How far are we from personalized gene expression prediction using sequence-to-expression deep neural networks?

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

[1] Paul G. Allen School of Computer Science and Engineering, University of Washington, WA, USA, 98195
[2] Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, USA, 60612
[3] 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, 10032

[4] Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA, 16354 .

[5] Canadian Institute for Advanced Research, Toronto, ON, Canada, MG5 1ZB

[*] These authors contributed equally
[$] Co-senior authorship, to whom the correspondence should be addressed to: mchikina@gmail.com, saramos@cs.washington.edu

Abstract

Deep learning (DL) methods accurately predict various functional properties from genomic DNA, including gene expression, promising to serve as an important tool in interpreting the full spectrum of genetic variations in personal genomes. However, systematic out-of-sample benchmarking is needed to assess the gap in their utility as personalized DNA interpreters. Using paired Whole Genome Sequencing and gene expression data we evaluate DL sequence-to-expression models, identifying their critical failure to make correct predictions on a substantial number of genomic loci, highlighting the limits of the current model training paradigm.

Main

Deep learning (DL) methods have recently become the state-of-the-art in a variety of regulatory genomic tasks[1–6]. By adapting convolutional neural networks (CNNs), these models take as input sub-sequences of genomic DNA and predict as outputs functional properties such as epigenomic modifications[2,7], 3D interactions[5,8], and gene expression values[1,9,10]. A key insight has been to formulate model training and evaluation such that genomic regions are treated as data points, resulting in millions of training sequences in a single Reference genome to optimize model parameters[11,12]. This training approach yields accurate predictions of context specific functional profiles from the Reference genome sub-sequences. However, an extraordinary promise of these sequence-based models is to make out-of-sample predictions for sequence alternatives across individuals at any locus, each presenting a unique combination of genetic variants – a combinatorial space that is simply not feasible to evaluate with current experimental assays[4,6]. Using natural genetic variation in expression quantitative trait loci (eQTL) studies, previous analysis has shown some promise in the ability of these models to make such out-of-sample predictions[1,13]. On the other hand, benchmarking efforts using massively parallel reporter assay (MPRA) experiments has shown varying degrees of accuracy[1,14]. However, experimental noise and context differences between in-vivo training datasets and MPRA in-vitro experiments pose significant challenges. Here, we use paired Whole Genome Sequencing (WGS) and RNA-sequencing from a cohort of 839 individuals to systematically benchmark the utility of existing sequence-to-expression DL models trained on the Reference genome in *in-vivo* out-of-sample prediction of gene expression across a diverse set of sequence alternatives.

We focus our evaluation on Enformer[1], the top-performing sequence-to-expression CNN-based model that follows the current standard training paradigm, utilizing genomic regions across a single Reference genome to learn the relevant sequence patterns for gene expression, TF binding, and histone modifications across hundreds of cell types in a multitask framework (**Fig. 1A**). First, we evaluated Enformer's prediction of population-average gene expression in the cerebral cortex from its test regions in the Reference Genome (Methods). We observe a Pearson correlation R=0.58 (**Fig. 1B, S1**, R=0.51 for Enformer's test set; Methods) between observed and predicted gene expression across genes which is consistent with previous reports[1,15].

Highlighting the ability of this model to perform out-of-sample predictions in selective cases, we examine the case of a gene related to DNA replication: *DDX11*. *DDX11*'s expression in the cortex is highly heritable according to standard linear analysis (heritability r^2=0.8), and this can

be attributed to a single causal single-nucleotide variant (SNV) using statistical fine-mapping[13]. Using WGS data, we created 839 input sequences of length 196,608bp centered at TSS, one per individual for the gene (**Fig. 1C**). Each individual's input sequence contains all their observed SNVs (**Fig. S2**). We observed accurate predictions of gene expression across individuals for the *DDX11* gene (*Pearson R*=0.85, **Fig. 1D**). Further, *in-silico* mutagenesis (ISM) at this locus showed that Enformer utilizes a single causal SNV with high correlation to gene expression (eQTLs) in making its predictions (**Fig. 1E**), the same causal SNV that was identified through statistical fine-mapping with Susie[13]. Thus, at this locus, Enformer is able to identify the causal SNV amongst all those in LD, and in addition provides hypotheses about the underlying functional cause, in this case the extension of a repressive motif (**Fig. S3**).

However, the impressive out-of-sample predictions on *DDX11* proved to be the exception rather than the rule, suggesting that current models do not consistently generalize to within-locus variation. When we compared the predicted to observed expression levels across individuals for 6,825 cortex-expressed genes that we were able to test, we found a large distribution in the *Pearson's R* (**Fig. 2A, Table S1**). While the model was able to predict gene expression with statistical significance for 598 genes ($FDR_{BH}=0.05$, Methods), the predictions were anti-correlated with the true gene expression values for 195 of these genes (33%). For example, predictions for *GSTM3* gene expression values are anti-correlated with actual observed gene expression across individuals (**Fig. 2B**). The results are similar when we select the best output track that matches the cerebral cortex ("CAGE, cortex, adult") or fine-tune the predictions using an elastic net model (**Fig. S4**, Methods). As well, model ensembling, whereby we averaged over model predictions on shifted sub-sequences and reverse and forward strands, slightly improved the performance in some cases, but our analysis indicates that the sign of the prediction would not be impacted by this type of ensembling approach (**Fig. S5**). When we focused this analysis on 184 genes with known causal SNVs according to previous eQTL analysis[13], again we observed that while Enformer can make significant predictions, the predicted expression levels are anti-correlated for 80 (43%) of these genes (**Fig. S6A, Table S1**). Overall, these results imply that the model fails to correctly attribute the variants' direction of effect (*i.e.,* whether a given variant decreases or increases gene expression level).

We then compared the prediction performance with a popular linear approach called PrediXcan[16]. PrediXcan constructs an elastic net model per 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. For a fair comparison, we used a "pre-trained" prediXcan model on GTEx data[16] and applied it to ROSMAP samples. Hence neither Enformer nor PrediXcan have seen the ROSMAP samples prior to their application. For the 1,570 genes where an elastic net model was available, performance of Enformer is substantially lower than PrediXcan (Mean R Enformer = 0.02, Mean R PrediXcan = 0.26 **Fig. 2C, Table S1**). Further, the Elastic net model did not have the same challenge with mis-prediction of the direction of SNV effect. We note that Enformer predictions were evaluated against eQTLs in the original study using SLDP regression demonstrating improved performance over competing models in terms of z-score. Our results are not in contradiction with these findings. The SLDP approach computes the association of effects

genome-wide; taking a conservative estimate for the degrees of freedom to be the number of independent LD blocks (1,361[17]) a z-score of 7 would correspond to an R^2 of 0.034.

Interestingly, when we compared the absolute Pearson R values across genes between Enformer and PrediXcan, we observed a substantial correlation (R=0.58, **Fig. S6B**), implying that genes whose expression values from genotype across subjects can be predicted well by PrediXcan overlap the set of genes where Enformer assesses a relationship between SNVs and expression. However, Enformer is not able to determine the sign of SNV effects accurately (hence a very low mean R value between observed and predicted gene expression of 0.02).

To explore the causes for the negative correlation between Enformer predictions and the observed gene expression values we applied two explainable AI (XAI) techniques on all genes with a significant correlation to observed expression values (abs(R)>0.2, **Fig. 2A**): ISM and gradients (Grad) [18–20]. Both XAI methods decompose the nonlinear neural network into a linear function whose weights approximate the effect and direction of every SNV to the prediction (Methods). While there was a moderate correlation between attributions computed with Grad and ISM (mean *Pearson R* = 0.28, **Fig. S7**), we found that linear decomposition with ISM generated a better approximation of Enformer's predictions (**Fig. S8**), and was able to accurately approximate Enformer's predictions for 95% of the examined genes ($R>0.2$, $p<10^{-8}$).

For each gene, based on its ISM attributions, we determined the main SNV driver(s) that dominate the linear approximation of the differential gene expression predictions across individuals (Methods). Across the 256 examined genes, we found that 32% have a single SNV driver, and the vast majority (85%) have five or fewer drivers (**Fig. S9, Table S2**). To understand how these driver SNVs cause mispredictions, we directly computed the SNV direction of effect by contrasting the gene expression levels across people when stratified by the SNV's genotypes (Methods), referred to as the eQTL effect size. We classified Enformer-identified driver SNVs into "supported" and "unsupported" categories based on the agreement of SNVs ISM attribution sign with the direction of effect according to the eQTL analysis. For example, *GSTM3* has two common driver SNVs and their predicted direction of effect was unsupported by the observed gene expression data (**Fig. 2D**). For all 256 inspected genes, we found that mispredicted genes had almost exclusively unsupported driver SNVs **(Figure 2E**), confirming that the selected drivers are in fact 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, we analyzed the genomic sequences around driver SNVs. We did not find any enrichment for specific sequence motifs (**Fig. S10**). 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, Fig. S11**). We looked at Grad attributions along the entire sequence (**Fig 2G top, S12**) and ISM attributions for large windows around the TSS (**Fig 2G, bottom**, Supplementary materials) and found that the area around the TSS not only contained distinguishable learned sequence motifs but also both the strongest positive and strongest negative attributions outside of apparent learned motifs. We observe that the majority of the SNVs that drive the significant positive and negative correlations to the observed expression do not fall into one of these distinguishable motifs but instead in regions of spurious attributions where training data was likely not sufficient to deduce the regulatory logic (**Fig. S13, Table S3**).

In summary, our results suggest that current models trained on a single Reference genome often fail to correctly predict the direction of SNV effects because most SNVs do not fall into the major motifs that are learned from the Reference genome. Therefore, we hypothesize that training similar models on larger, personalized datasets will increase the *causal genetic* variance that these models see during training, improving their sensitivity in predicting gene expression variation in a locus-dependent manner. Indeed, models like PrediXcan that only consider SNV values and not their sequence context are mainly statistical powered to make predictions for a limited set of genes[21]. A major factor that contributes to their lack of statistical power is their inability to learn and generalize biological *mechanisms* like the sequence motifs that explain the principles underlying the observed correlation between genetic variation and gene expression. Revising how sequence-based NN models are trained promises to enable accurate *in-silico* predictions about the cellular impact of genetic variants across the entire genome, because in contrast they can learn and generalize the regulatory logic of gene expression as encoded in the genomic sequence.

Figure Legends

Figure 1. Evaluation of Enformer across genomic regions and select loci. (A) Schematic of the training approach implemented by Enformer and other sequence-based CNN models. 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 TF binding (ChIP-Seq). (B) Population-average gene expression levels in cerebral cortex (averaged in ROSMAP samples, n=839) for expressed genes (n=13,397) shown on the x-axis. Enformer's prediction of gene expression levels for cortex based on the Reference genome sequences centered at TSS of each gene (196Kb) is shown on the y-axis. Enformer's output tracks are fine-tuned with an elastic net model (see Methods). (C) Schematic of the per-locus evaluation strategy. Personal genomes are constructed for each individual by inserting their observed SNVs into the Reference genome. The personalized sequences centered at the TSS of gene DDX11 are used as input to Enformer. (D) Prediction of cortex gene expression levels for individuals in the ROSMAP cohort. Each dot represents an individual. Output of Enformer is fine-tuned using an elastic net model. E) In-silico mutagenesis (ISM) values for all SNVs which occur at least once in 839 genomes within 98Kb of DDX11 TSS. SNVs are colored by minor allele frequency (MAF). The border of the "driver" SNV is shown in red and its size is proportional to its impact on the linear approximation (Supplementary Methods).

Figure 2. Evaluation of Enformer on prediction of gene expression across individuals. (A) Y-axis shows the Pearson R coefficient between observed expression values and Enformer's predicted values per-gene. X-axis shows the negative log10 p-value, computed using a gene-specific null model (Supplementary Method). The color represents the predicted mean expression using the most relevant Enformer output track ("CAGE, adult, brain"). Red dashed line indicates FDRBH=0.05. (B) Prediction of cortex gene expression levels ("CAGE, adult, brain" track) in the ROSMAP cohort (n=839) for the GSTM3 gene, x-axis shows the observed gene expression values. (C) Pearson R coefficient between PrediXcan predicted versus observed expression across 839 individuals (x-axis) versus Enformer's Pearson R values on the same sample (y-axis). Red lines indicate threshold for significance (abs(R)>0.2), darker colored dots are significant genes from panel A. Green cross represents the location of the mean across all x- and y-values. (D) ISM value (x-axis) versus eQTL effect size (y-axis) for all SNVs within the 196Kb input sequence of the GSTM3 gene. Red circles represent SNVs that drive the linear approximation to the predictions. SNVs are defined as supported or unsupported based on the

concordance with the sign of the eQTL effect size. (E) Fraction of supported driver SNVs per gene (y-axis) versus Pearson's R values between Enformer's predictions and observed expressions (x-axis). (F) Number of driver SNVs within the 1000bp window to the TSS. Main drivers are the drivers with the strongest impact on linear approximation, shown in different colors. (G) Top: Gradient attributions (grey) across the entire sequence of the GSTM3 gene with location of all SNVs and driver SNVs. Bottom: 300bp window around the TSS with ISM attributions normalized by the estimated standard deviation across the entire sequence. Most significant connected motifs are framed in red. Main driver shown as magenta triangle.

Software and intermediate results

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

Accession Codes

Genotype, RNA-seq, and DNAm data for the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) samples are available from the Synapse AMP-AD Data Portal https://www.synapse.org/#!Synapse:syn2580853/discussion/default as well as RADC Research Resource Sharing Hub at www.radc.rush.edu.

Acknowledgements

We thank David R. Kelley for helpful comments on this manuscript. We thank the participants of ROS and MAP for their essential contributions and gift to this project. This work has been supported by many different NIH grants: P30AG10161, P30AG72975, R01AG15819, R01AG17917, U01 AG046152, U01AG61356, R01 AG057911, R01 AG061798, RC2AG036547, U01 AG058589 and U01 AG072572.

Contributions

Conceived the study: SM, MC. Study design: SM, AS1, MC. Data generation and quality control analyses: BN, AS2, CG, PLD, ST, DAB. Analyses and interpretation: AS1, AS2, BN, SM, MC. Wrote the initial draft: SM, AS1, BN. Read and provided comments on the manuscript: MC, BN, AS2, PLD, CG, ST, DAB.

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. 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).
8. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* **17**, 1111–1117 (2020).
9. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **31**, 107663 (2020).
10. 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).
11. 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).

12. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).

13. 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).

14. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).

15. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *bioRxiv* 2022.09.15.508087 (2022) doi:10.1101/2022.09.15.508087.

16. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

17. MacDonald, J. W., Harrison, T., Bammler, T. K., Mancuso, N. & Lindström, S. An updated map of GRCh38 linkage disequilibrium blocks based on European ancestry data. *bioRxiv* 2022.03.04.483057 (2022) doi:10.1101/2022.03.04.483057.

18. 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.* (2022) doi:10.1038/s41576-022-00532-2.

19. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

20. 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, 06--11 Aug 2017).

21. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. *bioRxiv* 2022.05.07.491045 (2022) doi:10.1101/2022.05.07.491045.

Figure 1

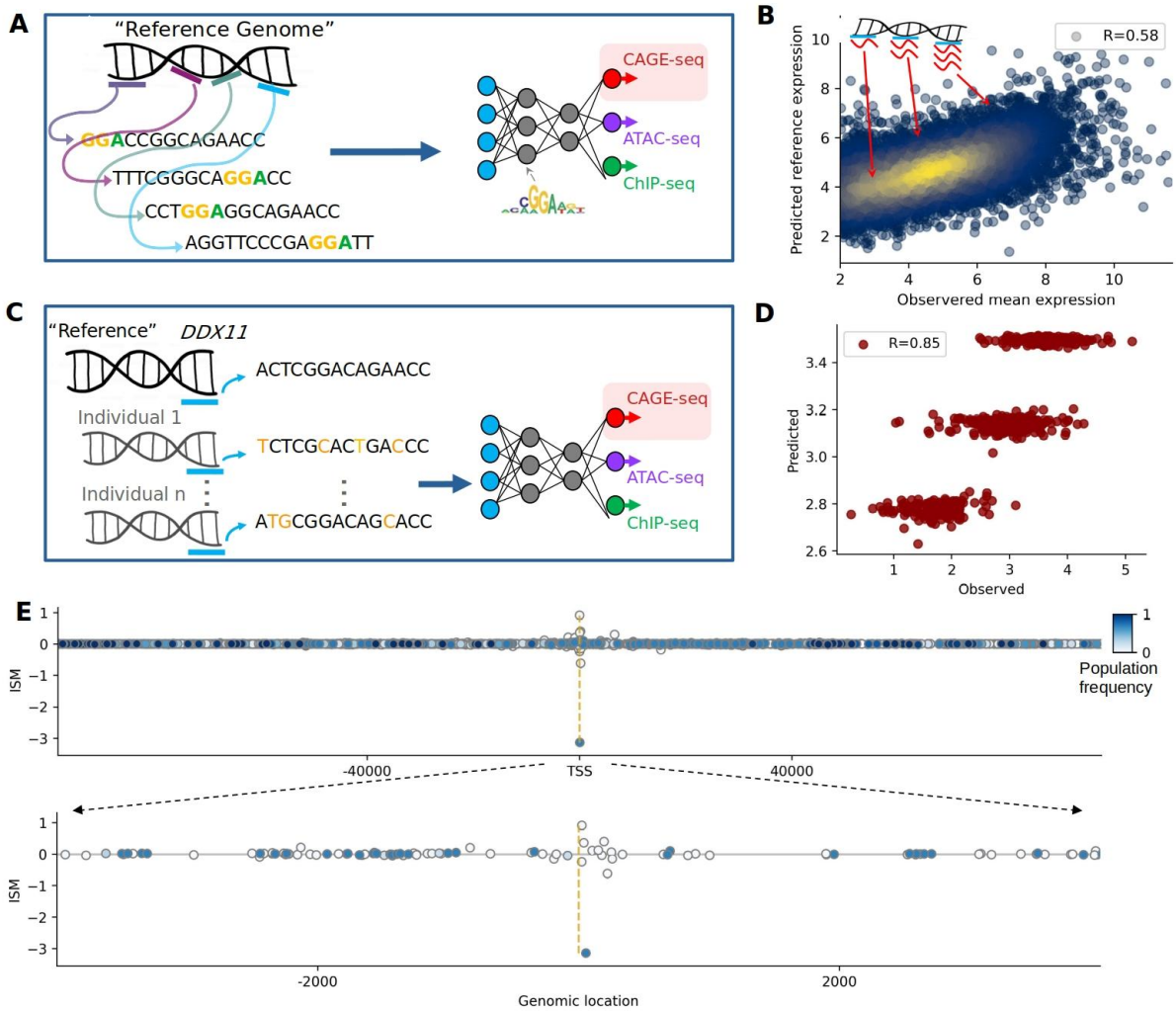

Figure 1. Evaluation of Enformer across genomic regions and select loci. (A) Schematic of the training approach implemented by Enformer and other sequence-based CNN models. Different genomic regions from the Reference genome are treated as data points. Genomic DNA underlying a given region is the input to the model, and the model learns to predict various functional properties including gene expression (CAGE-seq), chromatin accessibility (ATAC-Seq), or TF binding (ChIP-Seq). (B) Population-average gene expression levels in cerebral cortex (averaged in ROSMAP samples, n=839) for expressed genes (n=13,397) shown on the x-axis. Enformer's prediction of gene expression levels for cortex based on the Reference genome sequences centered at TSS of each gene (196Kb) is shown on the y-axis. Enformer's output tracks are fine-tuned with an elastic net model (see Methods). (C) Schematic of the per-locus evaluation strategy. Personal genomes are constructed for each individual by inserting their observed SNVs into the Reference genome. The personalized sequences centered at the TSS of gene *DDX11* are used as input to Enformer. (D) Prediction of cortex gene expression levels for individuals in the ROSMAP cohort. Each dot represents an individual. Output of Enformer is fine-tuned using an elastic net model. E) *In-silico* mutagenesis (ISM) values for all SNVs which occur at least once in 839 genomes within 98Kb of *DDX11* TSS. SNVs are colored by minor allele frequency (MAF). The border of the "driver" SNV is shown in red and its size is proportional to its impact on the linear approximation (Supplementary Methods).

Figure 2

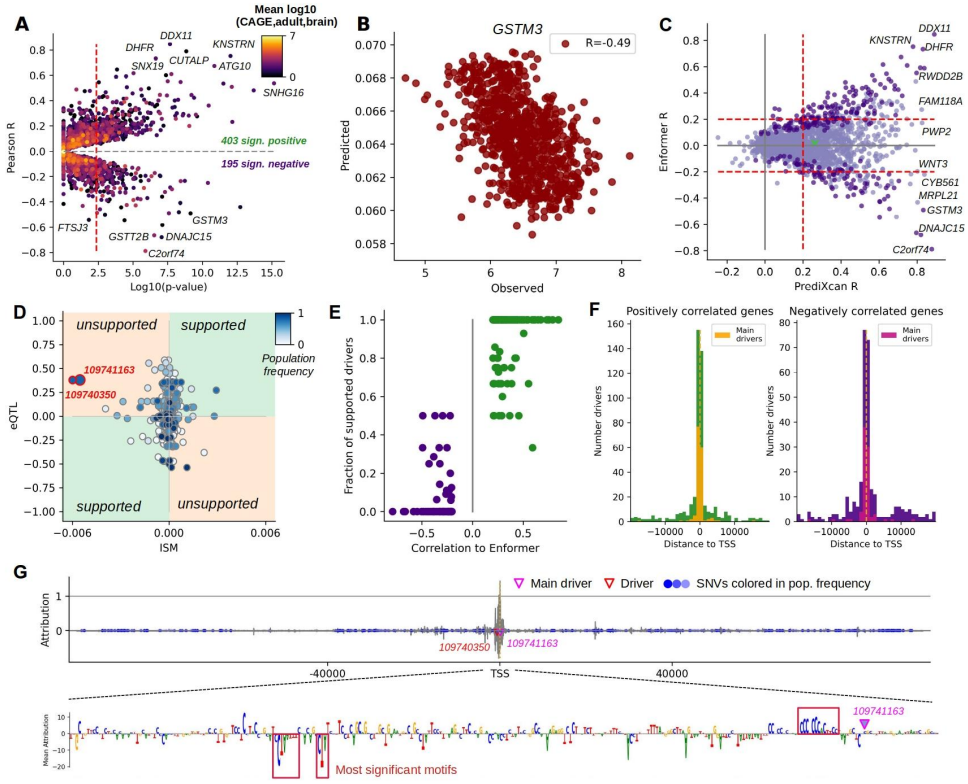

Figure 2. Evaluation of Enformer on prediction of gene expression across individuals. (A) Y-axis shows the Pearson R coefficient between observed expression values and Enformer's predicted values per-gene. X-axis shows the negative log10 p-value, computed using a gene-specific null model (Supplementary Method). The color represents the predicted mean expression using the most relevant Enformer output track ("CAGE, adult, brain"). Red dashed line indicates FDR_{BH}=0.05. (B) Prediction of cortex gene expression levels ("CAGE, adult, brain" track) in the ROSMAP cohort (n=839) for the *GSTM3* gene, x-axis shows the observed gene expression values. (C) Pearson R coefficient between PrediXcan predicted versus observed expression across 839 individuals (x-axis) versus Enformer's Pearson R values on the same sample (y-axis). Red lines indicate threshold for significance (abs(R)>0.2), darker colored dots are significant genes from panel A. Green cross represents the location of the mean across all x- and y-values. (D) ISM value (x-axis) versus eQTL effect size (y-axis) for all SNVs within the 196Kb input sequence of the *GSTM3* gene. Red circles represent SNVs that drive the linear approximation to the predictions. SNVs are defined as supported or unsupported based on the concordance with the sign of the eQTL effect size. (E) Fraction of supported driver SNVs per gene (y-axis) versus Pearson's R values between Enformer's predictions and observed expressions (x-axis). (F) Number of driver SNVs within the 1000bp window to the TSS. Main drivers are the drivers with the strongest impact on linear approximation, shown in different colors. (G) Top: Gradient attributions (grey) across the entire sequence of the *GSTM3* gene with location of all SNVs and driver SNVs. Bottom: 300bp window around the TSS with ISM attributions normalized by the estimated standard deviation across the entire sequence. Most significant connected motifs are framed in red. Main driver shown as magenta triangle.