nature genetics

Letter

https://doi.org/10.1038/s41588-023-01524-6

Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings

Received: 13 March 2023

Accepted: 8 September 2023

Published online: 30 November 2023

Check for updates

Alexander Sasse[1,7], Bernard Ng[2,7], Anna E. Spiro[1,7], Shinya Tasaki[2], David A. Bennett[2], Christopher Gaiteri[2,3], Philip L. De Jager[4], Maria Chikina[5] ✉ & Sara Mostafavi[1,6] ✉

Deep learning methods have recently become the state of the art in a variety of regulatory genomic tasks[1–6], including the prediction of gene expression from genomic DNA. As such, these methods promise to serve as important tools in interpreting the full spectrum of genetic variation observed in personal genomes. Previous evaluation strategies have assessed their predictions of gene expression across genomic regions; however, systematic benchmarking is lacking to assess their predictions across individuals, which would directly evaluate their utility as personal DNA interpreters. We used paired whole genome sequencing and gene expression from 839 individuals in the ROSMAP study[7] to evaluate the ability of current methods to predict gene expression variation across individuals at varied loci. Our approach identifies a limitation of current methods to correctly predict the direction of variant effects. We show that this limitation stems from insufficiently learned sequence motif grammar and suggest new model training strategies to improve performance.

Sequence-based deep learning methods are emerging as powerful tools for a variety of functional genomic prediction tasks. These models take as input genomic DNA and learn to predict context-dependent functional outputs such as transcription factor binding[2,8,9], chromatin state[10–13] and gene expression values[1,14]. State-of-the-art models can reproduce experimental measurements with a high degree of accuracy and enable mechanistic insights through their learned DNA features[1,2,15]. Yet, the true potential of these sequence-based models lies in their ability to predict outcomes for arbitrary sequence inputs—a space too large for experimental methods to fully explore. While partial evaluations through expression quantitative trait locus (eQTL)[1,16] studies or massively parallel reporter assays (MPRAs)[17] have shown promise, the broader application of these models as personalized DNA interpreters has not been comprehensively assessed. We address this by conducting extensive analyses using paired whole genome sequencing (WGS) and cerebral cortex RNA-sequencing (RNA-seq) data from the ROSMAP datasets[7] with measurements from 839 individuals. Our study bridges the gap between the known potential and the actual performance of these models in personalized genomics interpretation.

To start, we focus our evaluation on Enformer[1], the top-performing deep learning model. Enformer is trained to predict various functional outputs from (*cis*) sub-sequences from the reference genome. This training approach allows Enformer and other deep learning models

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [2]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA. [3]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA. [4]Center for Translational & Computational Neuroimmunology, Department of Neurology, and the Taub Institute for the Study of Alzheimer's Disease and the Aging Brain, Columbia University Irving Medical Center, New York, NY, USA. [5]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA. [6]Canadian Institute for Advanced Research, Toronto, Ontario, Canada. [7]These authors contributed equally: Alexander Sasse, Bernard Ng, Anna E. Spiro. ✉e-mail: mchikina@gmail.com; saramos@cs.washington.edu

https://doi.org/10.1038/s41588-023-01524-6

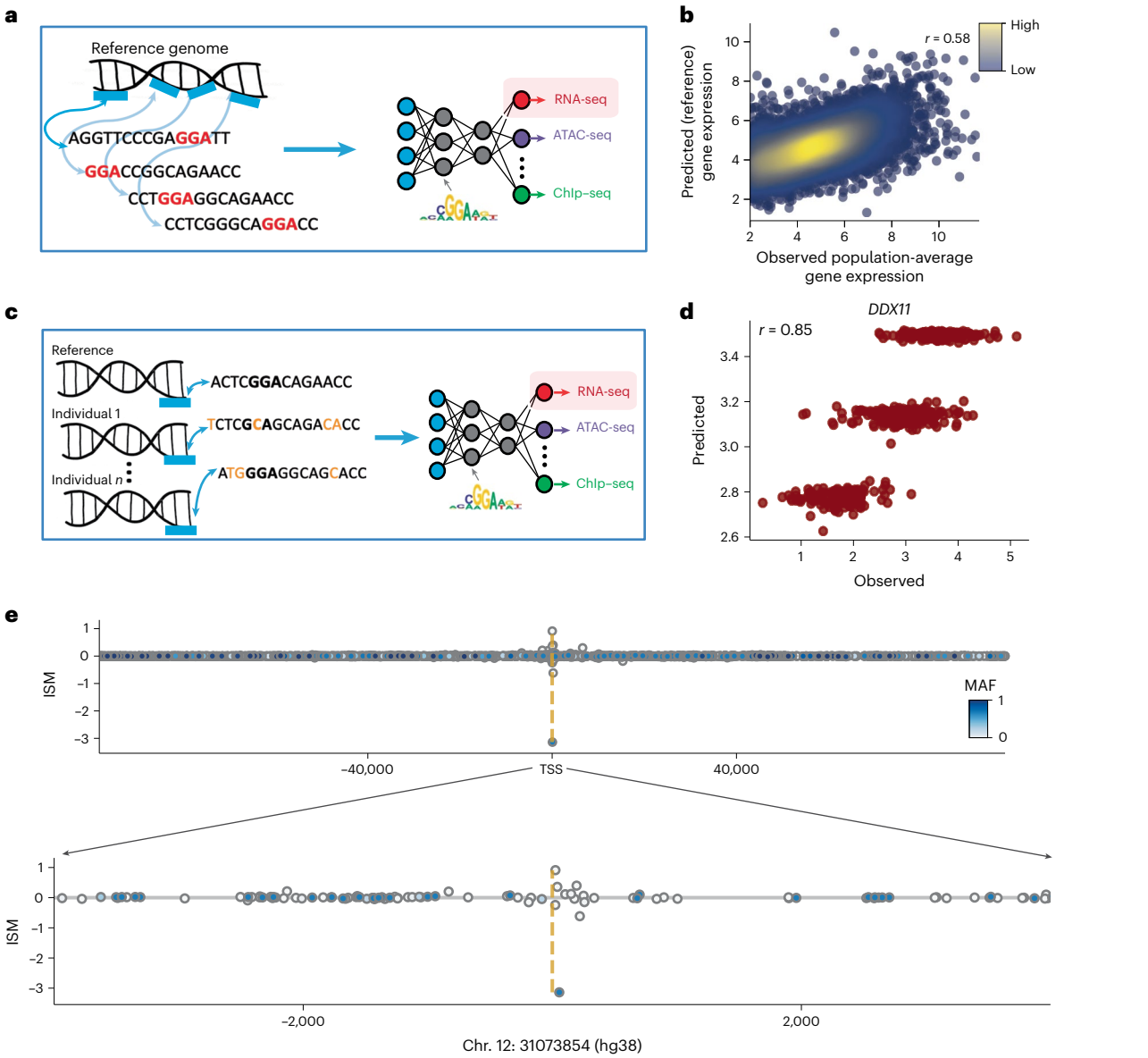


Fig. 1 | Evaluation of Enformer across genomic regions and select loci. **a**, Schematic of the reference-based training approach. Different genomic regions from the reference genome are treated as data points. Genomic DNA underlying a given region is the input to the model, and the model learns to predict various functional properties, including gene expression (CAGE–seq), chromatin accessibility (ATAC-seq) or transcription factor binding (ChIP–seq). **b**, Population-average gene expression levels in the cerebral cortex (averaged in ROSMAP samples, $n = 839$) for expressed genes ($n = 13{,}397$) versus Enformer's predictions. **c**, Schematic of the per-locus evaluation strategy. **d**, Predicted and observed *DDX11* gene expression levels in the cortex for individuals in the ROSMAP cohort ($n = 839$). Each dot represents one individual. The output of Enformer is fine-tuned using an elastic net model (Methods). **e**, ISM values for all SNVs that occur at least once in 839 genomes within 98 kb of the *DDX11* TSS. SNVs are colored according to minor allele frequency (MAF).

to identify short DNA sub-sequences (motifs) that are shared across the genome and exploits variations in motif combinations across genomic regions to make context-dependent predictions. As a control experiment, we used the pre-trained Enfomer model, provided it with sub-sequences around the transcription start site (TSS) from the reference genome and evaluated its predictions on population-average gene expression ($n = 13{,}397$ expressed protein-coding genes) from the cerebral cortex (Fig. 1a,b). To account for the differences between the data types that were used during Enformer's training and our study, we used a fine-tuning strategy, whereby we trained an elastic net model on top of the predictions from Enformer's output tracks (Methods). Consistent with the expectation for this type of evaluation, we observed good prediction accuracy as measured by Pearson's correlation coefficient (Pearson's r) = 0.58 (Fig. 1b). The results were similar when we restricted the analysis to a smaller set of genes ($n = 3{,}401$) overlapping Enformer's test regions (Pearson's $r = 0.51$; Supplementary Fig. 1).

While Enformer is not explicitly trained on genetic variation data, once trained, it holds promise that it has learned the *cis*-regulatory logic of gene expression and so can predict the impact of arbitrary genetic variation on its outputs. To evaluate its performance in this setting, which is distinct from the cross-genome performance evaluated above, we applied Enformer to predict individual-specific gene expression levels based on personal genomic sequences (Methods and Fig. 1c). As a positive example, we present results for a highly heritable gene, *DDX11* (heritability $r^2 = 0.8$). *DDX11*'s variance in expression across individuals can be attributed to a single causal single-nucleotide variant (SNV) using statistical fine-mapping[16]. Using WGS data, we created 839 input sequences of length 196,608 bp centered at the TSS, one per

https://doi.org/10.1038/s41588-023-01524-6

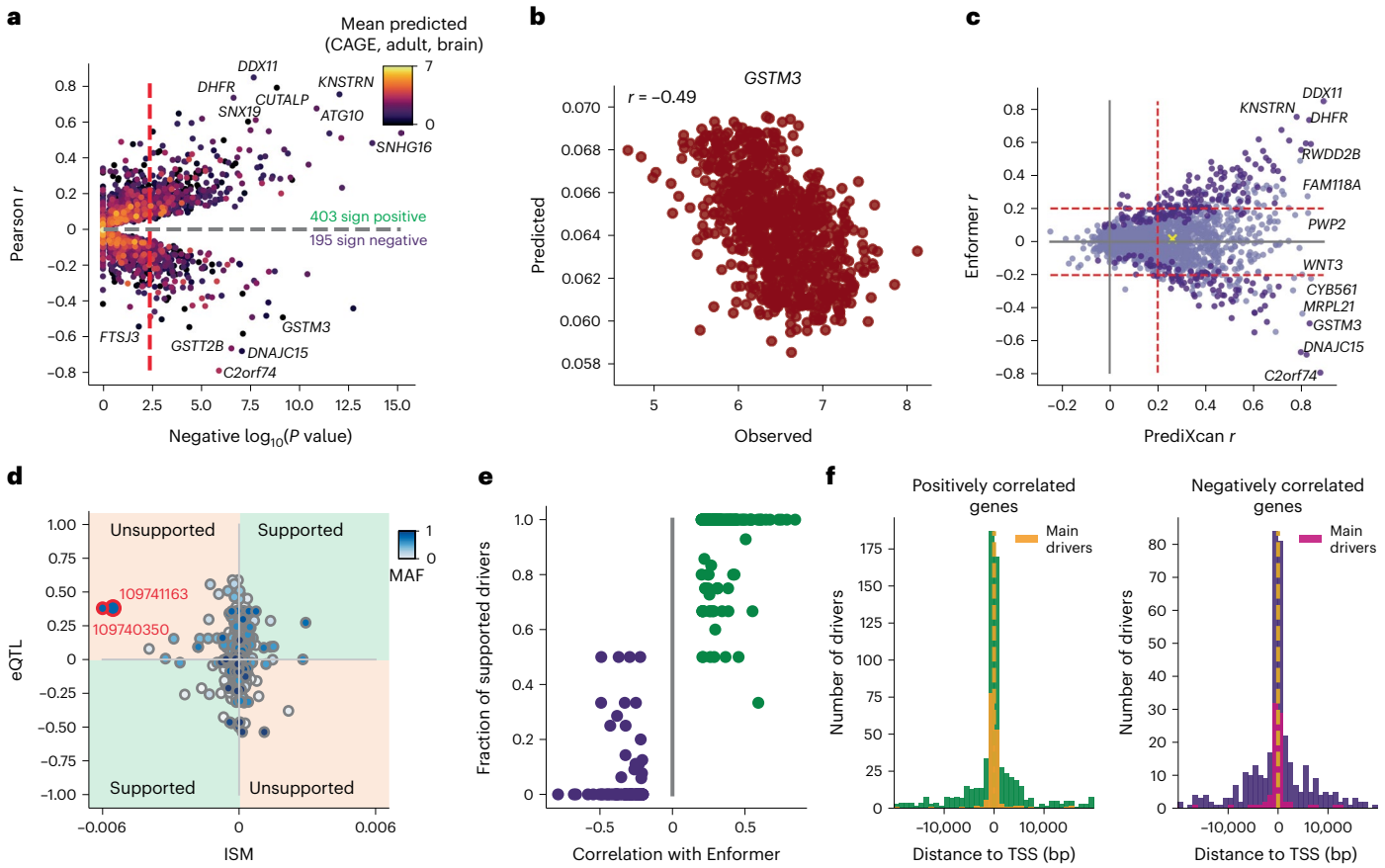


Fig. 2 | Evaluation of Enformer on prediction of gene expression across individuals. a, The *y* axis shows the Pearson's *r* coefficient between observed expression values and Enformer's predicted values per gene ($n = 6{,}825$ genes). The *x* axis shows the negative $\log_{10}$ (*P*-value), computed using a gene-specific null model (Methods; one-sided *t*-test, permutation analysis with $n = 50$ independent samples per gene). The colors represent the predicted mean expression using the most relevant Enformer output track ('CAGE, adult, brain'). The red dashed line indicates $FDR_{BH} = 0.05$. **b**, The *y* axis shows the prediction from Enformer's 'CAGE, adult, brain' track across individuals for the *GSTM3* gene ($n = 839$). The *x* axis shows the observed gene expression values. **c**, Pearson's *r* coefficients between PrediXcan-predicted versus observed expression across individuals are shown on the *x* axis, Enformer's Pearson's *r* coefficients are shown on the *y* axis. Red lines indicate the threshold for significance (abs(*r*) > 0.2, Bonferroni-corrected nominal *P*-value). Darker colored dots are significant genes from **a**. The yellow cross represents the location of the mean across all *x* and *y* values. **d**, ISM value versus eQTL effect size for all SNVs ($n = 706$ with MAF > 0.01) within the 196 kb input sequence of the *GSTM3* gene. Red circles represent driver SNVs. SNVs are defined as supported or unsupported based on the concordance of the sign of the eQTL effect size. **e**, Fraction of supported driver SNVs per gene (*y* axis) versus Pearson's *r* coefficients between Enformer's predictions and observed expression values (*x* axis) ($n = 161$ positively correlated genes, $n = 87$ negatively correlated genes). **f**, Number of driver SNVs within the 1,000-bp window centered on the TSS. The main drivers are those with the strongest impact on linear approximation, shown in different colors. Left plot, $n = 983$ driver SNVs; right plot, $n = 564$ driver SNVs.

individual for the gene (Fig. 1c Supplementary Fig. 2 and Methods). Applying Enformer to these input sequences we observed a Pearson's correlation coefficient of 0.85 ($P < 1 \times 10^{-200}$) between predicted and observed gene expression levels across individuals (Fig. 1d). Further, in silico mutagenesis (ISM) at this locus showed that Enformer used a single SNV with high correlation to gene expression (eQTLs) in making its predictions (Fig. 1e). This SNV is the same causal SNV that was identified through statistical fine-mapping with Susie[16]. Thus, at this locus, Enformer can identify the causal SNV among all those in linkage disequilibrium (LD), and in addition provides hypotheses about the underlying functional cause, in this case, the extension of a repressive motif (Supplementary Fig. 3).

However, the impressive predictions on *DDX11* proved to be the exception rather than the rule. When we tested 6,825 cortex-expressed genes, we found a large distribution in the Pearson's *r* (Fig. 2a and Supplementary Table 1; minimum $r = -0.76$, maximum $r = 0.84$, mean = 0.01). Surprisingly, while the predictions were significantly correlated with observed expression for 598 genes (Benjamini–Hochberg false-discovery rate (FDR_{BH}) = 0.05; Methods), they were significantly anti-correlated with the true gene expression for 195 (33%) of these genes. For example, predicted *GSTM3* gene expression values were anti-correlated with the observed values ($r = -0.49$; $P < 1 \times 10^{-200}$; Fig. 2b). We performed several sensitivity analyses to which these results proved robust (Methods and Extended Data Fig. 1): (1) these results were not sensitive to output track fine-tuning, (2) to model ensembling as done in Enformer or (3) to subsetting the analysis to a smaller set of genes that have easily detectable causal variants based on statistical fine-mapping (Supplementary Table 1). Overall, these results imply that the model fails to correctly attribute the variants' direction of effect (that is, whether a given variant decreases or increases gene expression level).

We then compared Enformer against a widely used linear approach called PrediXcan[18]. PrediXcan constructs an elastic net model for each gene from *cis* genotype SNVs across individuals. Unlike Enformer, PrediXcan is explicitly trained to predict gene expression from variants, but it does not take into account variants that were not present in its training data and cannot output a prediction for unseen variants. While the models are conceptually different, the PrediXcan model gives a lower bound on the fraction of gene expression variance that can be predicted from genotype. Further, genes that are significantly predicted with the

PrediXcan should have at least one causal variant somewhere in the genomic region used for the predictions, thus providing a substantial set of loci for assessing Enformer's predictions. We used a previously published PrediXcan model that was trained on Genotypic tissue Expression (GTEx) cerebral cortex data[18] and applied it to ROSMAP samples. Hence, neither Enformer nor PrediXcan had seen the ROSMAP samples before their application. As shown in Fig. 2c, for the 1,570 genes where PrediXcan's elastic net model was available, the performance of Enformer was substantially lower (921 significantly predicted genes by PrediXcan versus 162 by Enformer; mean r Enformer = 0.02, mean r PrediXcan = 0.26; Supplementary Table 1). Further, PrediXcan did not have the same challenge with misprediction of the direction of the SNV effect (that is, all of PrediXcan's significantly predicted genes had a positive correlation between predicted and observed values). When we ignored the sign of Enformer's correlation values, we observed that both models, despite their conceptual differences, showed some predictive power for the same genes (r = 0.58; Supplementary Fig. 4). This supports the observation that Enformer can identify genes at which genetic variation across individuals significantly impacts gene expression values, but unlike PrediXcan, Enformer is not able to determine the sign of SNV effects accurately. We note that Enformer predictions were evaluated against eQTLs in the original study using signed linkage disequilibrium profile (SLDP) regression[1,19], demonstrating improved performance over competing models in terms of z score; however, this previous result was based on evaluation across the genome, and was not locus specific as we report here.

To investigate whether these observations are specific to Enformer or more broadly apply to sequence-based deep learning models that follow the same training recipe, we trained a simple convolutional neural network (CNN) that takes as input sub-sequences from the reference genome centered at the TSSs of genes (40 kb) and predicts population-average RNA-seq gene expression in the cerebral cortex as output (Methods). This CNN could predict population-average gene expression in the cortex with similar accuracy as Enformer (r = 0.57; Extended Data Fig. 2a), yet it has the same challenge with the direction of the predictions across individuals (Extended Data Fig. 2b). Thus, our results on Enformer are likely to generalize to other sequence-based deep learning models trained in the same way.

To explore the causes for the negative correlation between Enformer predictions and the observed gene expression values we used two explainable artificial intelligence (AI) approaches: ISM and input-Gradient (Supplementary Methods 2). These approaches approximate the output of a nonlinear neural network with a linear function that weights the contribution of each SNV through a process referred to as feature attribution. First, we confirmed that this approximation was reasonable for 95% of the examined genes (Supplementary Figs. 5 and 6). For each gene, based on its ISM attributions, we determined the main SNV driver(s) that dominate the differential gene expression predictions across individuals (Supplementary Methods 2). Across the 256 examined genes, we found that 32% had a single SNV driver, and the vast majority (85%) had five or fewer drivers (Supplementary Fig. 7 and Supplementary Table 2) that determine the direction and correlation with the observed expression values. To understand how these driver SNVs cause mispredictions, we classified Enformer-identified driver SNVs into 'supported' and 'unsupported' categories based on the agreement of the SNV's ISM attribution sign with the direction of effect according to the eQTL analysis (Methods). For this analysis, we computed marginal eQTL effect sizes, which do not distinguish causal variants from others in LD. However, it is important to note that the Enformer model is entirely agnostic to LD structure as it was trained with a single reference genome. As such, Enformer predictions by construction assume a causal interpretation of the identified driver variants. Thus, a comparison of Enformer-identified driver variants is informative because sign discordance between the two strongly suggests that the Enformer effect is incorrect. On the other hand, the reverse analysis is not interpretable: an eQTL with a large marginal effect can have a low Enformer effect because it is not causal. As an example of sign discordance analysis, *GSTM3* had two common driver SNVs identified by Enformer, yet their predicted direction of effect was unsupported based on the SNVs signed eQTL effect sizes (Fig. 2d). For all 256 inspected genes, we found that mispredicted genes had almost exclusively unsupported driver SNVs, whereas correctly predicted genes indeed had supported driver SNVs (Fig. 2e). This analysis thus confirms that the small number of driver SNVs per gene is the cause of Enformer's misprediction for the sign of the effect.

To investigate whether these unsupported attributions are caused by systematically erroneous sequence-based motifs that Enformer learns from the training data, we analyzed the genomic sequences around driver SNVs. We did not find any enrichment for specific sequence motifs (Supplementary Fig. 8). When we plotted the location of SNV drivers along the input sequences, we found that most drivers were located close to the TSS (Fig. 2f, Supplementary Figs. 9 and 10 and Supplementary Methods 3), supporting a recent report[17] that showed that current sequence-based deep learning models mainly predict gene expression from genomic DNA close to the TSS, despite using larger input DNA sequences. Further, when we analyzed ISM values in windows around the driver SNVs, we observed that the majority did not fall into coherent 'attributional motifs' (short regions of sequence with consistent attribution) as would be expected if the model were picking up on biologically meaningful regulatory mechanisms (Supplementary Fig. 11, Supplementary Table 3 and Supplementary Methods 4).

In summary, our results suggest that current sequence-based deep learning models trained on the input-output pair of a single reference genome often fail to correctly predict the direction of SNV effects on gene expression. We further show that current neural network models perform worse than simple baseline approaches such as PrediXcan in predicting the impact of genetic variation across individuals. For future development, we recommend that new models be assessed not only on genome-wide statistics of absolute causal eQTL effect sizes, but also on the per-gene agreement between the sign and the size of the predicted and measured effect of causal variants.

We hypothesize that two complementary strategies will be fruitful for improving the prediction of gene expression across individuals. First, current models are trained on sequences from a single reference genome and learn sequence features that explain gene-to-gene expression variation; they thus have not been explicitly trained to learn how locus-dependent genetic variation impacts gene expression. The mechanisms that explain gene-to-gene variation may be distinct from those that explain interpersonal variation. For example, while promoter logic is important to determine which genes are expressed within a cell type, long-range interaction appears to be much more important for interpersonal variation[17]. Thus, training on input–output pairs of diverse genomes and their corresponding gene expression measurements may be a way to increase sequence variation and learn these effects for accurate personalized predictions. Second, current methods do not accurately model all of the biochemical processes that determine RNA abundance. For example, post-transcription RNA processing (whose dependence on sequence is mediated via RNA–protein or RNA–RNA interactions) is entirely ignored. While including datasets that explicitly measure post-transcriptional regulatory processes and long-range interaction may improve the modeling of these effects[4,6], it is also possible that, with sufficiently large paired WGS and gene expression training datasets, the resulting models will implicitly learn these mechanisms as long as they impact gene expression variation across individuals.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01524-6.

References

1. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
2. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
3. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**, 389–403 (2019).
4. Zhou, J. et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).
5. Zhou, J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat. Genet.* **54**, 725–734 (2022).
6. Park, C. Y. et al. Genome-wide landscape of RNA-binding protein target site dysregulation reveals a major impact on psychiatric disorder risk. *Nat. Genet.* **53**, 166–173 (2021).
7. De Jager, P. L. et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 180142 (2018).
8. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
9. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
10. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* https://doi.org/10.1038/s41592-022-01562-8 (2022).
11. Maslova, A. et al. Deep learning of immune cell differentiation. *Proc. Natl Acad. Sci. USA* **117**, 25655–25666 (2020).
12. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (2022).
13. Kim, D. S. et al. The dynamic, combinatorial *cis*-regulatory lexicon of epidermal differentiation. *Nat. Genet.* https://doi.org/10.1038/s41588-021-00947-3 (2021).
14. 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).
15. Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W. & Mostafavi, S. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-022-00532-2 (2022).
16. Wang, Q. S. et al. Leveraging supervised learning for functionally informed fine-mapping of *cis*-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
17. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**, 56 (2023).
18. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
19. Reshef, Y. A. et al. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).

Methods

No specific ethics approval was needed to conduct the current study.

WGS and RNA-seq datasets

We used $n = 839$ individuals with available WGS (blood) and RNA-seq (cerebral cortex) data from the ROS and MAP cohort studies[20] (previously described[21], also see Supplementary Methods). The 839 samples were from distinct individuals. Both studies were approved by the Institutional Review Board of Rush University Medical Center. All participants signed an informed and repository consent form and an anatomic gift act. Besides the availability of both WGS and cortex RNA-seq data after pre-processing, no other inclusion criteria were used.

Predicting gene expression with Enformer

Population-average gene expression. We centered the reference genome (GRCh38) around the gene's TSS (Gencode v27) and extracted the genomic sequence in the 196,608-bp window, which was then used as input to Enformer v1 (April 2022). We performed this analysis for 13,397 brain-expressed genes (for computational reasons, a random set of 6,825 genes among these were used in per-individual analyses described below). To use the outputs of Enformer predictions, we closely followed the previous methodology. Specifically, for a given input sequence, Enformer makes predictions for 5,313 human output tracks and 986 bins. The predictions were obtained for all 5,313 human output tracks, as the sum of log values from the three central 128-bp bins (bin numbers 447, 448 and 449) for each output track. We performed two types of summarization of the output tracks: (1) directly using the single track that best matched our RNA-seq gene expression data ('Cortex, adult, brain') and (2) using an elastic net model, trained on the predictions from all tracks and all expressed genes (that is, a matrix of 5,313 tracks × 13,397 brain-expressed genes), to predict population-average gene expression for adult cortex (using GTEx data). As discussed further in the 'Sensitivity analysis' section the results from these two types of analyses proved similar. Finally, we also performed our evaluation analysis on a smaller set of genes ($n = 3,401$) that overlapped test regions not used to train the Enformer model (Supplementary Fig. 1).

Predicting gene expression across individuals. For each individual and each gene, we constructed a personalized DNA sequence input (196,608 bp) from phased WGS data (separated maternal and paternal DNA sequence inputs were constructed for each individual and each gene). As above, we summed the log-transformed predicted values for the three central 128-bp bins (bin numbers 447, 448 and 449) for each output track. We used two methods to predict final gene expression: (1) 'fine-tuning' and 2) direct selection of a single track most representative of cortex gene expression data ('CAGE, brain, adult'). For fine-tuning, we trained an elastic net model to linearly weight all of Enformer's 5,313 human output tracks to predict population-average gene expression in the cerebral cortex, using the GTEx RNA-seq data (cortex). Specifically, the elastic net model was fit to predict population-average gene expression levels in the cortex from Enformer's predictions when the reference sequence centered at each gene's TSS was used as input. Enformer was used to make separate predictions from the maternal and paternal sequences. For each individual and each gene, we averaged the predictions from the maternal and paternal sequences.

Statistics and reproducibility

We used a sample size of 839 independent individuals to assess the significance of the model's predictions. This sample size is sufficient to assess significance across individuals and per gene, based on previous eQTL analyses[22,23]. We also note that no data from the complete initial dataset (where both WGS and RNA-seq samples passed quality control) were excluded from the analyses. Permutation analysis was used to complement the standard FDR_{BH} and Bonferroni-corrected P-values.

Deriving the gene-specific null distribution. Predicted gene expression for the 839 individuals is a function of SNV genotype for each gene and individual. Thus, we can linearly approximate Enformer's predictions for each gene and each individual as the weighted sum of the SNVs present in that individual for the given gene[18]. Therefore, to create a null distribution for predictions of gene expression value for each individual and each gene, we assigned random attribution weights to each SNV present in the given individual. Specifically, we sampled random normally distributed weights for every SNV within the 196,608-bp window around the TSS and summed them for each individual genotype as the random gene-specific predictions. For each gene, we generated 50 random predictors from which we derived the mean and standard deviation of the absolute Pearson's correlation to the observed expression values. To assign P-values to Enformer's correlation to observed gene expression, we used a one-sided t-test and the Benjamini–Hochberg procedure to target a FDR of 0.05.

Sensitivity analysis. We performed three types of sensitivity analyses to ensure our cross-individual prediction results were robust. First, we compared the predictions from a single relevant track (CAGE, cortex, adult) with the results when we fine-tuned the predictions with the elastic net model described above (trained on average gene expression prediction from all tracks, using data from GTEx) (Extended Data Fig. 1a). Second, we performed model ensembling, whereby we averaged model predictions on shifted sub-sequences and reverse and forward strands, but this did not impact the sign of significant correlations in ~96% of cases (Extended Data Fig. 1b). Third, we focused the analysis on 184 genes with known causal SNVs according to previous eQTL analysis[16], and again observed that while Enformer can make significant predictions, the predicted expression levels were anti-correlated for 80 (43%) of these genes (Extended Data Fig. 1c and Supplementary Table 1).

Training and testing of a simple CNN

Our simple CNN was trained on genes that were not located within the regions of the Enformer' test set. During training, we used sequences of length 40,001 bp from the reference genome centered at the TSS as input to the model and predicted mean log gene expression from the ROSMAP dataset (dorsolateral prefrontal cortex). The length of the input sequence was informed by a recent study[17]. This CNN has a very shallow architecture; it consists of a single convolutional layer with 900 kernels of width 10 and rectified linear unit (ReLU) activation. We applied a single average pooling layer of size 900 bp that reduces the input of the network to 44 segments. We then applied a single hidden layer of size 200 with ReLU activation before predicting the mean gene expression of the given gene. For training, we used mean squared error loss and the Adam optimizer with a learning rate of 0.001 and default hyperparameters. Then, for a random set of 190 individuals, we constructed a maternal and a paternal genomic sequence by inserting all the variant alleles within ± 20,000 bp of the TSS into the reference sequence. We then made separate predictions for the maternal and paternal sequences and averaged them for every individual. We computed the Pearson's correlation coefficient between the predicted and observed expression values for these 190 individuals and compared the absolute Pearson's r to the value that we would expect from our gene-specific null model for variants within ± 20,000 bp of the TSS.

Driver variant attribution scores using input-gradient and ISM

To explore the causes for the negative correlation between Enformer predictions and the observed gene expression values we applied two explainable AI techniques on all genes with a significant correlation value ($abs(r) > 0.2$; Fig. 2a): ISM and gradients[9,15,24]. Please see the Supplementary Methods for details on the rationale and methodology, as well as the procedure for identifying the Enformer 'driver SNVs' for predictions from WGS data.

Computing eQTL values and sorting drivers into supported and unsupported drivers

We computed eQTL effect sizes for a given SNV as the slope of the linear regression solution that predicts gene expression across individuals from this SNV genotype, that is, for individuals with two copies of the major allele (genotype 0), those with one copy of the major allele (genotype 1) and those with two copies of the minor allele (genotype 2). The slope of the regression with the genotype of each SNV represents how much expression changes with an additional copy of the minor allele. Positive or negative slopes determine the direction of the SNV effect on gene expression. Based on the eQTL effect size and ISM attribution values for each SNV, one can distinguish between supported and unsupported drivers. Supported drivers' attributions have the same sign as the eQTL effect size, whereas unsupported drivers have the opposite sign.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Genotype and RNA-seq data for the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) samples are available from the Synapse AMP-AD Data Portal (accession code syn2580853) as well as the RADC Research Resource Sharing Hub at www.radc.rush.edu.

Code availability

Scripts for running the analyses presented, as well as intermediate results, are available from https://github.com/mostafavilabuw/EnformerAssessment (ref. 25).

References

20. Bennett, D. A. et al. Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
21. Mostafavi, S. et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat. Neurosci.* **21**, 811–819 (2018).
22. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
23. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
24. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) Vol. 70 3319–3328 (PMLR, 2017); https://doi.org/10.5281/zenodo.8274879
25. Sasse, A, Ng, B, & Spiro, E. A. mostafavilabuw/EnformerAssessment: EnformerEvaluationV1. *Zenado* https://doi.org/10.5281/zenodo.8274879 (2023).

Acknowledgements

We thank D. R. Kelley for helpful comments on this manuscript. We thank the participants of ROS and MAP for their essential contributions and gifts to this project. This work has been supported by many different NIH grants, including P30AG10161 (to D.A.B.), P30AG72975 (to D.A.B.), R01AG15819 (to D.A.B.), R01AG17917 (to D.A.B.), U01AG46152 (to D.A.B. and P.L.D.), U01AG61356 (to D.A.B. and P.L.D.), R01AG057911 (to C.G.), R01AG06179 (to C.G.) and R01AG036836 (to P.L.D.), as well as a CIFAR research fellowship and an NSERC Discovery Grant (to S.M.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Conceived the study: S.M. and M.C. Study design: S.M., A.S. and M.C. Data generation and quality control analyses: B.N., A.E.S., C.G., P.L.D., S.T. and D.A.B. Analyses and interpretation: A.S., A.E.S., B.N., S.M. and M.C. Wrote the initial draft: S.M., A.S. and B.N. Read and provided comments on the manuscript: M.C., B.N., A.E.S., P.L.D., C.G., S.T. and D.A.B. Supervised the project: S.M. and M.C.

Competing interests

The authors declare no competing interests.

Additional information

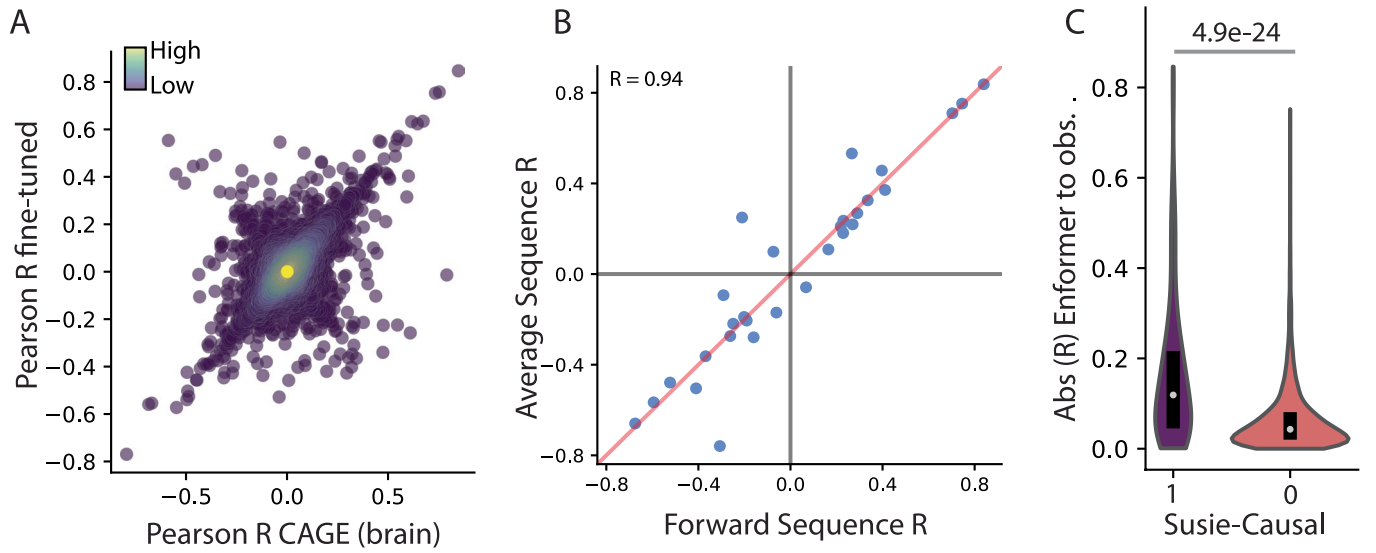
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01524-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01524-6.

Correspondence and requests for materials should be addressed to Maria Chikina or Sara Mostafavi.

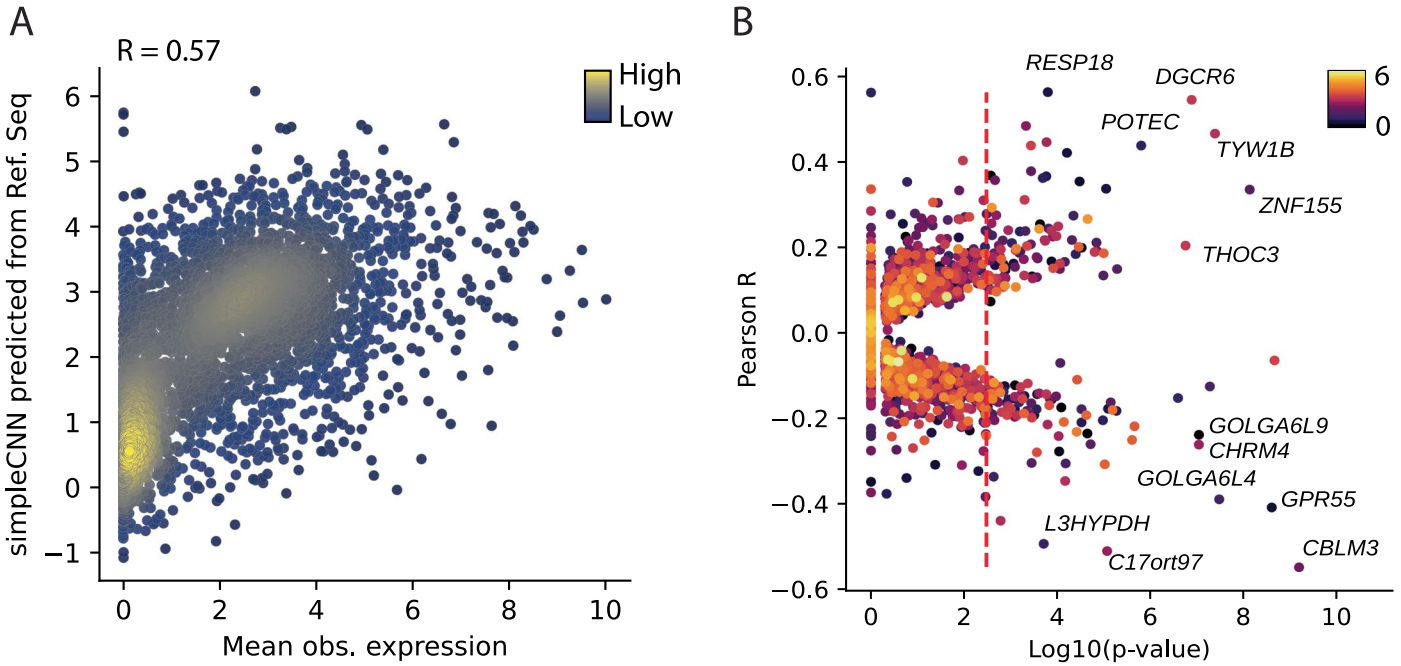
Peer review information *Nature Genetics* thanks Kaur Alasoo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-023-01524-6



Extended Data Fig. 1 | Sensitivity analysis for Enformer Predictions. (**a**) Density plot, where each dot represents a gene (n = 13,397). X-axis shows Pearson's *r* coefficients for Enformer predictions for the single most relevant track ('CAGE,brain,adult') and y-axis shows the fine-tuned cortex model from all human tracks. Color depicts local density. (**b**) Pearson's *r* coefficients across 839 individuals between observed expression and the predicted CAGE track from a single forward-stranded input sequence centered at the TSS (x-axis) versus the average over forward-stranded sequences which were shifted by −3, −2, −1, 0, 1, 2, 3 bp, and a reverse-stranded input sequence centered at the TSS (y-axis). Data shown for a random subset of loci (n = 30). Orange line: diagonal line where x and y-axis have the same value. The correlation coefficient between values on x-axis and y-axis is R = 0.94 (**c**) Absolute Pearson's *r* coefficients between Enformer predictions and observed gene expression for sets of genes with one causal SNP and all others. Causal genes determined by the Susie algorithm ('Susie-Causal'). Edges of the box indicate the 25th and 75th percentiles, and the central mark indicates the median (N1 = 183 genes fine-mapped with Susie, N2 = 6625 genes without fine-mapped variants, two-sided Wilcoxon rank-sum test, for each gene R coefficient computed using n = 839 individuals).

https://doi.org/10.1038/s41588-023-01524-6



Extended Data Fig. 2 | Performance of the shallow CNN model. (**a**) Density plot of observed population-average expression of test set genes (n = 3,401 genes) in cerebral cortex versus simple CNN's predicted gene expression from the Reference sequences. This plot only displays genes which could be assigned to Enformer's test set. Colors depict local density. (**b**) Y-axis shows Pearson's *r* correlation coefficients between observed expression values and a simple CNN's predicted values per individual. X-axis shows the negative log10 p-value computed with a gene-specific Null model (one-sided T-test, n = 50 independent samples per gene; Supplementary Method). The color represents the predicted mean expression. Red dashed line indicates FDRBH = 0.05.

nature portfolio

Corresponding author(s): Sara Mostafavi
Maria Chikina

Last updated by author(s): 07/28/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software used in data collection.
Data analysis	RNA-seq data from ROS and MAP cohorts were preprocessed using edgeR (3.28.0) and limma (3.17). WGS data was preprocessed using Picard LiftOver (3.0.0) and Eagle (2.4.1). We used Enformer software (https://github.com/deepmind/deepmind-research/tree/master/enformer, Version 1.0.0). Our software and scripts can be found at: https://github.com/mostafavilabuw/EnformerAssessment (DOI: https://zenodo.org/record/8274879)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Methods and Supplementary Methods describe data and availability as required.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We consider sex as a biological variable that was adjusted for in RNA-seq data. We do not consider gender.
Reporting on race, ethnicity, or other socially relevant groupings	We do not consider race or ethnicity in our analysis.
Population characteristics	Population characteristics include: age and sex
Recruitment	This study did not collect any new data. The procedure for recruiting and collecting ROS and MAP subjects are cited and described previously.
Ethics oversight	ROS and MAP studies were approved by Institutional Review Board of RUSH University Medical Center.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size of 839 independent subject and 6824 genes were used in this study. The number of subjects was informed by previous eQTL studies as cited in text.
Data exclusions	No samples where excluded from the analysis in this study.
Replication	We performed several sensitivity analyses to ensure replication of results, these include: 1) forward and backward passes in obtaining prediction results, 2) model ensembling 3) comparing model performance according to different ways of summarizing the output tracks.
Randomization	Data was randomly split to train/test/validation folds.
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging