and miRNAs, overall, the collective impacts of RNA features on the generative process of DEs for age and sex were greater than those of promoter features in most tissues (Supplementary Fig. 8).

**Regulatory basis of gene co-expression relationships.** Co-expression analysis is a frequent component of transcriptome studies, as gene-to-gene co-expression relationships are regarded as functional units of the transcriptional system[31]. We thus examined if the DEcode models could detect known gene co-expression relationships. These tests are both a potential validation of the person-specific DEcode predictions and a means to explore the biological basis of co-expression. We found that the gene co-expression relationships in the predicted gene expression profiles separated gene pairs with positive and negative correlations in the actual gene expression data in each tissue (Fig. 6a). Furthermore, the predicted gene expression profiles also detected inter-tissue co-expression relationships (Fig. 6b). The accuracy of these results motivated us to investigate key factors driving co-expression, via the DEcode predictions. RNA features alone could explain co-expression relationships better than promoter features in most tissues (Extended Data Fig. 2), which again suggested the significant contribution of RNA features to person-specific transcriptomes.

To further assess the capability of DEcode to decipher the mechanisms leading a specific co-expression relationship, we focused on the co-expression of *LAPTM5* and *CD53*, which were robustly

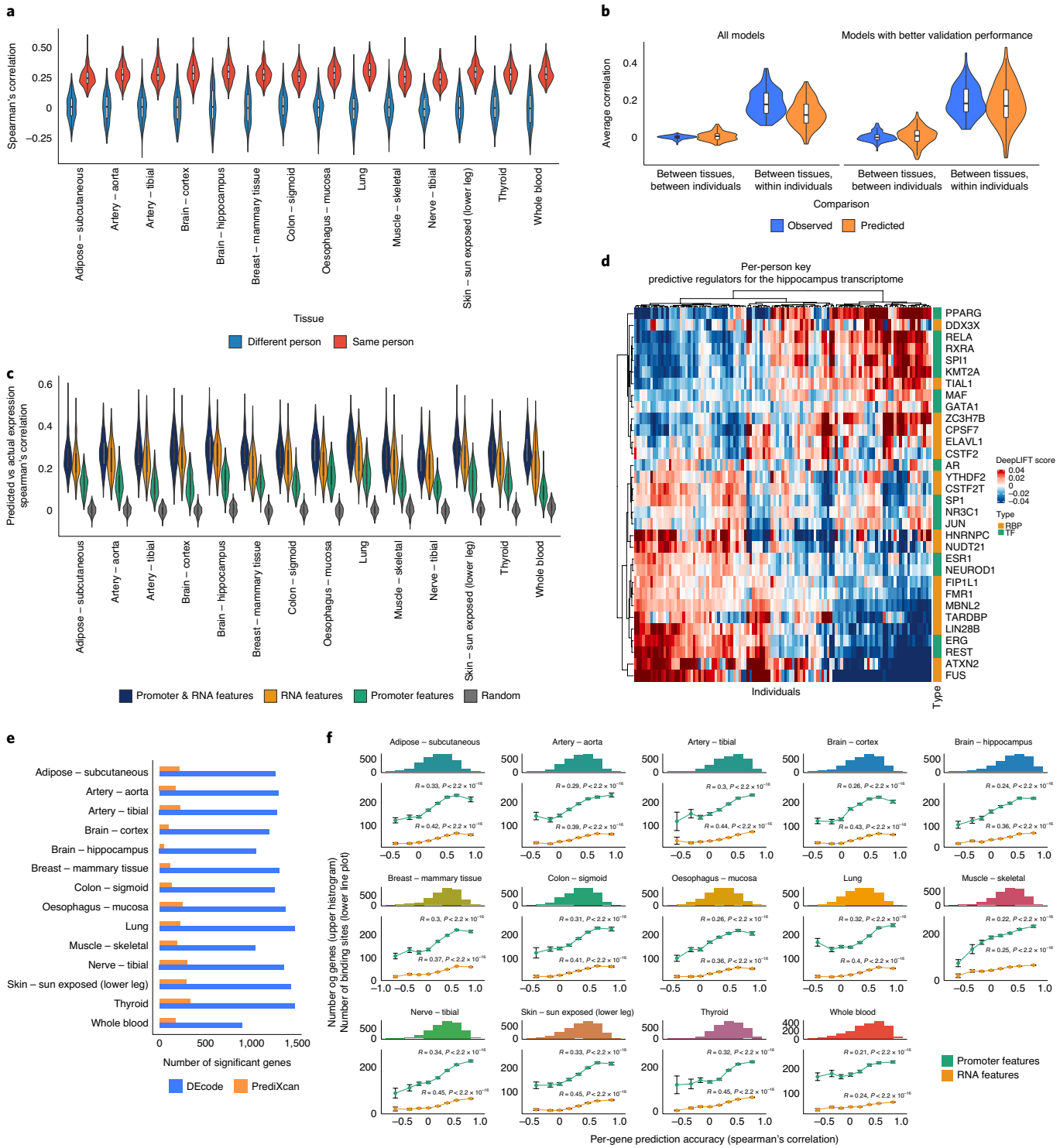



**Fig. 4 | Performance of the person-specific models. a**, Predictive performances of the person-specific models for the actual data from the same individuals and unrelated random individuals. **b**, The person-specific models predicted person-specific expressions shared across tissues. **c**, Performances of the models trained with a distinct feature set. **d**, Per-person key predictive regulators for the hippocampus transcriptome. We selected the top five key predictors of the hippocampus transcriptome for each individual. Their DeepLIFT scores are displayed as a heatmap. **e**, Comparison of per-gene predictive accuracy between DEcode and PrediXcan. The number of genes that showed a positive Pearson's correlation between predicted and actual gene expression levels at FDR of 5% was calculated for each method. The testing genes on chromosome 1 were used in this comparison. **f**, The per-gene prediction accuracy is associated with the number of features present in RNAs and promoters. The histograms show Pearson's correlations between the predicted and actual expression for each gene. The line plots show the average numbers of RNA and promoter features of genes in each bin of the histogram. Spearman's correlation values between the number of features and per-gene correlations are displayed in the line plots. Error bars indicate standard errors.

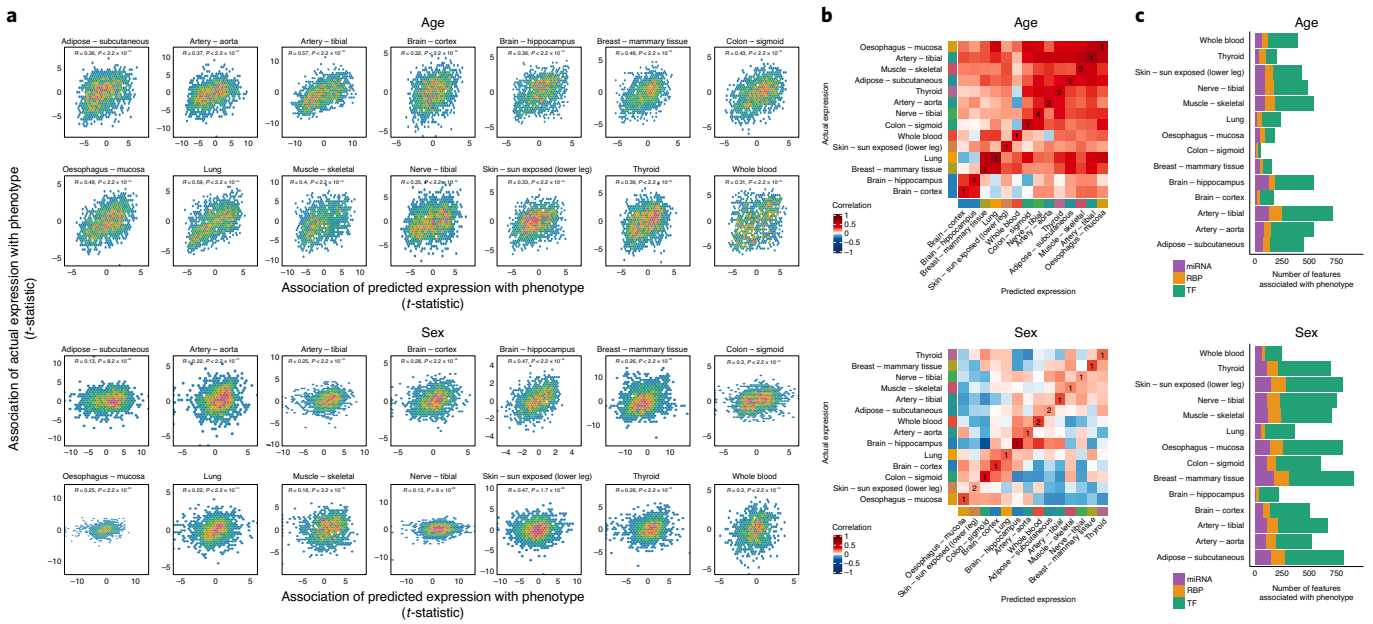



**Fig. 5 | Application of the person-specific models to analyse phenotype-related gene signatures. a**, Scatter plots showing the relations between the associations of genes with age and sex using predicted expression and those using the actual expression. Spearman's correlations between *t*-statistics based on the predicted and actual gene expression are displayed in the scatter plots. **b**, Pairwise Spearman's correlations between the predicted and actual associations of genes with age and sex in all tissues. The numbers in the diagonal elements of the heatmap indicate the ranks of similarity of the predictions with the actual observations in the corresponding tissues. **c**, The regulators whose DeepLIFT scores are associated with age or sex. The Benjamini–Hochberg procedure was used to control the FDR at 5% for each phenotype.

co-expressed both in the simulated expression data and the actual data in all tissues except whole blood. Using the trained model, we simulated the consequences of disruptions of promoter and mRNA features. The co-expression relationship was weakened when the features near the transcriptional start site (TSS) and 1,000 bp downstream of the TSS in *LAPTM5* or near the TSS and 500 bp upstream of the TSS in *CD53* were removed (Fig. 6c). These observed effects were reasonable, as many TFs bind to these regions. We further examined the specific regulators for the co-expression relationships by simulating knockout (KO) effects of regulators. The in silico KO experiments revealed that immune-related TFs such as SPI1 and TBX21 potentiated the co-expression relationships consistently across multiple tissues (Fig. 6d). To validate if these regulators indeed induced the co-expression relationships, we conducted a mediation analysis, that is, an orthogonal computational method to infer the effect of regulators on downstream targets. The set of 10 regulators together mediated up to 94% of covariance, which was significantly greater than the same number of randomly picked regulators (Fig. 6e).

**Molecular regulations for frequently DE genes.** A recent meta-analysis of over 600 human transcriptome data revealed that some genes are more likely to be detected as DE genes than others in diverse case–control studies[32]. From this observation, Crow et al. formulated the 'DE prior', a global ranking of a gene's generic likelihood of being DE. The genes with a high DE prior rank were significantly more enriched with frequently DE genes, as compared to other functional gene sets, such as those contained in GO. However, the regulatory origin behind the ranking of these highly responsive genes has yet to be uncovered. Therefore, we used DEcode to examine whether the DE prior rank could be generated by gene regulatory interactions, and to identify critical regulators for frequently DE genes. The ability of DEcode to predict global DE prior ranks was highly significant ($P < 2.2 \times 10^{-16}$) and practically relevant (Spearman's $\rho = 0.53$) (Extended Data Fig. 3a). Furthermore, DEcode was able to identify genes with high (90th percentile and greater) DE prior probability (area under the receiver operating characteristic curve = 0.81, 95% confidence interval = 0.78–0.84) (Extended Data Fig. 3b). Re-training the model with randomized inputs indicated that TF–promoter interactions were the major factors explaining the DE prior rank. To further characterize TFs that contributed to the prediction, we defined critical TFs based on the DeepLIFT score (Supplementary Table 6) and performed pathway analysis on them. We found that critical TFs were enriched for cancer or inflammatory-related pathways (FDR < 5%) such as pathways in cancer (fold = 3.1, $P = 4.2 \times 10^{-5}$), the JAK-STAT signalling pathway (fold = 6.8, $P = 4.8 \times 10^{-5}$), the chemokine signalling pathway (fold = 7.3, $P = 1.4 \times 10^{-4}$), and acute myeloid leukaemia (fold = 4.5, $P = 3.6 \times 10^{-4}$) (Supplementary Table 7). This result is consistent with the disease-related data context for the DE prior, which is 62% cancer-related and 23% inflammatory-related. Supported by the ability to predict DE prior ranks, and by the consistency of these results, this application of DEcode illustrates how it goes beyond DE gene lists to uncover the major key drivers for generating DE.

## Discussion

We have introduced the DEcode framework, which integrates a wealth of genomic data into a unified computational model of transcriptome regulations, to predict multiple transcriptional effects in tissue- and person-specific transcriptomes. Systems biology analysis of these results has provided biological insights regarding the regulatory mechanisms of the transcriptome. For example, it suggested that significant contributions of post-transcriptional regulators for explaining tissue-specific DE, differential transcript usage and even individual differences in transcriptomes and thus post-transcriptional controls might be key mechanisms with which to fine-tune the transcriptome in response to environmental and genetic factors.

We designed the DEcode framework as multi-task learning that predicts the transcriptomes of multiple samples simultaneously,

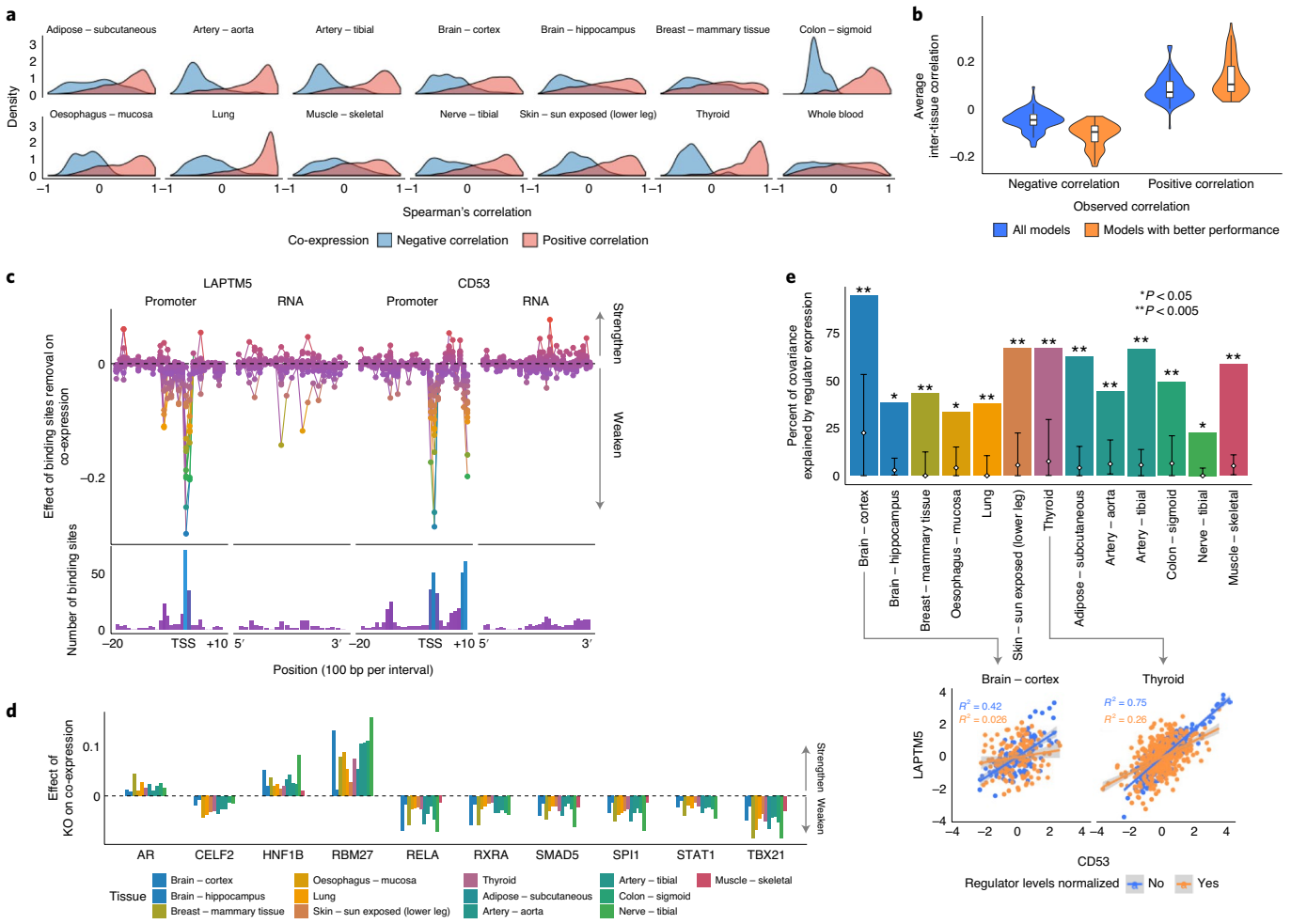



**Fig. 6 | Regulatory basis of gene co-expression relationships. a**, Co-expression relationships in the predicted gene expression. We defined the ground-truth co-expression relationships as gene pairs with absolute Spearman's correlation greater than 0.7 in the actual testing data. The density of Spearman's correlation between the co-expressed gene pairs in the predicted gene expression data was estimated based on the Gaussian kernel. **b**, Inter-tissue gene co-expression relationships in the predicted gene expression. We defined the ground-truth co-expression relationships as gene pairs with absolute Spearman's correlation greater than 0.5 in the actual test data. The violin plots show Spearman's correlation of the inter-tissue co-expressed gene pairs based on the predictions from all models or the models whose performances on validation data were greater than the 50% percentile in each tissue. **c**, The effect of binding site removal on co-expression between *LAPTM5* and *CD53*. **d**, The key regulators for co-expression between *LAPTM5* and *CD53*. **e**, Percent of the co-expression relationship explained by the expression levels of the key regulators. The white diamonds and error bars in the bar indicate the averages and 95% percentiles of the percent of variance explained by randomly picked regulators, respectively. The scatter plots and linear regression lines show the effect of the key regulators on co-expression.

with shared feature extraction layers, as opposed to single-task learning, which builds a prediction model specific to each sample. The multi-task design not only reduces the learning time, as feature extraction layers are optimized for all samples jointly and training data are generated only once, but also improves the prediction accuracy for both tissue-specific expression and person-specific expression (Supplementary Fig. 9). This improvement probably arises because multi-task learning is able to extract RNA and promoter features that are generalizable across samples, preventing overfitting to a specific sample. In addition, the multi-task design benefits when comparing differential contributions of key predictive features between samples, as it shares the same feature extraction processes, which is a critical step to achieving a biologically meaningful interpretation of the DEcode model.

Transcriptome analysis often identifies DE genes and then assesses the enrichment of functional genes such as TF targets, one by one. The person-specific DEcode model offers several comparative advantages over this traditional approach. First, DEcode can take into account the effects of multiple regulators simultaneously, rather than one at a time. Second, DEcode can estimate person-specific regulators' activities, which can be used to identify regulators associated with a phenotype of interest. Third, DEcode can simulate the consequence of KO perturbations for each gene. This step can reduce the number of candidate key drivers of gene expression changes by an order of magnitude or more, and facilitates the design of follow-up experiments. Indeed, key predictors nominated based on DEcode show significant and robust overlaps with various disease-causing and -related gene sets. Therefore, DEcode can extract more actionable information from transcriptome data, which will benefit a variety of transcriptome studies.

Looking toward even more expansive applications, the DEcode framework has the flexibility to incorporate other types of genomic information such as raw sequences, DNA methylation, histone marks and RNA modifications, and can also be extended to other organisms. Thus, the DEcode framework provides a direct bridge between accumulating genomic big data and individual transcriptome studies, allowing researchers to predict molecules that control DE associated with any condition or disease.

## Methods

**Transcriptome data processing.** To prepare the gene expression data used for model training, we downloaded the median gene TPM (transcripts per million) values from 53 human tissues from the v7 release of GTEx portal (https://gtexportal.org). We kept 27,428 genes that expressed more than two TPM in at least one tissue and $\log_2$-transformed the TPM with the addition of 0.25 to avoid a negative infinity. We then calculated the median $\log_2$(TPM) across 53 tissues and $\log_2$(fold-change) relative to the median of all tissues. The processed gene-level expression data comprised 27,428 genes with 54 columns, including relative fold-changes for 53 tissues and the median $\log_2$(TPM) across 53 tissues. To compile transcript-level data, we downloaded the individual-level transcript TPM from the GTEx portal and computed the median transcript TPM by tissue. We processed the transcript data in the same way as for the gene-level data. The resulted transcript-level data included 79,647 transcripts that corresponded to 23,813 genes. To build person-specific DEcode models, we obtained the gene-level TPM for each individual in 14 tissues from the GTEx portal. We filtered out low-expressed genes in each tissue and kept genes that expressed more than one TPM in at least 50% of samples. Then, we $\log_2$-transformed TPM with the addition of 0.25 and then quantile-normalized the $\log_2$(TPM). Finally, we removed the effects of technical covariates, including rRNA rate, intronic rate and RIN (RNA integrity number) via linear regression for each gene, followed by quantile normalization.

**Promoter and RNA binding features.** To generate RNA and DNA feature matrices, we downloaded the genomic locations of the binding sites of 171 RBPs from POSTAR2[14] as of October 2018, 218 miRNAs from TargetScan Release 7.2[15] and 826 TFs from GTRD[13] as of October 2018. We then mapped the binding sites of RBPs, miRNAs, and TFs to promoters and exons defined in the GTF file provided by the GTEx portal. A promoter region of each gene was defined as the region from 2,000 bp upstream of the TSS to 1,000 bp downstream of the TSS. We only used interactors that bind to promoters or RNA-coding regions of at least 30 genes, or transcripts as the predictors in each model. To reduce the size of the input, an RNA-coding region and a promoter region of each gene were binned with 100 bp intervals and the number of bases bound to each RBP, miRNA or TF was counted in each interval. This step generated RNA and DNA feature matrices for each gene described in Fig. 1.

**Training tissue-specific models.** To train the gene-level model of tissue-specific expression, we reserved all 2,705 genes coded on chromosome 1 as the testing data and the rest of the genes were randomly split into training data (22,251 genes) and validation data (2,472 genes). In the case of the transcript model, we used all 7,631 transcripts coded on chromosome 1 as the testing data; the rest of the transcripts were randomly split into training data (64,978 transcripts) and validation data (7,038 transcripts). This cross-chromosome train–test split prevented information leaking from intra-chromosomal interactions and potential overlaps of regulatory regions. The individual binding features (TF, RBP and miRNA) were normalized by their maximum values across genes. The relative fold-changes for 53 tissues were scaled together to set the standard deviation to one, and the median $\log_2$(TPM) was separately scaled to set the standard deviation to one. These steps were conducted for the training data first and then the same scaling factors were used for the validation and testing data to avoid information leaking from those data. We constructed and trained DL models using Keras (version 2.1.3) with a TensorFlow (version 1.4.1) backend. Hyper-parameters were optimized using hyperopt (version 0.2)[33] based on the mean-squared error against the validation data. The detailed structure of the model is described in Supplementary Fig. 10. The training was done using mini-batches of 128 training examples with a learning rate of 0.001 for the Adam optimizer. The number of maximum training epochs was set to 100, with early stopping of 10 based on validation loss. This training cycle was repeated 10 times and the best model for the validation data was selected as the final model (Supplementary Fig. 11). All models were trained using TITAN X Pascal graphics processing units (Nvidia).

**Comparison of DEcode and ExPecto.** To perform a fair comparison between DEcode and ExPecto[11], we used 18,550 genes that were commonly included in both studies and trained models with the same set of genes for training and evaluation. Because the ExPecto model was originally built using genes on chromosome 8 as the testing data, we followed the same procedure, reserving all 714 genes coded on chromosome 8 as the testing data and the rest of the genes being randomly split into training data (16,052 genes) and validation data (1,784 genes). The epigenetic states estimated by ExPecto were downloaded from the ExPecto repository (https://github.com/FunctionLab/ExPecto) as of November 2019. Given the epigenetic states, we built a prediction model for tissue-specific gene expression for each tissue via XGBoost based on the training script downloaded from the ExPecto repository. We modified the original script so that the early stopping of the model optimization was decided based on the performance on the validation data instead of the testing data. This modification prevented overfitting of the model to the testing data. We used the same hyper-parameters for XGBoost as in the script. We built 10 models for each method using the same genes for training, validation and testing to predict gene expression in the 53 tissues.

**DeepLIFT score calculation.** To evaluate the importance of input features to the prediction, we calculated DeepLIFT (Deep SHAP) scores[17] using DeepExplainer implementation (version 0.27.0)[18]. The DeepLIFT method estimates the contribution of each input compared to a reference input in a trained DL model. To compute the contribution of the presence of a binding site, we used a reference that does not have any binding sites in both promoters and RNAs with the median length of all genes in the testing data. DeepLIFT scores follow a summation-to-delta property where the summation of input contributions (DeepLIFT scores) is equal to the difference in the predicted value compared to the prediction from the reference input. We calculated DeepLIFT scores for each gene and transcript in the testing data for each of the 53 outputs, then summed the scores over promoter or RNA regions for each feature, and finally averaged them over genes. To estimate the collective impact on differential expression across 53 tissues, we averaged the absolute DeepLIFT scores of all tissues for each regulator. To evaluate the regulator's contribution to the prediction of transcript usage, we calculated the variance of DeepLIFT scores across transcript variants for each gene and then averaged them over genes. We performed pre-ranked GSEA[34] to assess the GO terms in MSigDB v7.1[35] associated with key TFs and RBPs. The absolute DeepLIFT scores were log-transformed and standardized to make them close to a normal distribution and then applied to GSEA with 10,000 permutations. The enrichment map was utilized for visualization of the results[36].

**Disease genes.** The probability that a gene is intolerant for a LoF mutation was downloaded from release 0.3.1 of the ExAC portal (http://exac.broadinstitute.org). We defined genes with probability greater than 99% as intolerant of homozygous or heterozygous LoF mutations. Disease genes were obtained from the OMIM portal as of June 2019 (https://www.omim.org/). We excluded provisional gene-to-phenotype associations and genes associated with non-disease phenotypes, multifactorial disorders or infection. We obtained mouse-lethal genes from the Gene Discovery Informatics Toolkit (v1.0.0)[37], which provided pre-processed gene lists from the murine KO experiments registered in Mouse Genome Informatics[38] and the International Mouse Phenotyping Consortium[39]. The results of CRISPR screening for the genes essential for proliferation or viability conducted in the DepMap project[40] were downloaded from the Enrichr portal[41,42] as of June 2019 (https://amp.pharm.mssm.edu/Enrichr). The Enrichr portal provided two CRISPR screening results, conducted independently at the Broad Institute and the Sanger Institute. To reduce the false positives in the CRISPR screening, we used essential genes that were identified in both independent screenings. Protein–protein interaction networks were downloaded from eXpression2Kinases Web[43], which consists of 210,221 interactions across 15,716 proteins.

**Training person-specific models.** To train person-specific models, we utilized the same model structure as the tissue model, except that the number of model outputs was modified to match the sample size of the tissue. We reused the parameters of convolutional layers in the tissue model and only parameters in the fully connected layers were tuned (Supplementary Fig. 12). We used the same gene splits and the same procedure of normalization and scaling as the tissue model for training and evaluating models. We evaluated the model prediction for each individual separately based on validation data and filtered out the individual models that performed at less than the 50% percentile of all individual models for some analyses.

**Training PrediXcan models.** To build a prediction model for gene expression from genotype data, we trained PrediXcan[29] models with GTEx gene expression and genotype data. A QCed vcf file of GTEx genotype data called by whole-genome sequence was downloaded from dbGaP for 635 individuals. We filtered out variants with a missing rate greater than 1% and minor allele frequency less than 1% and kept 9,219,660 variants for PrediXcan. We followed the model building procedure employed in PredictDB (http://predictdb.org/), a repository of PrediXcan models, as of November 2019. Briefly, we randomly split the samples into five folds. Then, for each fold, we removed the fold from the data and used the remaining data to train an elastic-net model using 10-fold cross-validation to tune the lambda parameter. With the trained model, we predicted gene expression values for the hold out samples. We applied the PrediXcan method to predict the same gene expression data used for the person-specific DEcode models. We built PrediXcan model for each gene using variants located within 1 Mbp upstream and downstream of its TSS. A missing value of the genotype data was replaced with an average dosage of non-missing samples.

**DE analysis for age and sex.** Limma[44] was used to identify genes associated with age using gender as a covariate. The $\log_2$(TPM) values of genes in the testing data were used. We also tested the associations between DeepLIFT scores for predictors and age via limma to identify regulators for DE against age and sex. The Benjamini–Hochberg procedure was used to control the FDR at 5%.

**In silico binding site disruption experiment.** To simulate the consequence of the removal of binding sites for the expression of *LAPTM5* and *CD53*, we generated 10,000 synthetic inputs for each of *LAPTM5* and *CD53*, where all binding sites in each interval of its promoter and RNA were randomly removed. From each of

these synthetic inputs, we computed predicted expression values and correlated them with those of another gene without any disruptions in its binding sites. Then, we used multiple linear regression to associate the location of disrupted regions with the correlation values between *LAPTM5* and *CD53* to estimate the effects of the disruption in each region on the co-expression relationship.

**In silico KO experiment.** To simulate the effect of regulator KO on the expression of *LAPTM5* and *CD53*, we generated 10,000 synthetic inputs for each *LAPTM5* and *CD53*, where each protein or miRNA bound to its promoter or RNA was randomly removed from the feature matrices. From each of these synthetic inputs, we computed predicted expression values and correlated them with those of another gene without any removals in its feature matrices. We then used multiple linear regression to associate the KOs of regulators with the correlation values between *LAPTM5* and *CD53* to estimate the effects of the KO of each regulator on the co-expression relationship. We applied the Bonferroni correction to control multiple testing, and the regulators with corrected $P$ values of <0.05 in all tissues were chosen as the key drivers of the co-expression.

**Conditional independence test.** To validate the effect of the predicted drivers on co-expression, we conducted a mediation analysis using a conditional independence test. A mediation analysis evaluated the hypothesis whereby, if *LAPTM5* and *CD53* are co-expressed due to the predicted regulators, normalizing expression levels of the two genes by the expression levels of the regulators would decrease the co-expression relationships. Specifically, we regressed the actual $\log_2$(TPM) values of *LAPTM5* and *CD53* with the actual $\log_2$(TPM) values of the predicted drivers and computed $R^2$ (variance explained) between the residuals of two genes. The $R^2$ based on the actual gene expression and one from the residuals were compared to quantify the covariance explained by the predicted drivers. To evaluate the significance of this effect, we repeated this process 1,000 times with an equal number of randomly picked genes that have a binding site in *LAPTM5* or *CD53* as regressors.

**DEcode model for DE prior rank.** The DE prior rank was downloaded from https://github.com/maggiecrow/DEprior. In the DE prior rank, each gene has a probability-like value, where zero is the minimum and one is the maximum. To convert this value to a non-bounded scale, we applied the logit transformation to the DE prior value. We assigned a value of 10 to a gene that had an infinite value after logit transformation. We used the same gene splits as the GTEx tissue model, which resulted in 13,433 genes for training, 1,504 genes for validation and 1,674 genes for testing. We trained the DEcode model for DE prior rank using the same procedure as with the GTEx person-specific models. To evaluate the contribution of promoter and RNA features to the prediction, the model was also trained with randomized input features. Receiver operating characteristic curve analysis was performed using pROC R package[45] with a default setting. We performed pathway analysis of the TFs with a DeepLIFT score greater than the 90th percentile using KEGG pathways[46]. KEGG pathway gene sets were downloaded from MSigDB v6.1[35]. The enrichment significance was based on the results of the hypergeometric test, with 757 unique TF genes as a background, against KEGG pathways composed of at least five background genes. FDR was controlled at 5%. We manually curated the 159 disease-related datasets used in the construction of the DE prior ranking to determine the number of datasets related to cancer or inflammatory disease.

**Definition of boxplot elements.** We used the R ggplot2 package to draw the boxplots in this Article with the default setting. The upper, centre and lower lines of the boxplots indicate 75%, 50% and 25% quantiles, respectively. The upper and lower whiskers of the boxplot indicate 75% quantile + 1.5 × interquartile range (IQR) and 25% quantile − 1.5 × IQR.

## Data availability

The data sources for GTEx data, TF and RBP binding peaks, miRNA binding locations, disease-related genes, protein–protein interaction data, pathways, gene ontology and the DE prior rank that were used for model training and interpretation are available in Supplementary Table 8. Processed data are available through our Code Ocean capsule (https://doi.org/10.24433/CO.0084803.v1).

## Code availability

DEcode software and pre-trained models for tissue- and person-specific transcriptomes are available at www.differentialexpression.org, https://github.com/stasaki/DEcode and from our Code Ocean capsule (https://doi.org/10.24433/CO.0084803.v1). DEcode is licensed under a BSD 3-Clause license.

Received: 19 February 2020; Accepted: 15 June 2020;
Published online: 6 July 2020

## References

1. Lee, T. & Young, R. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
2. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
3. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).
4. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
5. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).
6. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
7. Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. *Cell* **169**, 1187–1200 (2017).
8. Avsec, Ž. et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.* **37**, 592–600 (2019).
9. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
10. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
11. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
12. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
13. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.* **45**, D61–D67 (2017).
14. Zhu, Y. et al. POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.* **47**, D203–D211 (2019).
15. Agarwal, V., Bell, G. W., Nam, J. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
16. Melé, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
17. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research* Vol. 70 (eds Precup, D. & Teh, Y. W.) 3145–3153 (ICML, 2017).
18. Lundberg, S. M. & Lee, S. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4765–4774 (2017).
19. Chong, J. A. et al. REST: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell* **80**, 949–957 (1995).
20. Imperato, M. R., Cauchy, P., Obier, N. & Bonifer, C. The RUNX1-PU.1 axis in the control of hematopoiesis. *Int. J. Hematol.* **101**, 319–329 (2015).
21. Soares, E. & Zhou, H. Master regulatory role of p63 in epidermal development and disease. *Cell. Mol. Life Sci.* **75**, 1179–1190 (2018).
22. Watt, A. J., Garrison, W. D. & Duncan, S. A. HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology* **37**, 1249–1253 (2003).
23. Lefterova, M. I., Haakonsson, A. K., Lazar, M. A. & Mandrup, S. PPARγ and the global map of adipogenesis and beyond. *Trends Endocrinol. Metab.* **25**, 293–302 (2014).
24. Ge, Z., Quek, B. L., Beemon, K. L. & Hogg, J. R. Polypyrimidine tract binding protein 1 protects mRNAs from recognition by the nonsense-mediated mRNA decay pathway. *eLife* **5**, e11155 (2016).
25. Wang, Y. et al. $N^6$-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.* **16**, 191–198 (2014).
26. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
27. Goh, K. et al. The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
28. Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
29. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
30. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
31. Gaiteri, C., Ding, Y., French, B., Tseng, G. C. & Sibille, E. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes Brain Behav.* **13**, 13–24 (2014).
32. Crow, M., Lim, N., Ballouz, S., Pavlidis, P. & Gillis, J. Predictability of human differential gene expression. *Proc. Natl Acad. Sci. USA* **116**, 6491–6500 (2019).
33. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. & Cox, D. D. Hyperopt: a Python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **8**, 014008 (2015).
34. Korotkevich, G., Sukhov, V. & Sergushichev, A. Fast gene set enrichment analysis. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/060012v2 (2019).
35. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

36. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
37. Dawes, R., Lek, M. & Cooper, S. T. Gene discovery informatics toolkit defines candidate genes for unexplained infertility and prenatal or infantile mortality. *NPJ Genom. Med.* **4**, 8–11 (2019).
38. Smith, C. L., Blake, J. A., Kadin, J. A., Richardson, J. E. & Bult, C. J. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res.* **46**, D836–D842 (2018).
39. Koscielny, G. et al. The International Mouse Phenotyping Consortium web portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.* **42**, D802–D809 (2014).
40. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576 (2017).
41. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
42. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
43. Clarke, D. J. B. et al. eXpression2Kinases (X2K) Web: linking expression signatures to upstream cell signaling networks. *Nucleic Acids Res.* **46**, W171–W179 (2018).
44. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
45. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).


## Acknowledgements

We thank L. Yu for managing access to Genotype-Tissue Expression (GTEx) data. The study was supported by NIH grants P30AG010161, R01AG061798 and R01AG057911. The GTEx Project is supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. The data used for the analyses described in this Article were obtained from the GTEx Portal on 1 October 2018.

## Author contributions

S.T. contributed to the conception and design of the study. S.T. performed the computational analysis. S.T., C.G., S.M. and Y.W. interpreted the results. S.T. wrote the first draft of the manuscript. All authors contributed to manuscript revision, then read and approved the submitted version.

## Competing interests

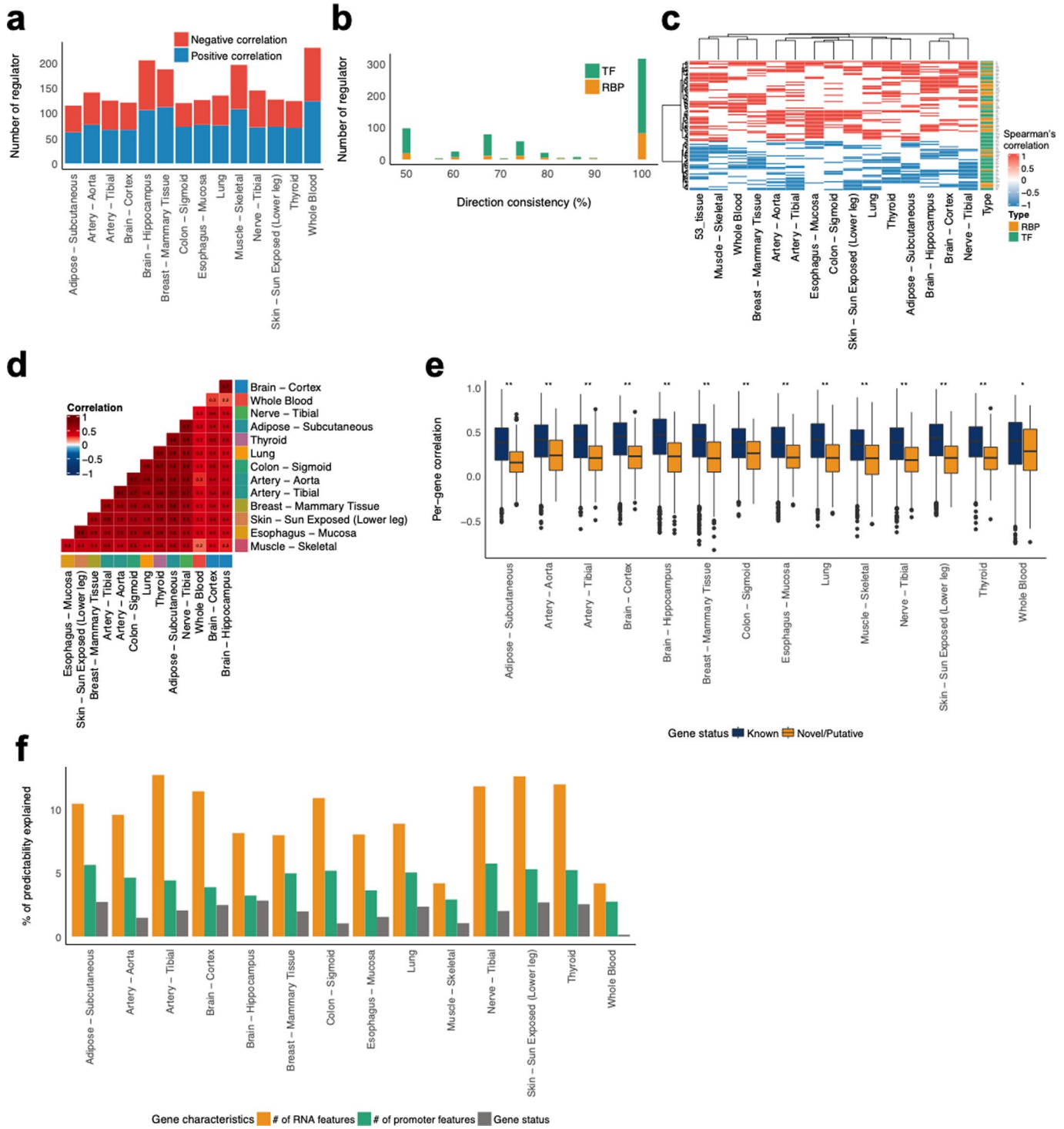
The authors declare no competing interests.

## Additional information

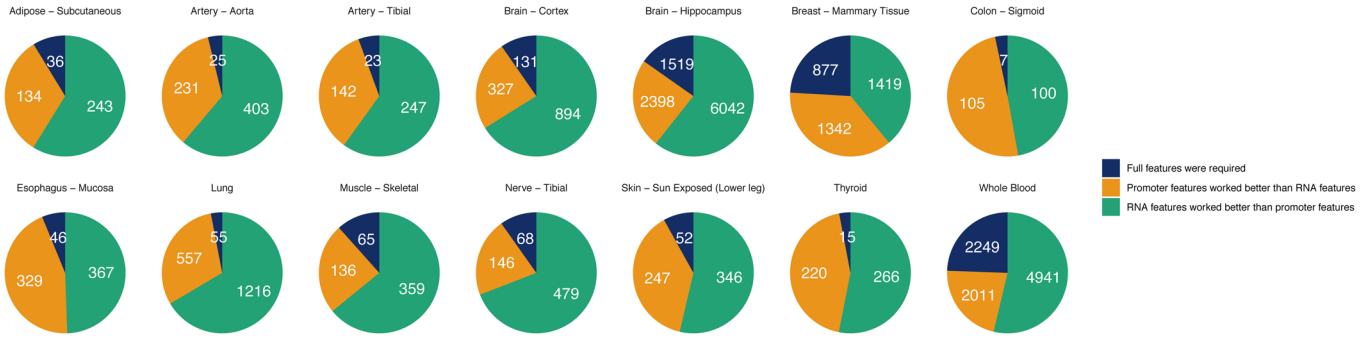
**Extended data** is available for this paper at https://doi.org/10.1038/s42256-020-0201-6.

**Supplementary information** is available for this paper at https://doi.org/10.1038/s42256-020-0201-6.

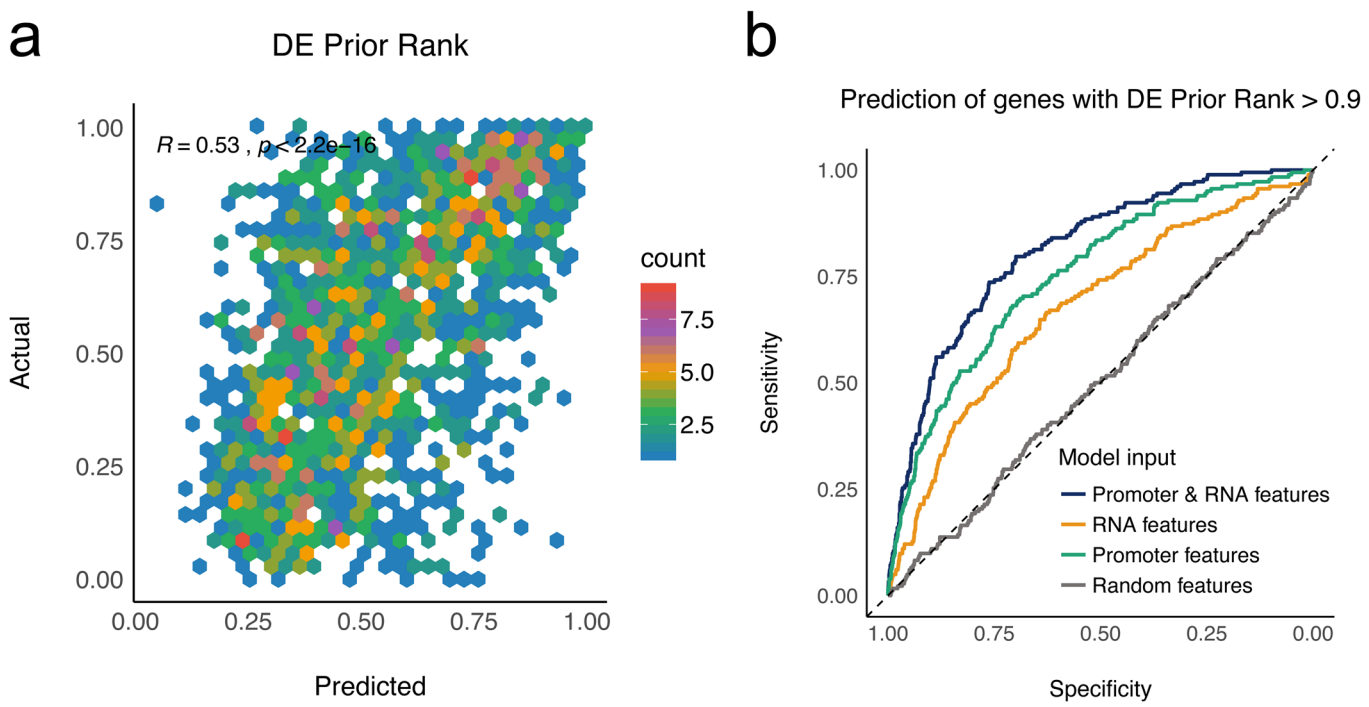
**Correspondence and requests for materials** should be addressed to S.T.




**Extended Data Fig. 1 | Characterizations of the person-specific models. a**, The consistency between the predictive contributions of regulators and their expression. Spearman's correlation was used to evaluate the relation between DeepLIFT scores vs log2-TPMs in each tissue. A relation with the absolute correlation > 0.3 and FDR < 5% was defined as significant. **b**, The consistency of predictive contributions across tissues. We selected the regulators that showed the significant relationships between their DeepLIFT scores with log2-TPMs in multiple tissues. Then, the consistency of directions of the correlations was assessed and visualized as a histogram. **c**, The actual correlation profile between DeepLIFT scores and log2-TPMs. We selected 99 regulators that showed consistent relationships in more than four tissues. **d**, The pairwise similarity of per-gene prediction accuracy between tissues. Spearman's correlation was used for this comparison. **e**, The association of per-gene prediction accuracy with the gene status. The gene status indicates whether genes are registered in multiple databases (Known) or only in the GENCODE database (Novel or Putative). **f**, Decomposition of variance in the per-gene prediction accuracy. To compare the variances in the per-gene prediction accuracy explained by the gene status, the number of promoter features, and RNA features, we used a variance decomposition method (lmg) implemented in relaimpo R package.




**Extended Data Fig. 2 | The major feature types contributed to the gene co-expression.** We defined the gene pairs with the absolute Spearman's correlation greater than 0.3 and the sign of the correlation matched with one with the ground truth as the successfully predicted gene pairs. The successfully predicted gene pairs of the model trained with the full set of features were split into three groups based on the performance of the models trained with only RNA features or promoter features.




**Extended Data Fig. 3 | DEcode predicts the regulatory principles behind frequently DE genes across diverse conditions. a**, The scatter plots showing the relations between predicted and actual DE prior rank. The predicted logit of DE prior rank was converted to probability and compared with actual DE prior rank with Spearman's correlation. **b**, The performances of the models trained with a distinct feature set. ROCs represent the performances of models in predicting genes with whose DE prior rank greater than 0.9.