Enformer

Deep Learning Reading Group, Dec 2nd, Eduarda Vaz

ARTICLES
https://doi.org/10.1038/s41592-021-01252-x

nature methods

Check for updates

OPEN

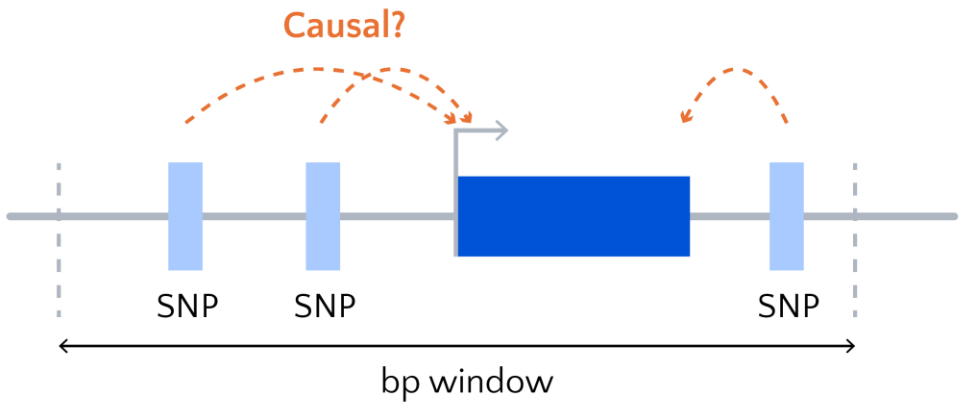
Effective gene expression prediction from sequence by integrating long-range interactions

Žiga Avsec[1]✉, Vikram Agarwal[2,4], Daniel Visentin[1,4], Joseph R. Ledsam[1,3], Agnieszka Grabska-Barwinska[1], Kyle R. Taylor[1], Yannis Assael[1], John Jumper[1], Pushmeet Kohli[1]✉ and David R. Kelley[2]✉

How noncoding DNA determines gene expression in different cell types is a major unsolved problem, and critical downstream applications in human genetics depend on improved solutions. Here, we report substantially improved gene expression prediction accuracy from DNA sequences through the use of a deep learning architecture, called Enformer, that is able to integrate information from long-range interactions (up to 100 kb away) in the genome. This improvement yielded more accurate variant effect predictions on gene expression for both natural genetic variants and saturation mutagenesis measured by massively parallel reporter assays. Furthermore, Enformer learned to predict enhancer–promoter interactions directly from the DNA sequence competitively with methods that take direct experimental data as input. We expect that these advances will enable more effective fine-mapping of human disease associations and provide a framework to interpret *cis*-regulatory evolution.

Moving beyond linear variant-based association models

Baseline approach **linear variant-based** model



$$y_{g,i} = \sum_j w_{j,g} \cdot x_{j,i} + \varepsilon_{i,g}$$

$y_{g,i}$ Expression level of gene g in individual i

$x_{j,i}$ Dosage of variant j in individual i

$w_{j,g}$ **Fit using linear model** (e.g., elastic net)

Commonly used model: **PrediXcan**

Limited to relatively common variants

⊗ No non-linear effects or interactions

It cannot generalize to unseen genes

Translational and increase information flow between distal elements

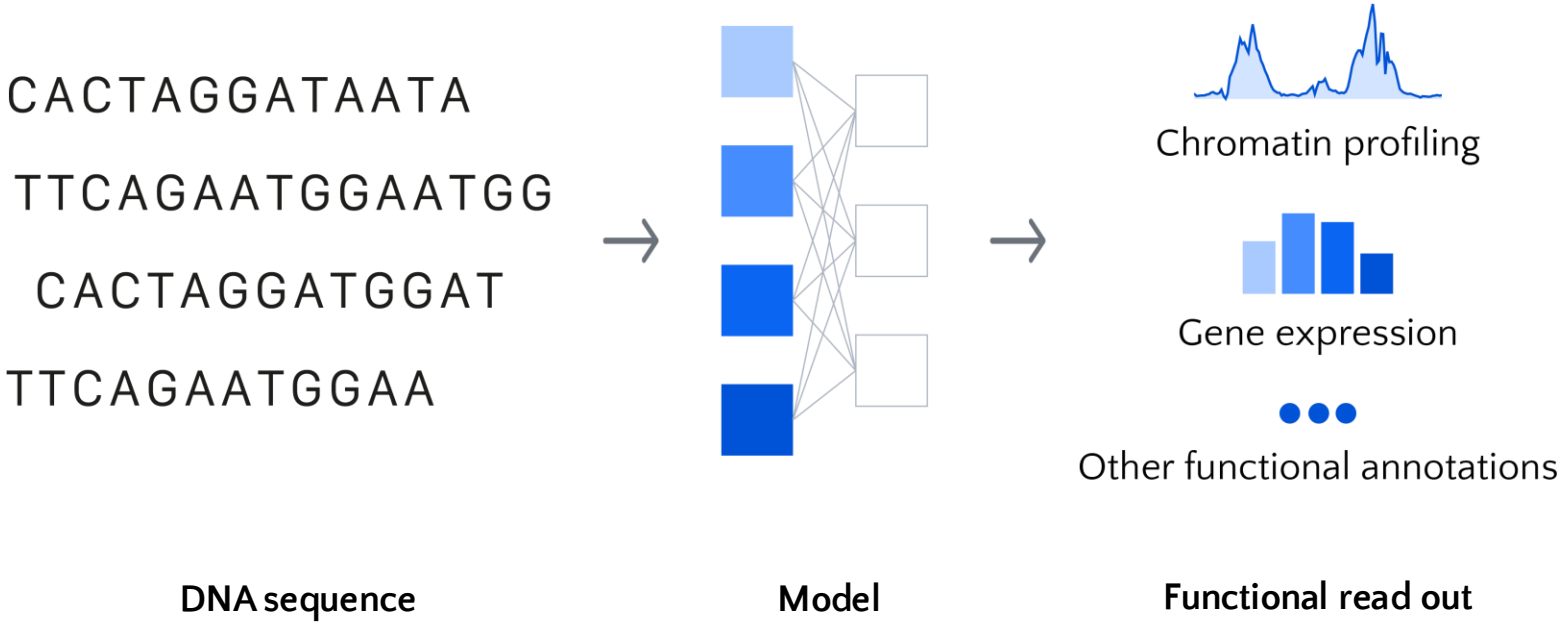
Challenge: LD confounds identification of causal variants because multiple SNPs in LD are statistically associated with a trait.

Solution: Deep learning models trained on sequence and multi-omics data can predict traits or diseases directly from DNA sequences, circumventing LD limitations by learning richer representations of genomic context.

Challenge: So far CNNs have a narrow receptive field of view.

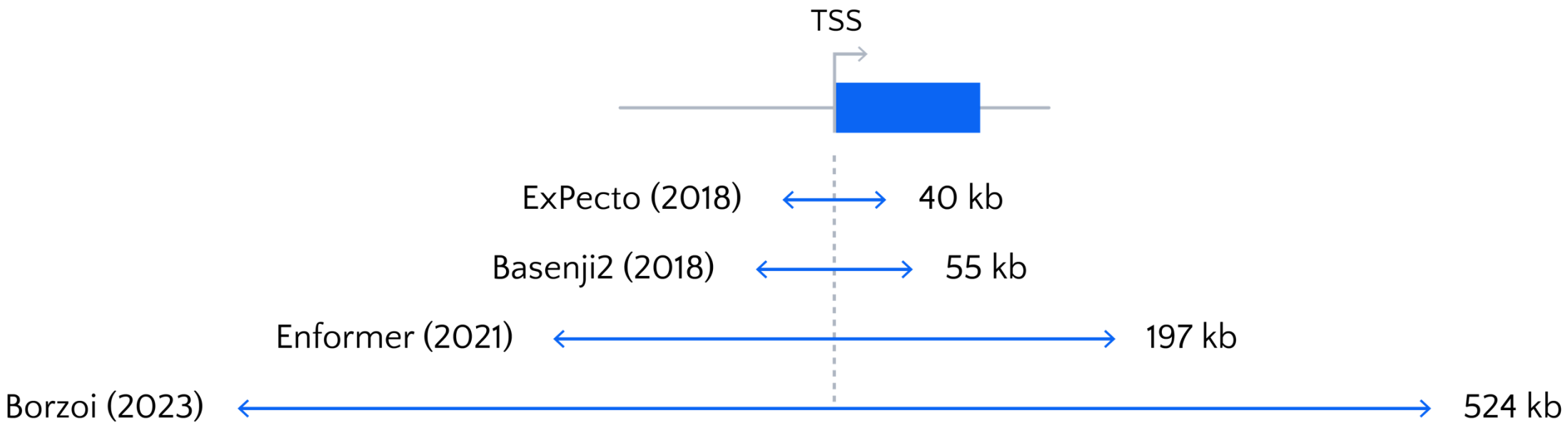
Solution: Use transformers.

The appeal of DNA sequence-to-function DL models

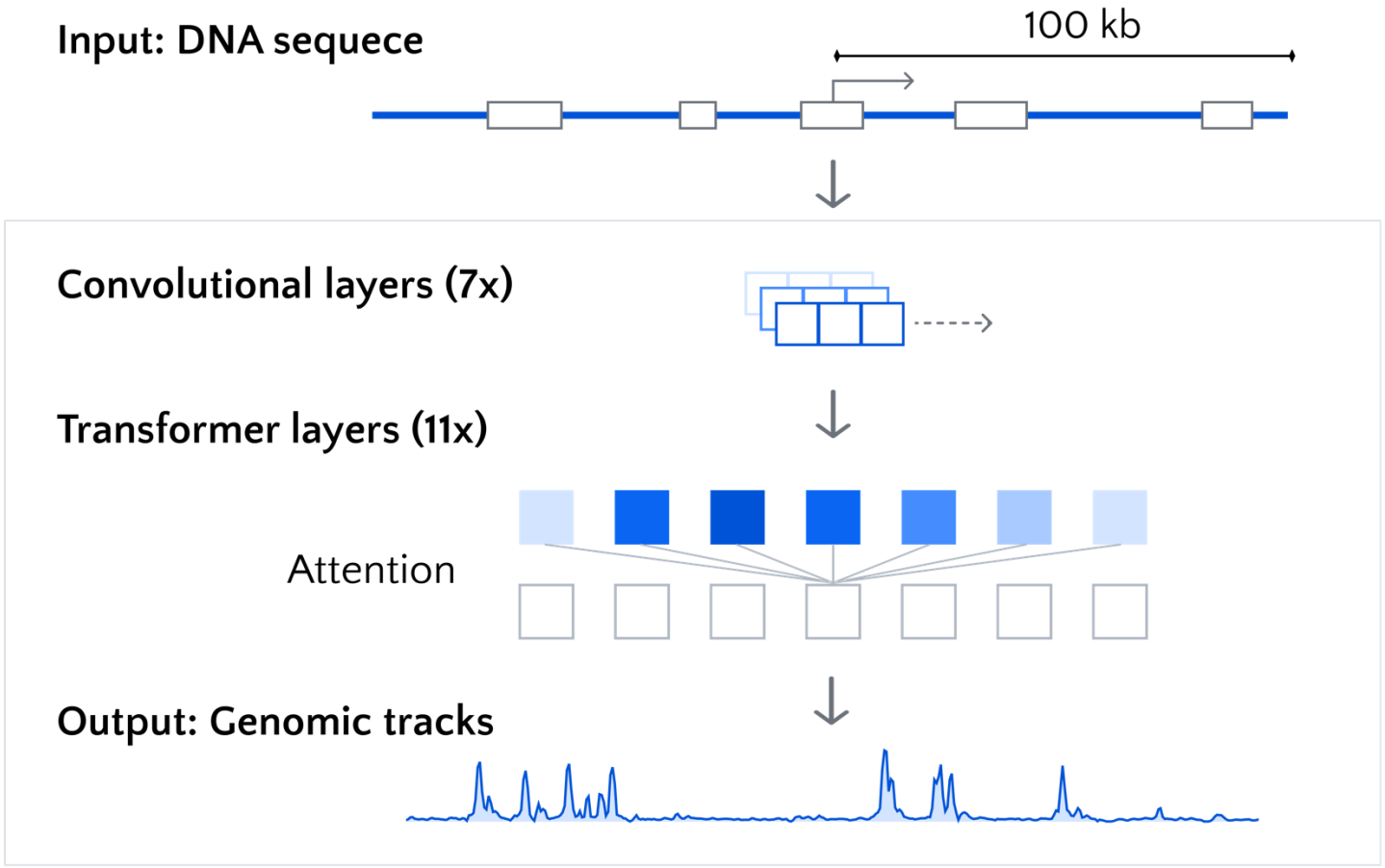


DNA sequence **Model** **Functional read out**

Sequence-to-function DL models have improved significant in recent years



A look into Enformer



Compress information reducing the input dimension for the transformer layers.

Capture long range dependencies and context between each region of the sequence.

Branching into 2 organisms-specific heads (human and mouse)

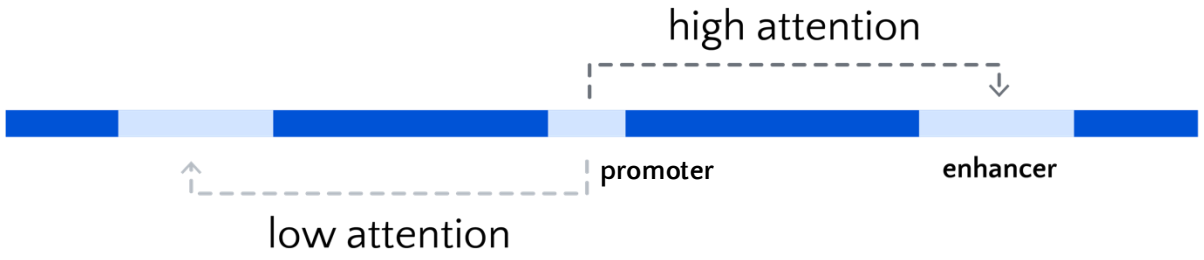
A look into Enformer

Input DNA

Convolutional layers

Input size is reduced by 120x

Attention

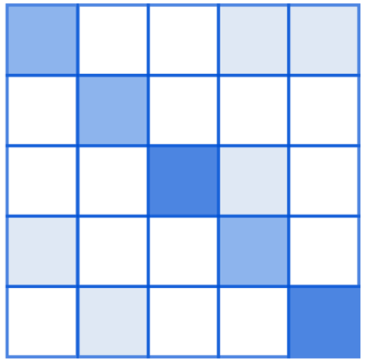


New length: 1536

Quadratic operation

1k tokens = 1M operations

1M tokens = 1T operations



A look into Enformer

Main hyperparameters

L: Number of transformer layers

C: number of channels

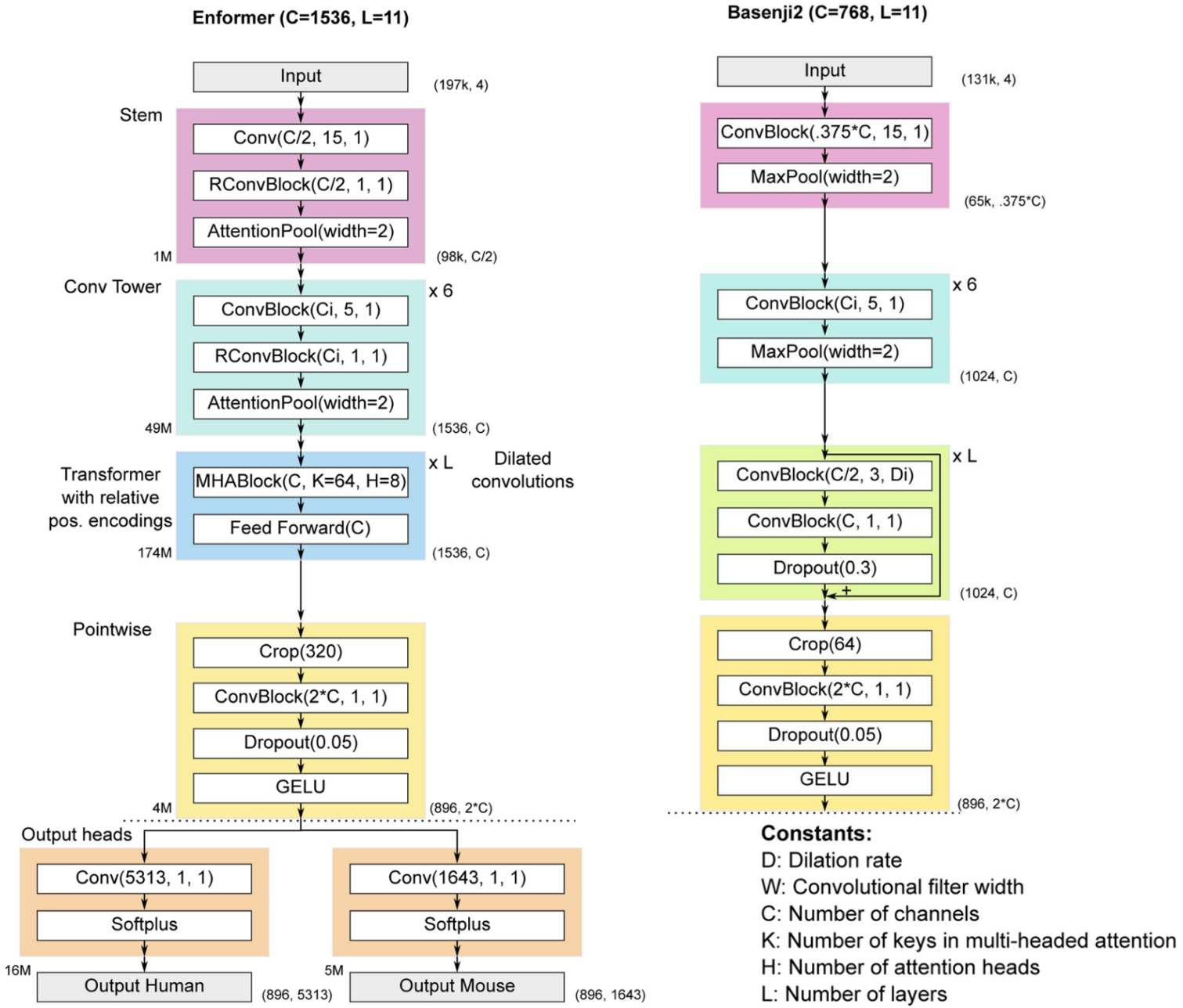
7x convolutional blocks (stem + conv tower): reduces the spatial dimension from 197k to 128bp.

Attention pooling is also an innovation that summarizes a contiguous chunk of the input seq.

Positional encoding is an embedding having both meaning and position of a token.

Attention block assigns high focus to the features that are more important (8 heads, each has different attention filters – separated set of weights).

Cropping layer trim 320 positions on each end size to avoid computing the loss on far ends.



1. Improves gene expression prediction in held-out genes

Main innovations

Longer input sequences

Transformer blocks instead of dilated convolutions

Attention pooling instead of max pooling

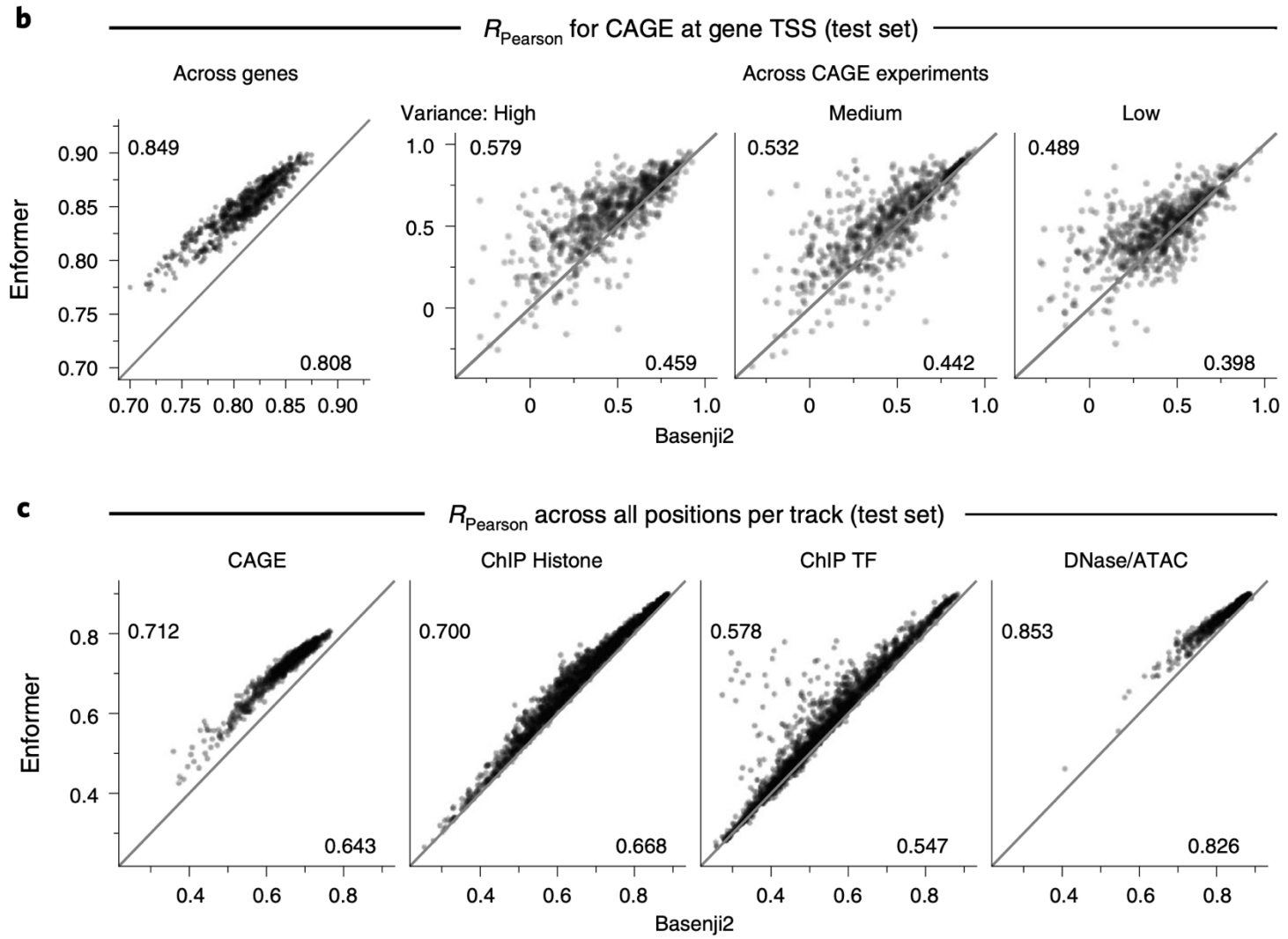
Twice as many channels

Cage measuring transcriptional activity

Histone modifications

TF binding

DNA accessibility



2. Enformer attends to cell-type-specific enhancers

What is Enformer attending to?

Computed **contribution scores** for several genes with CRISPRi-validates enhancers.

Contribution scores highlight the input sequences that are most predictive for the expression of a particular gene.

Input gradients (gradient x input)

- Tissue or cell type specific since they are computed with respect to a particular output CAGE sample.

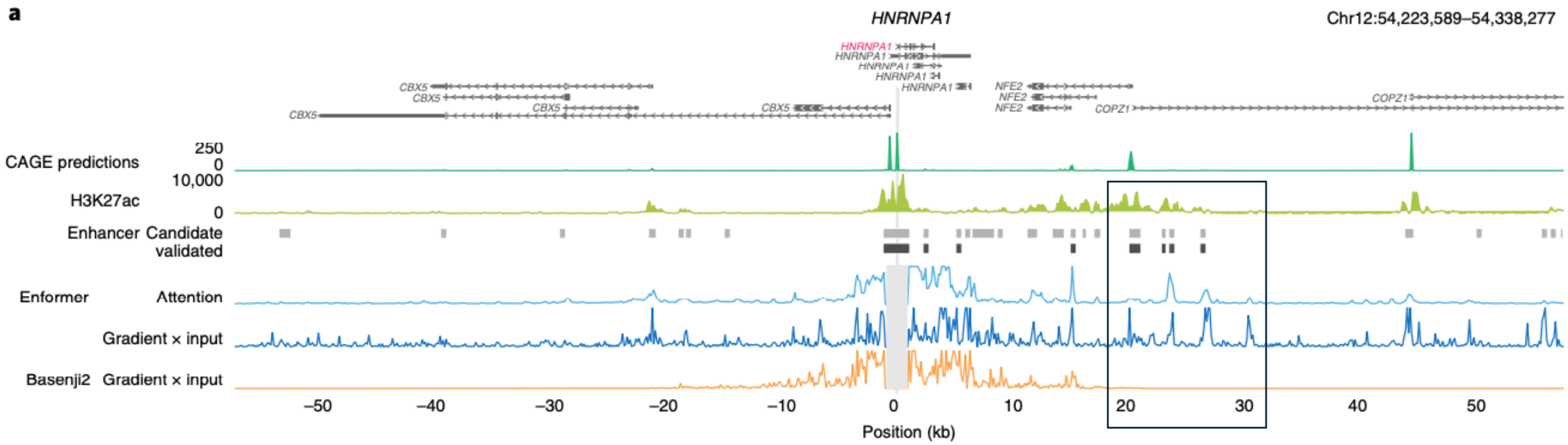
Attention weights

- Internal to the model and shared among all tissue and cell-type predictions.

In silico mutagenesis (ISM)

- Tissue or cell type specific since they are computed with respect to a particular output CAGE sample.

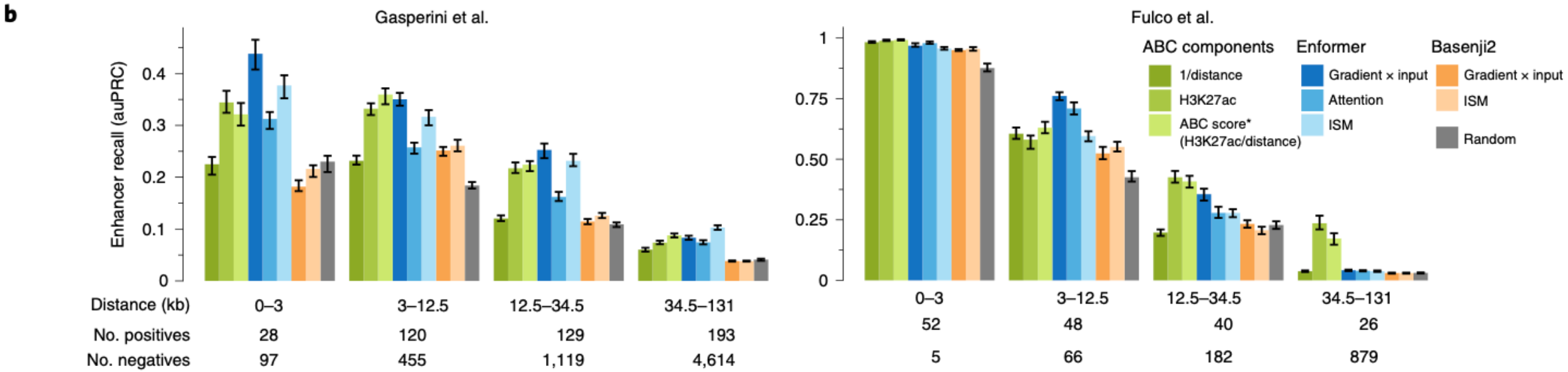
2. Enformer attends to cell-type-specific enhancers



HNRNPA1 locus for K562 cell line and correlation with histone H3 acetylated (ac) at lysine (K) at the 27th position (H3K27ac)

- H3K27ac is a hallmark of active chromatin (DNA is opened and accessible) and therefore is a marker to identify enhancers (regions of DNA that regulate gene transcription by recruiting transcription factors, RNA polymerase, etc.).

2. Enformer attends to cell-type-specific enhancers



Contribution scores across all tested enhancer-gene pairs in 2 large CRISPRi studies for K562 cell lines (>10,000 enhancers)

- Enformer prioritizes validated enhancer-gene pairs with higher accuracy than Basenji and the performance was comparable to ABC scores (enhancer prioritization method.)

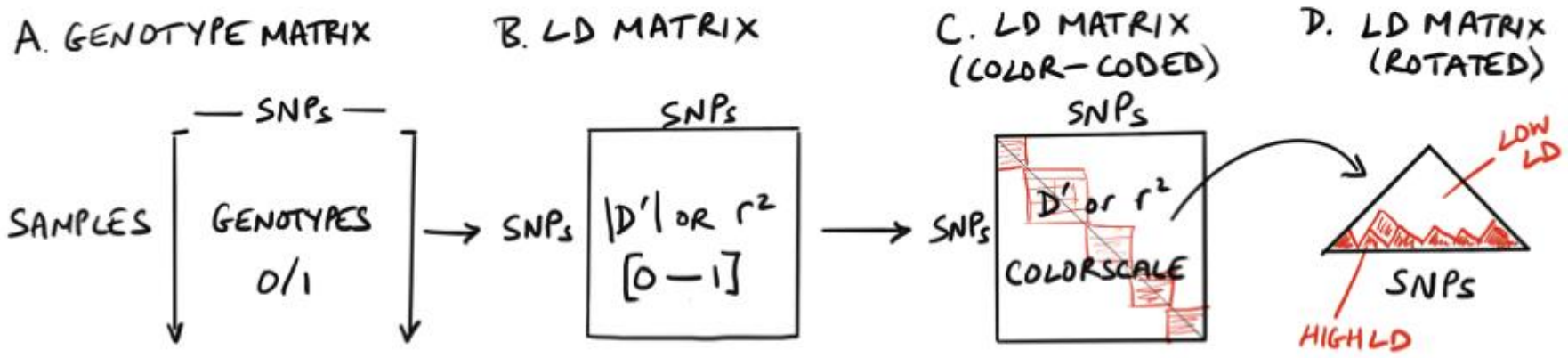
3. Enformer improves variant effect prediction on eGenes

Goal

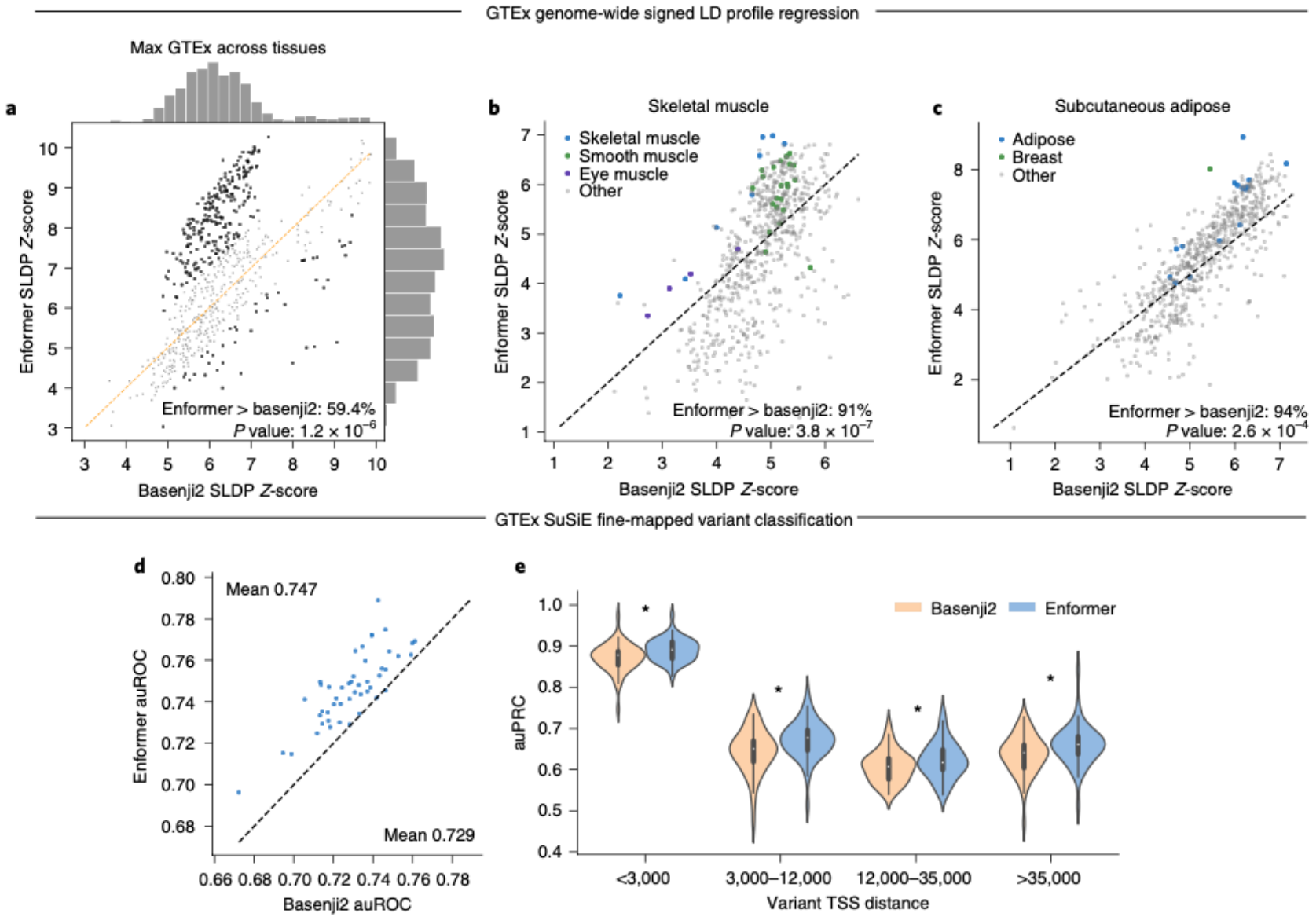
Predict the influence of genetic variants on cell-type-specific gene expression.

Challenge

Rippling effect of causal eQTL to nearby occurring genes due to linkage disequilibrium which transfers the causal eQTL effect to nearby co-occurring variants' measurements.



3. Enformer improves variant effect prediction on eGenes



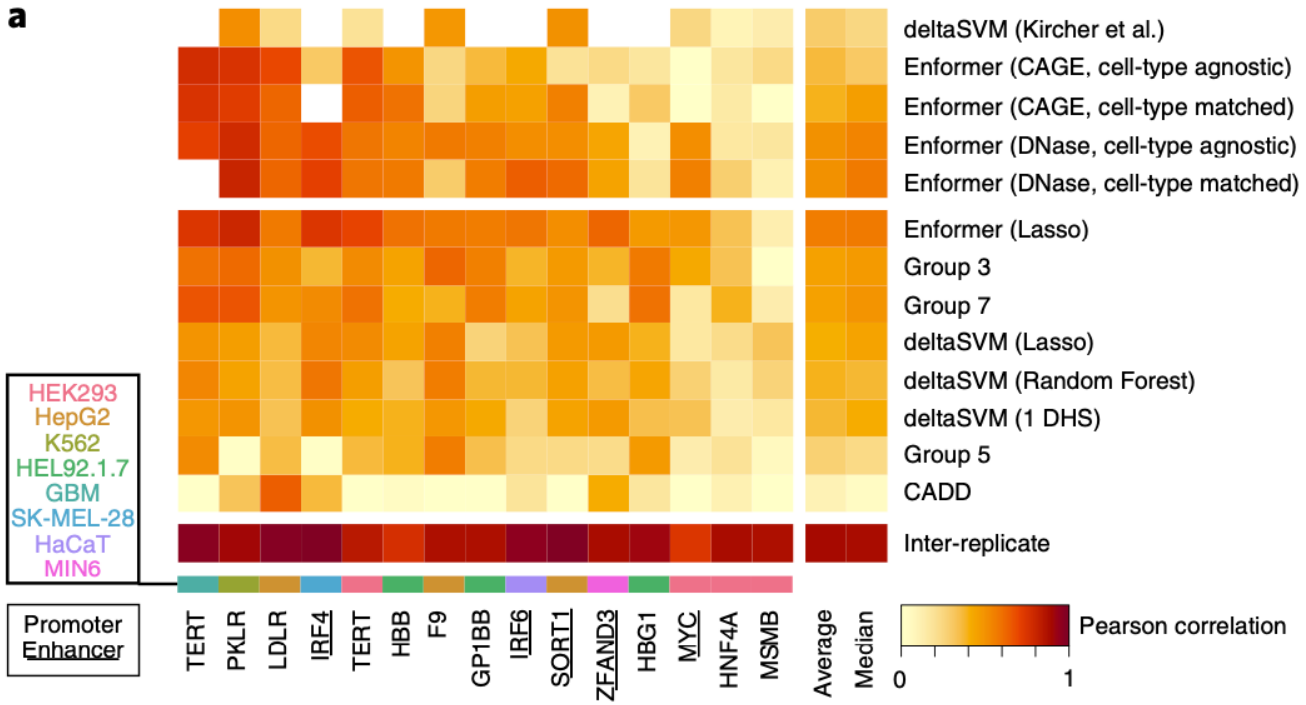
4. Enformer improves noncoding variant effect prediction as measured by saturation mutagenesis experiments

Saturation mutagenesis

Systematically introduce mutations at specific positions within a DNA sequence and measures the functional impact of each mutation.

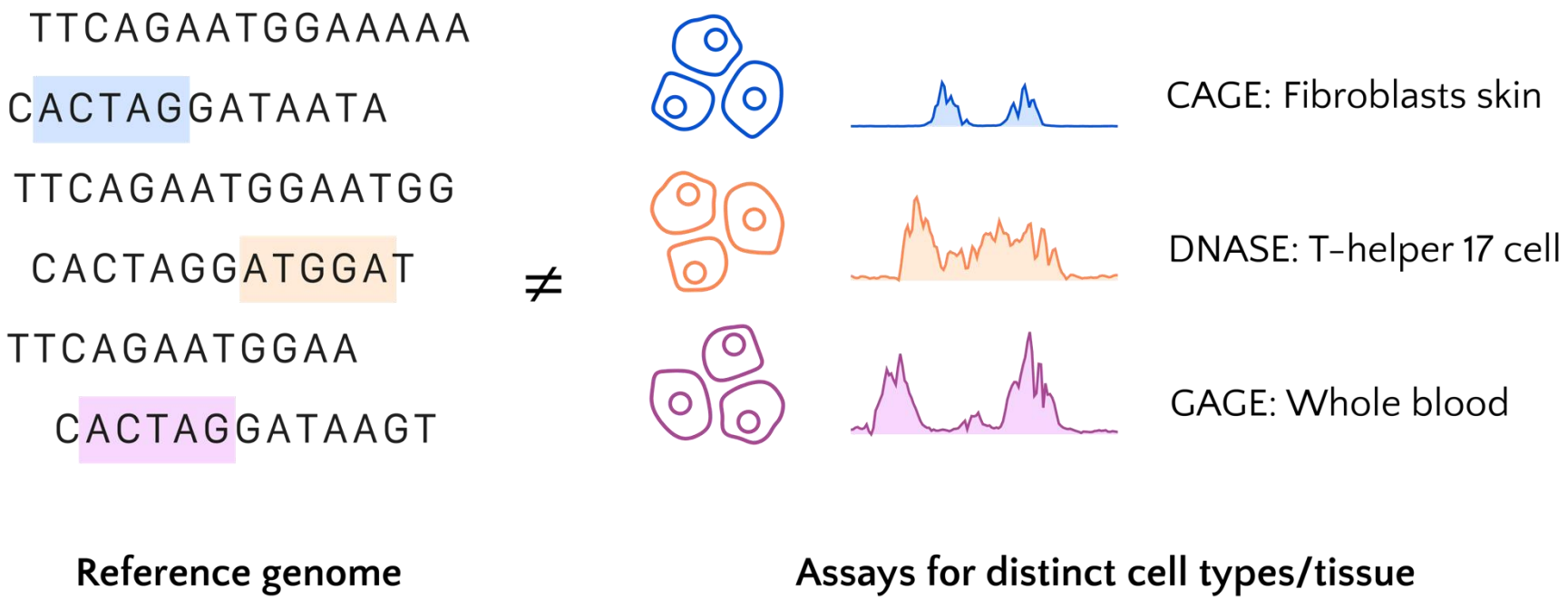
Dataset: MPRAs (massively parallel reporter assays)

Measurements of the functional effect of genetic variants through saturation mutagenesis of several enhancers and promotes in different types of cells.

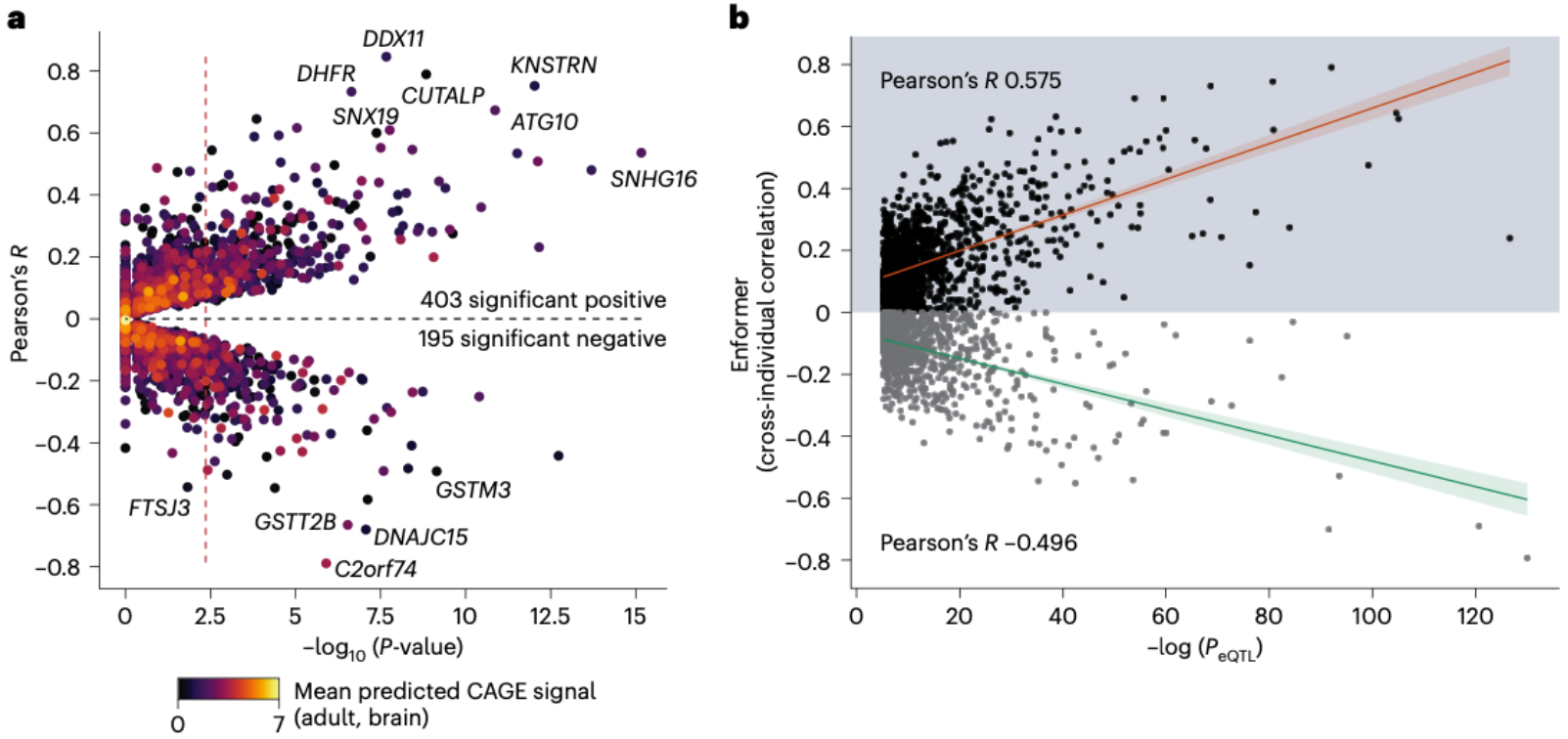


Existing sequence-to-function DL models poorly explain variation in expression across individuals

Training set: mismatch between sequence and assay/measurements and no genetic variation



Existing sequence-to-function DL models poorly explain variation in expression across individuals



a. RNA-seq from the cerebral cortex of 839 individuals for 13,397 genes (ROSMAP)

b. RNA-seq from lymphoblastoid cell lines of 421 individuals for 3,259 genes (Geuvadis)

[1] Benchmarketing of deep neural networks for predicting personal expression from DNA sequence highlights shortcomings (2023)
[2] Personal transcriptome variation poorly explained by current genomic deep learning models (2023)

Lessons learned from fine-tuning Enformer on personal genomes

From UCSF (Pollard lab) [1]

GTEx: n=**670** in whole blood expression
(300 genes, ~49kb input sequence)

From UW (Mostafavi lab) [2]

ROSMAP: n=**839** in cerebral cortex expression

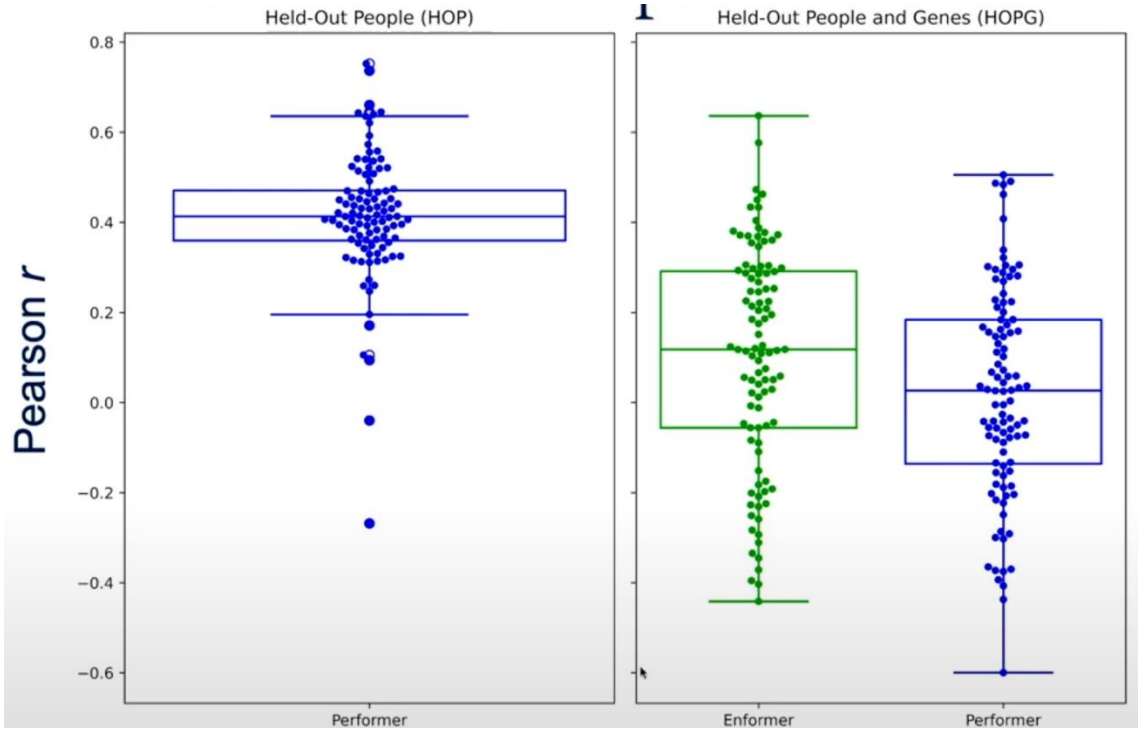
From UC Berkley (Ioannidis lab) [3]

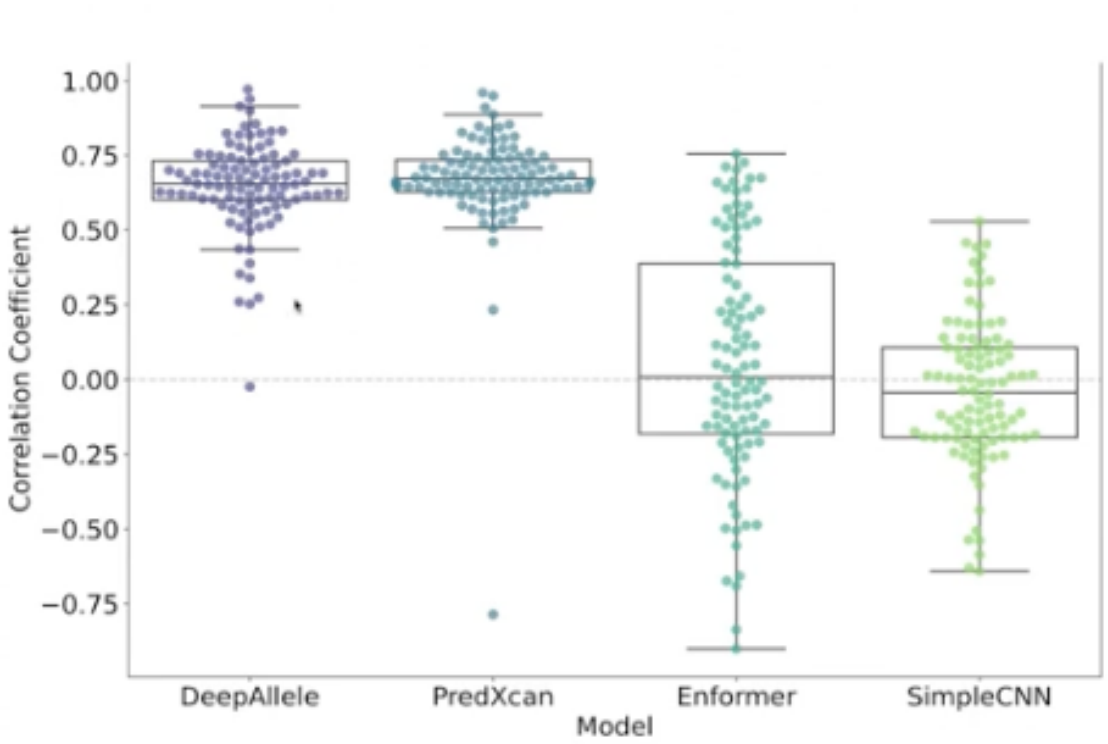
Geuvadis: n=**421** in lymphoblastoid cell lines
(200 genes, ~49kb input sequence)

1. Fine-tuning on personal genomes improves performance on held-out people achieving similar accuracy to linear models.
2. Seq2function models upweight putatively functional variants compare to linear models.
3. Fine-tuning doesn't improve performance on held-out genes compare to the original model.
4. Results are robust across studies, independent of model design.

[1] Deep Learning prediction of gene expression from personal genomes, 2024 (Pre-print)
[2] DeepAllele (Conference talk)
[3] Fine-tuning sequence-to-expression models on personal genome and transcriptome data, 2024 (Pre-print)

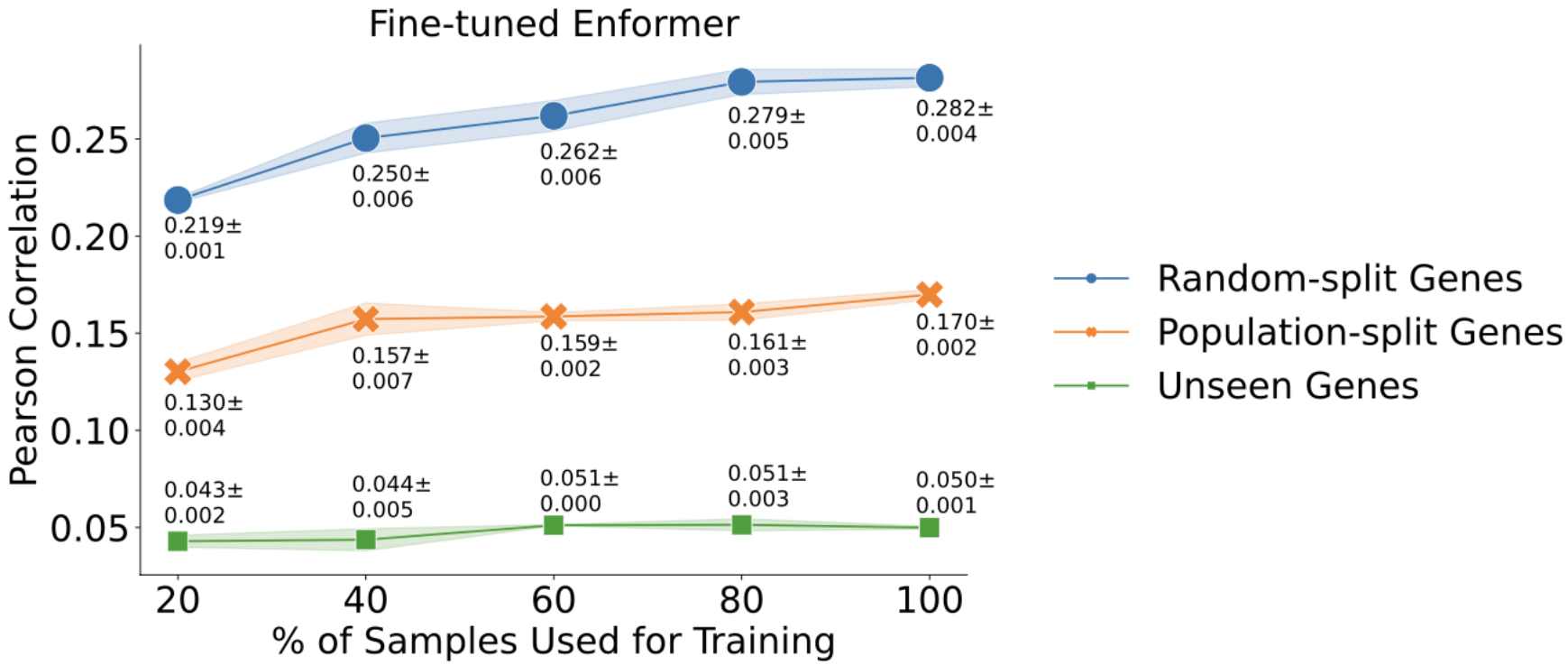
Lessons learned from fine-tuning Enformer on personal genomes





[1] Deep Learning prediction of gene expression from personal genomes, 2024 (Pre-print)
[2] DeepAllele (Conference talk)

Lessons learned from fine-tuning Enformer on personal genomes



[3] Fine-tuning sequence-to-expression models on personal genome and transcriptome data, 2024 (Pre-print)