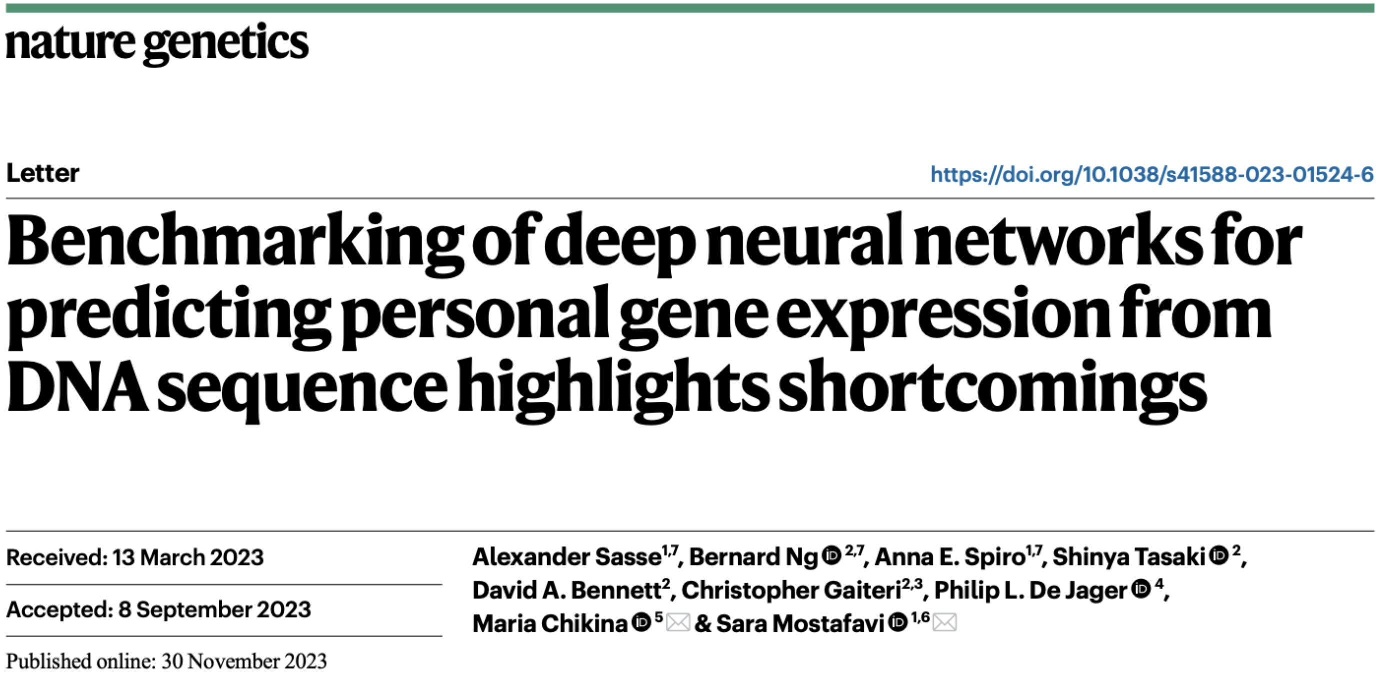
Weaknesses of Genomic LLMs

Deep Learning in Genomics Journal Club

Mahler Revsine

8 April 2025



nature genetics

Letter

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

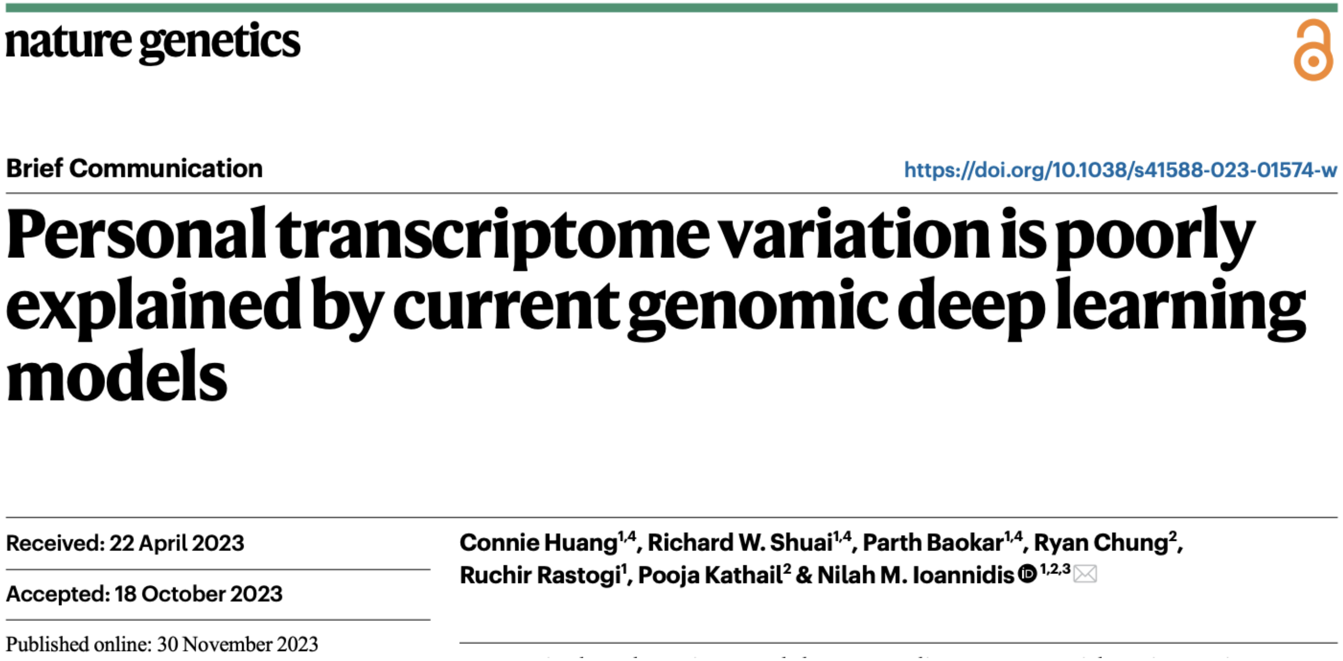
Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings

Received: 13 March 2023

Accepted: 8 September 2023

Published online: 30 November 2023

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]



nature genetics

Brief Communication

https://doi.org/10.1038/s41588-023-01574-w

Personal transcriptome variation is poorly explained by current genomic deep learning models

Received: 22 April 2023

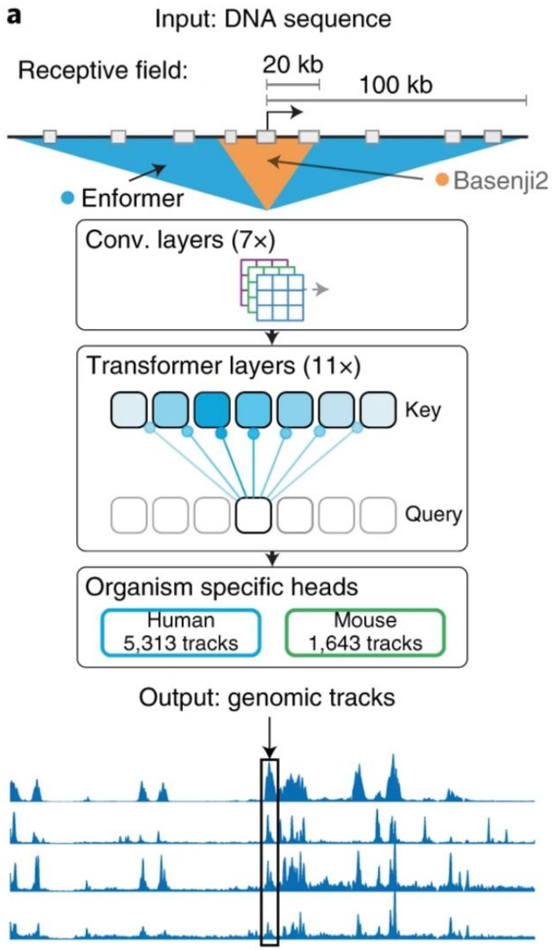
Accepted: 18 October 2023

Published online: 30 November 2023

Connie Huang[1,4], Richard W. Shuai[1,4], Parth Baokar[1,4], Ryan Chung[2], Ruchir Rastogi[1], Pooja Kathail[2] & Nilah M. Ioannidis[1,2,3]

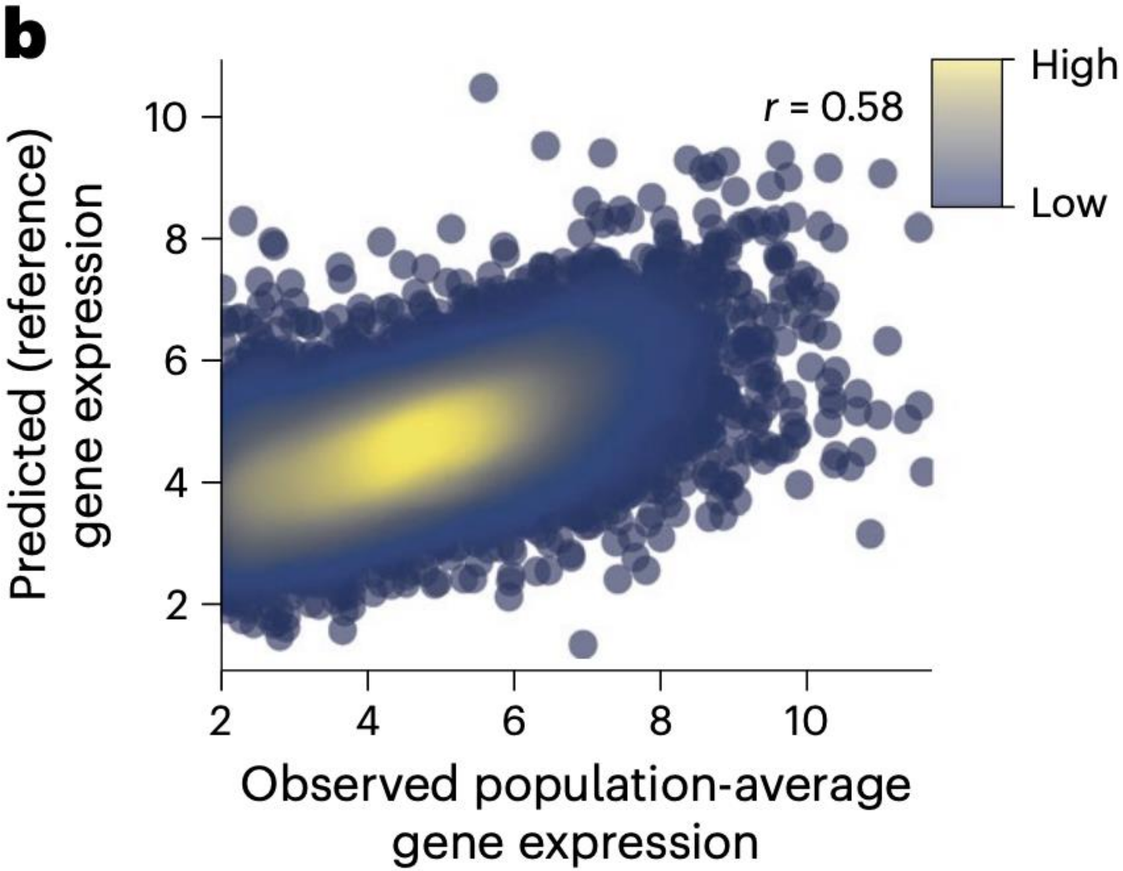
The main goal of genomic LLMs

- Sequence to function modelling
- Predict **gene expression** from corresponding DNA **sequence**
  - Or ATAC-seq, 3D structure, etc.
- Only possible at scale through **deep learning**
- Is this goal being fulfilled?



Avsec et al. *Nature Methods* (2021).

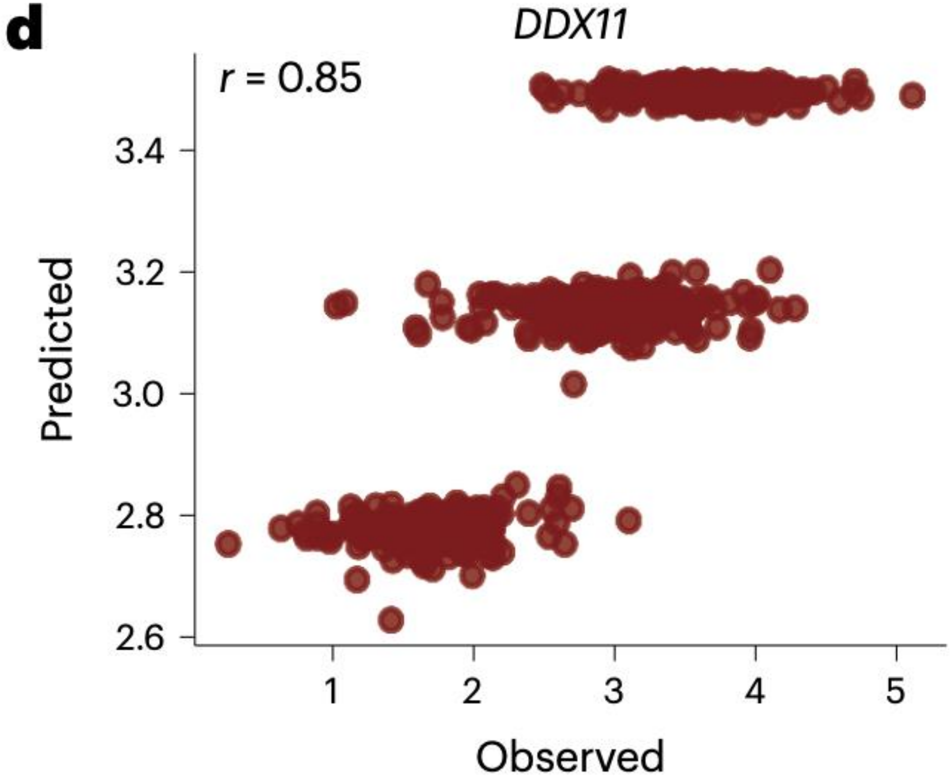
Enformer can predict average gene expression



- Example for Enformer
- 18k genes, avg. expression across 839 ROSMAP patients
- Pass in DNA sequences centered at gene TSS (transcription start site)
- Fit ElasticNet to outputs of all tracks to predict each gene
- Train on **reference** genome → predict **population-average** value

Sasse et al. *Nature Genetics* (2023).

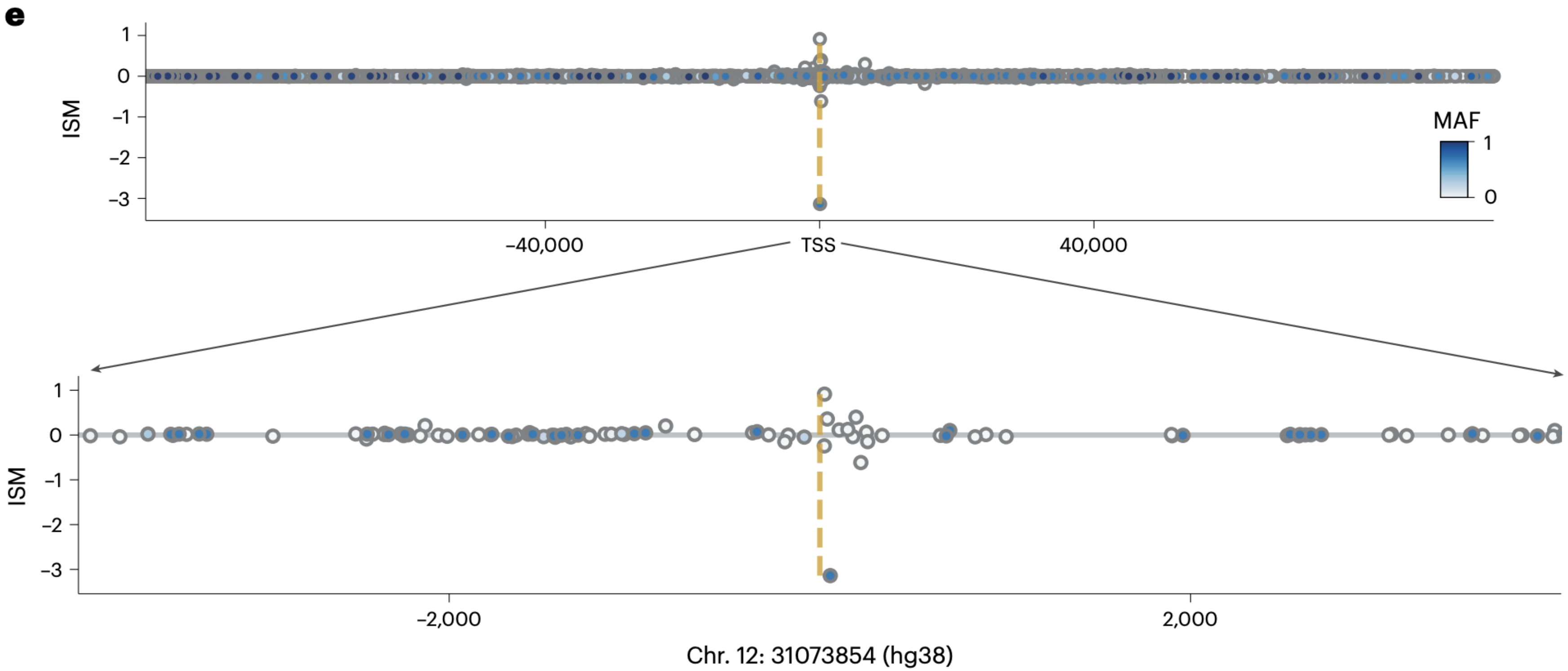
Enformer can sometimes predict eQTLs



- Observed vs. predicted gene expression **across individuals**
  - Each data point = 1 individual from ROSMAP (n=839)
  - Again fine-tuning with Elastic Net
- In this example gene, predictions capture individual-level variation
- *DDX11* is a protein-coding gene encoding a DEAD box protein

Sasse et al. *Nature Genetics* (2023).

Interpretation of predictions

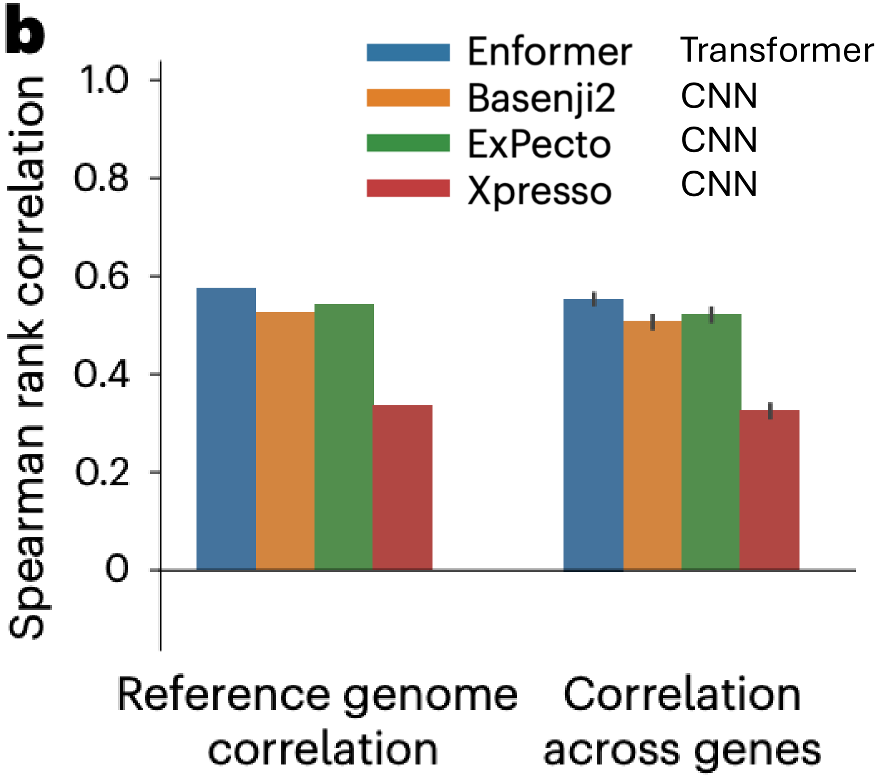


- Enformer finds the **single causal SNV** for *DDX11*
- In theory, these models can overcome linkage disequilibrium

Sasse et al. *Nature Genetics* (2023).

Similar pattern across other deep models

- Left: correlation between model prediction when trained on **reference genome** and **population-average** gene expression across Geuvadis (n=421)
- Right: correlation between model prediction when trained on **individual** genome and that **individual's** gene expression
- All four deep models can do these



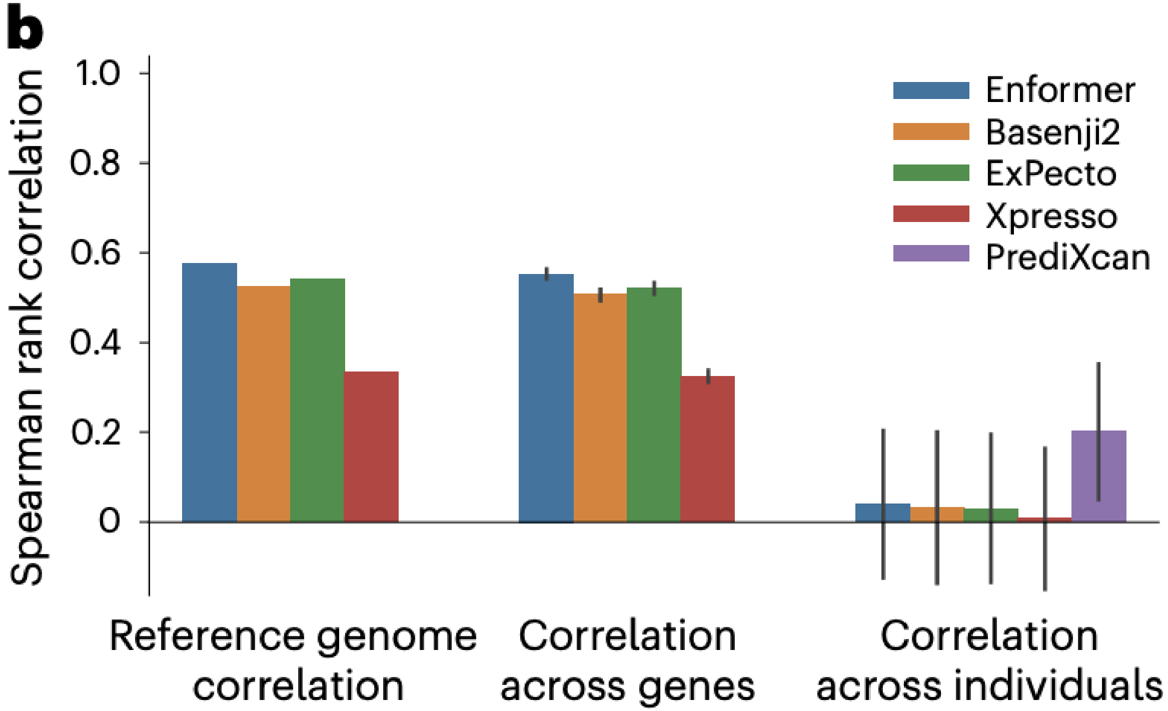
Huang et al. *Nature Genetics* (2023).

So can deep learning models predict gene expression from sequence?

Well….

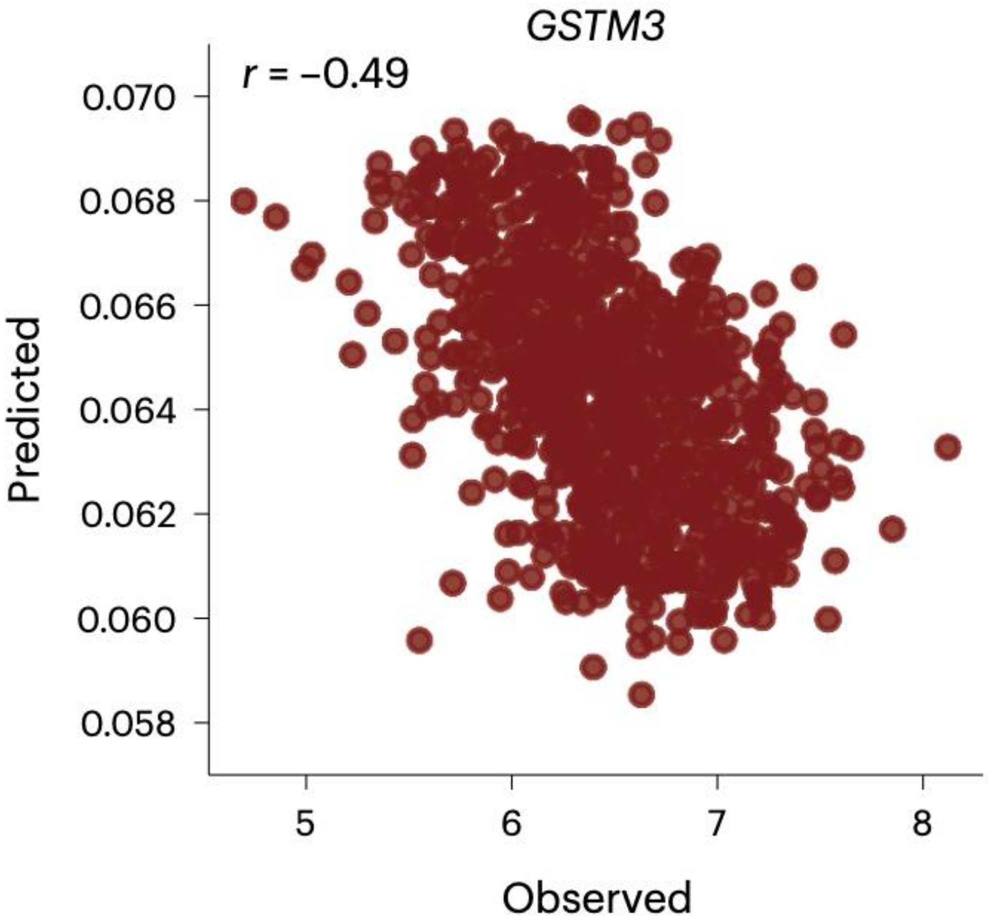
Models cannot predict individual variation

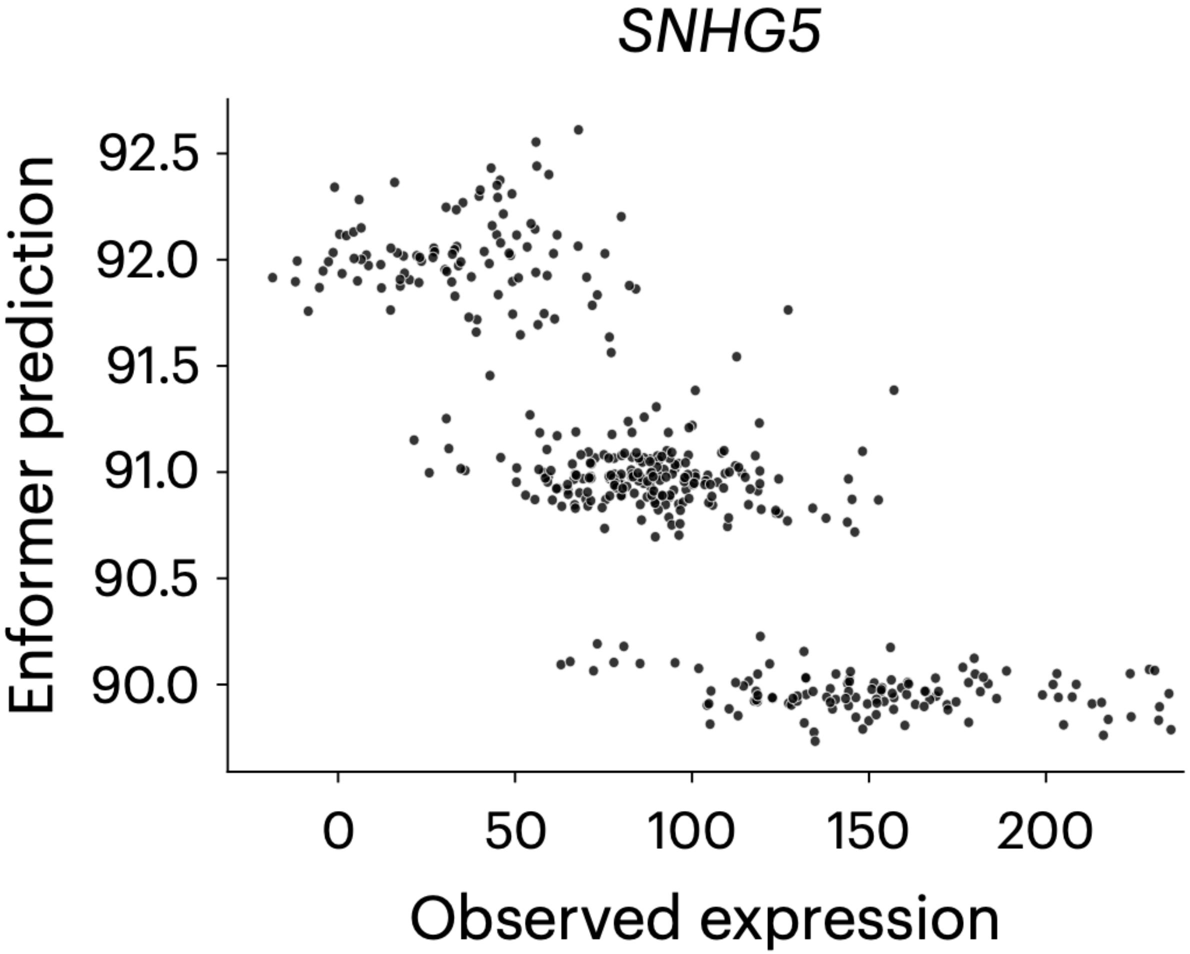
- Right: correlation between model outputs and gene expression **across individuals**
  - No correlation
- PrediXcan, a supervised linear model, is a lower bound for theoretical performance
  - Lots of signal is not being captured by deep learning models



Huang et al. *Nature Genetics* (2023).

Can cherry-pick negative examples too

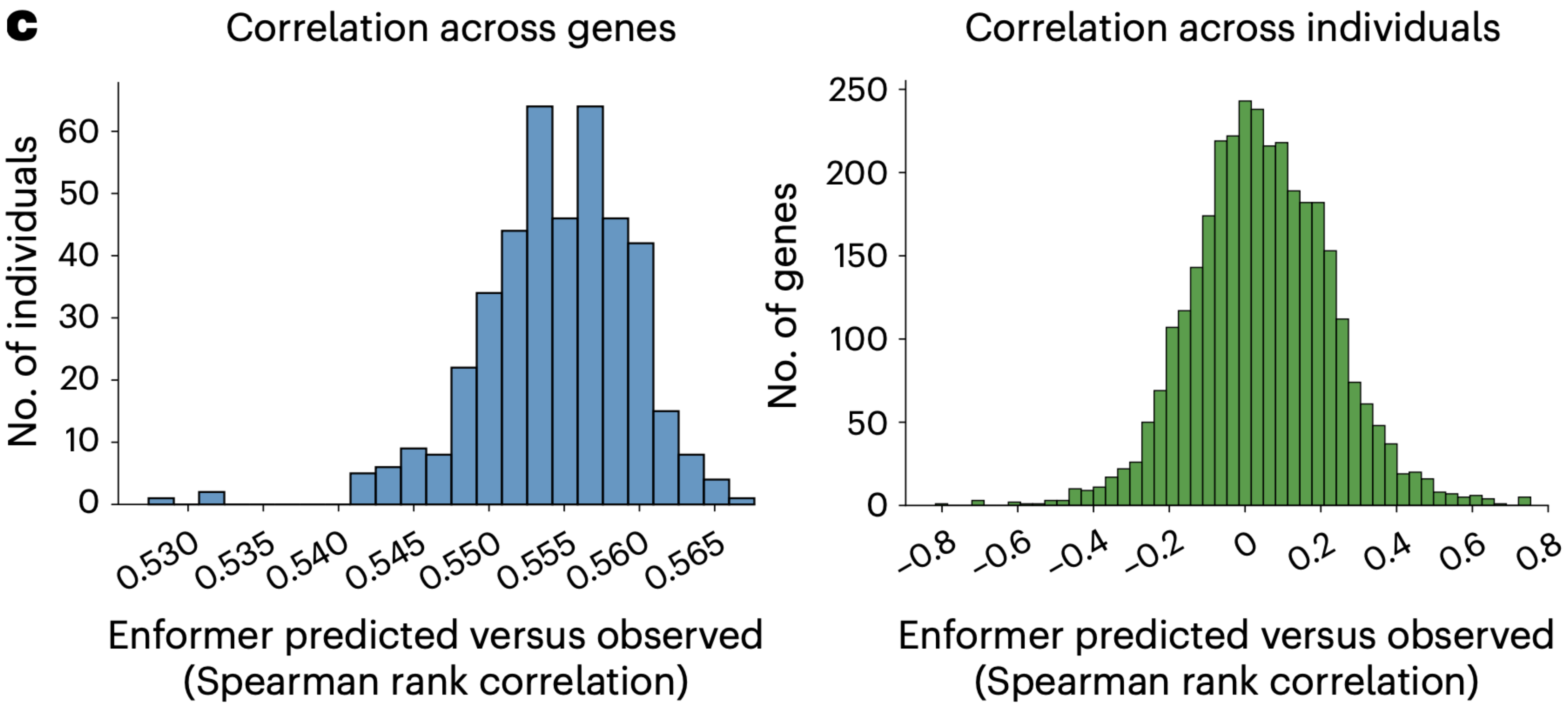




Sasse et al. *Nature Genetics* (2023).

Huang et al. *Nature Genetics* (2023).

Enformer's best example is an outlier

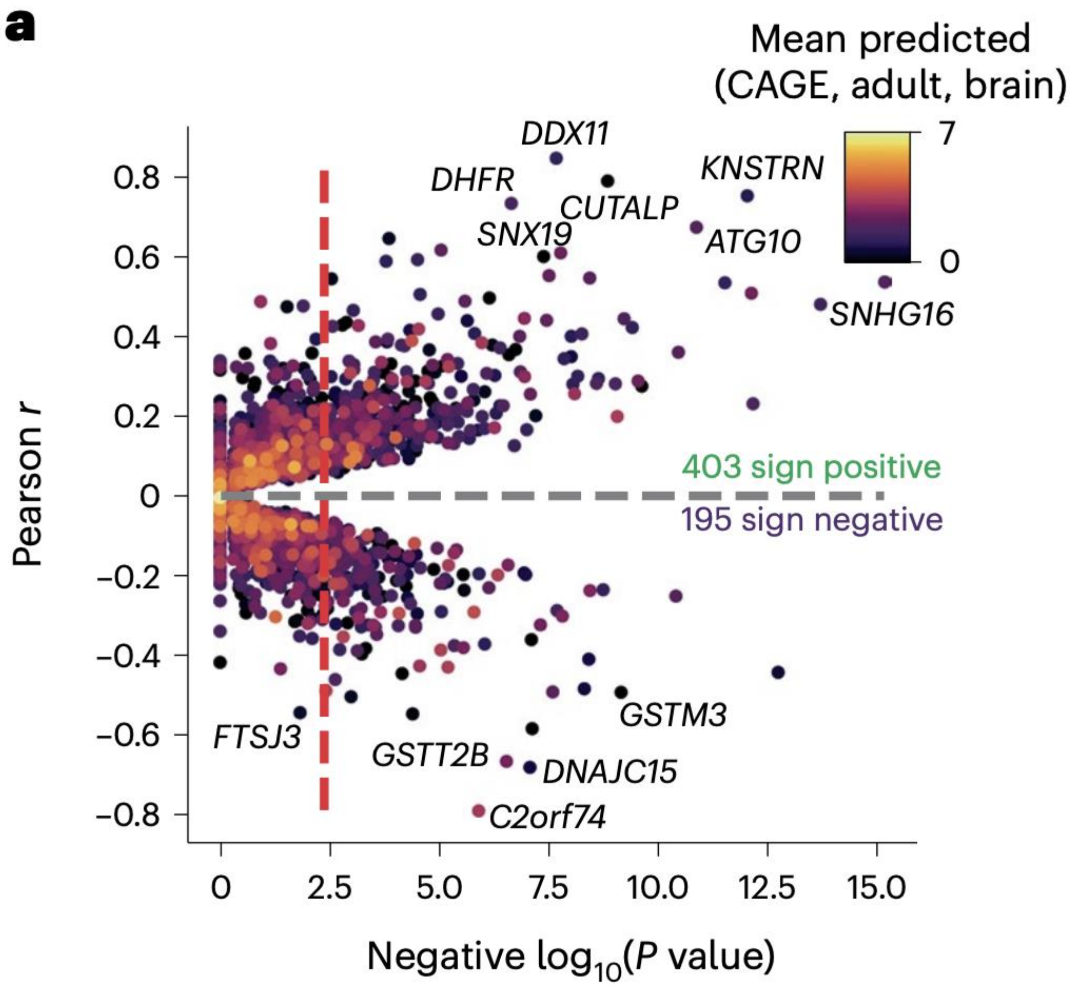


Predictions **across individuals** are essentially **random**

Huang et al. *Nature Genetics* (2023).

Enformer predictions are poor overall

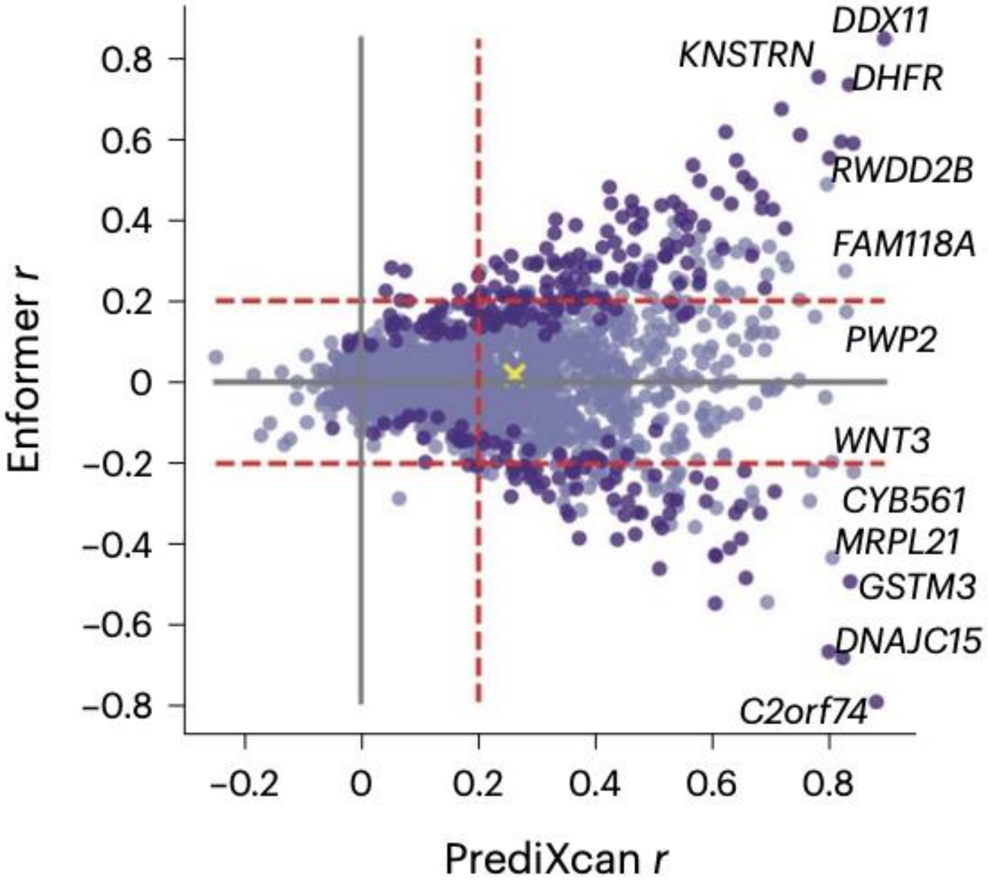
- Enformer predicting 6,825 brain cortex expressed genes
- Use output of Enformer’s most relevant CAGE track
- Vertical red bar = FDR of 0.05
- R values average to **0.01**
- **403** significant **positive** corr.s
- **195** significant **negative** corr.s
  - Predictions are often anti-correlated with ground truth



Sasse et al. *Nature Genetics* (2023).

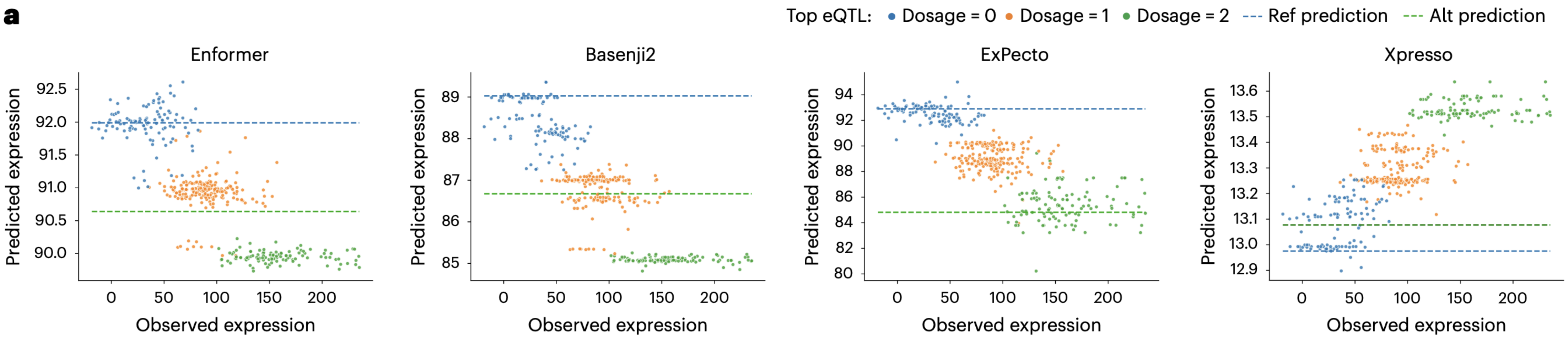
Linear model greatly outperforms Enformer

- PrediXcan, linear elastic net model, here trained on GTEx cerebral cortex data
  - **No** issue with **anti-correlation**
  - Finds 921 significant genes vs. 162 by Enformer
- All significant genes found by PrediXcan should have at least 1 causal variant in the input
  - Indicates loci that Enformer **missed**



Sasse et al. *Nature Genetics* (2023).

Deep models don’t agree on predictions

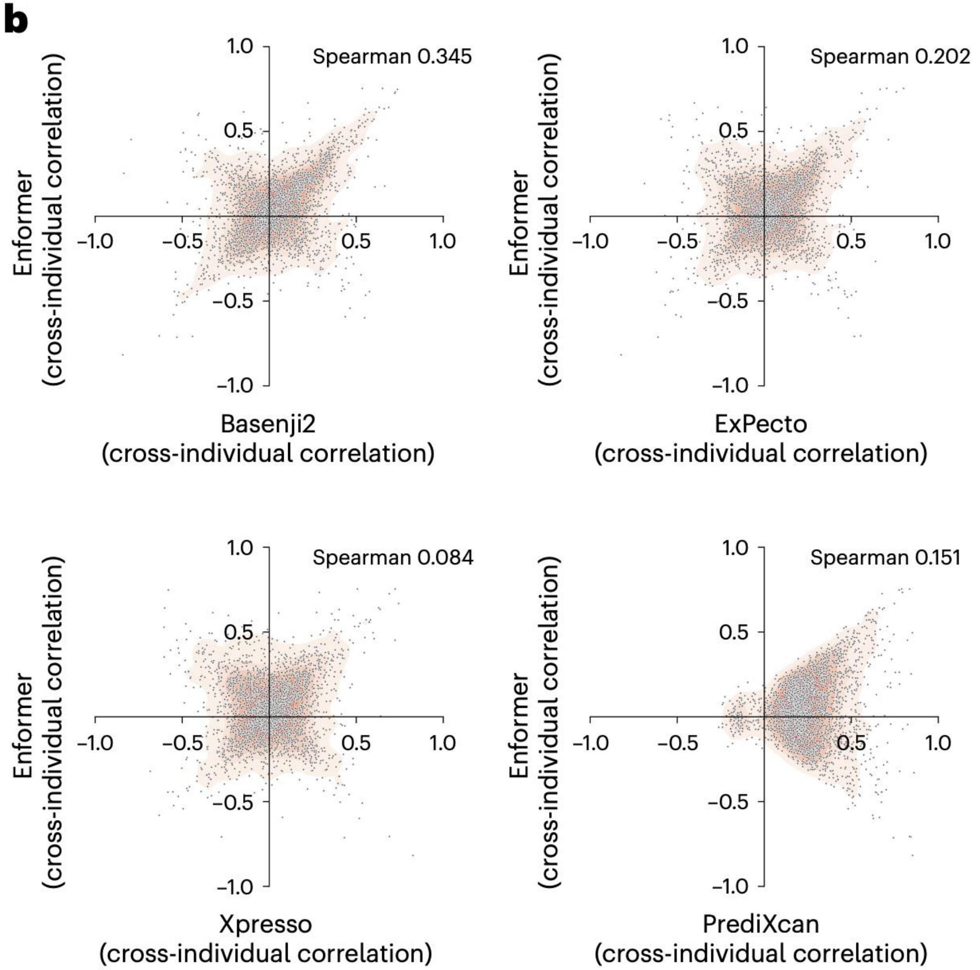


- Example gene *SNHG5*
- Enformer, Basenji2, and ExPecto are **wrong**, while Xpresso is **right**
- Suggests that certain genes are **not inherently harder**, since the models don’t get the same ones wrong

Huang et al. *Nature Genetics* (2023).

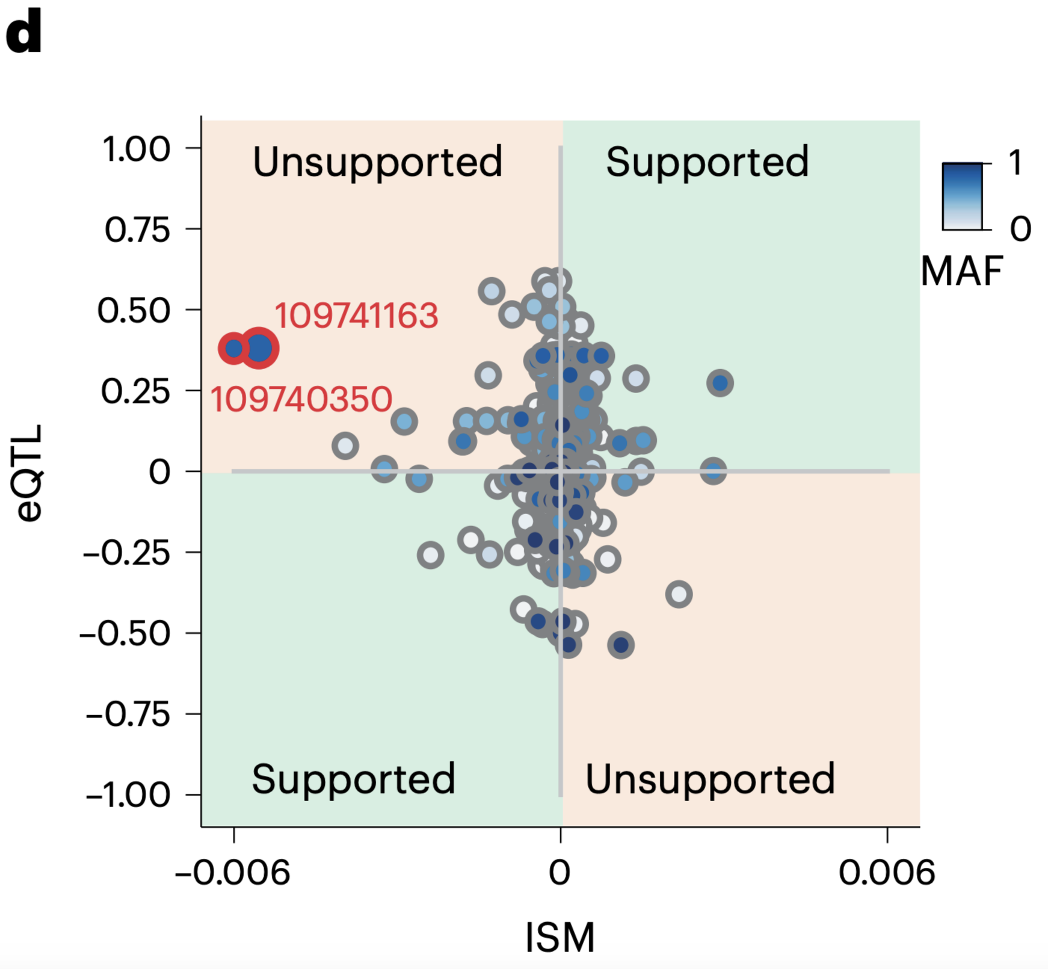
Models agree on magnitude of effect, not direction

- Models don’t agree on **direction** of effect
- However, they do generally agree on **magnitude**
  - X shape instead of circle
- Can tell when a variant is causal, but not the **type** of causality



Huang et al. *Nature Genetics* (2023).

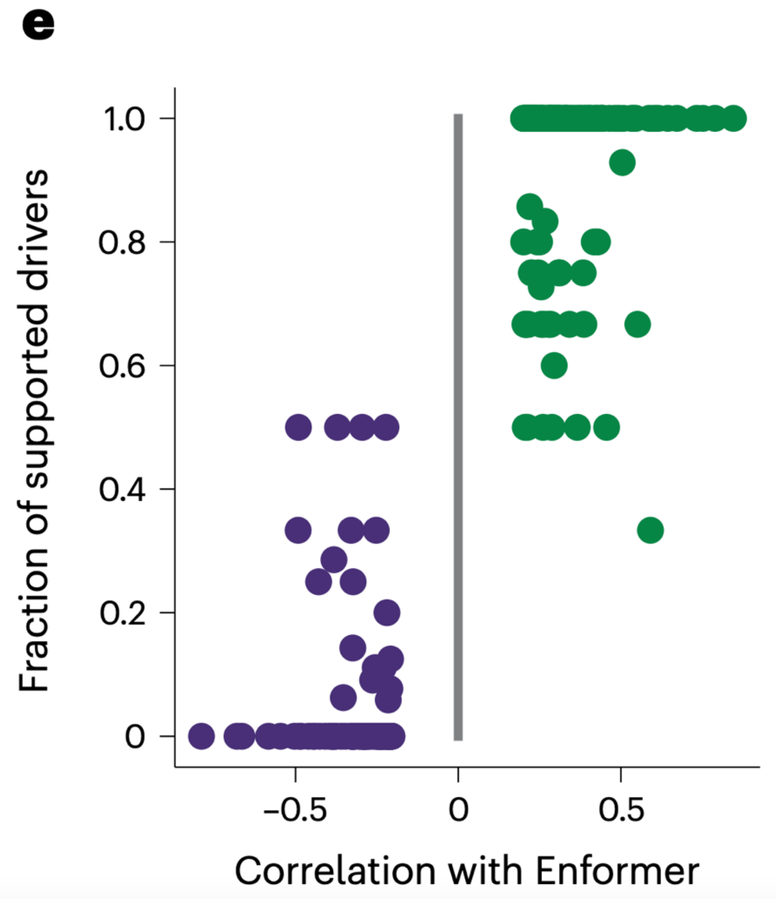
Enformer SNVs are not supported by eQTLs



- All SNVs (n=706) within 197kb window of example *GSTM3* gene
- x axis = **ISM** score (importance to Enformer predictions)
- y axis = **eQTL** effect size (R value between genotype and expression count)
- Two biggest ISM scores are **unsupported**
- Generally, **no agreement**

Sasse et al. *Nature Genetics* (2023).

Performance depends on finding the right SNVs

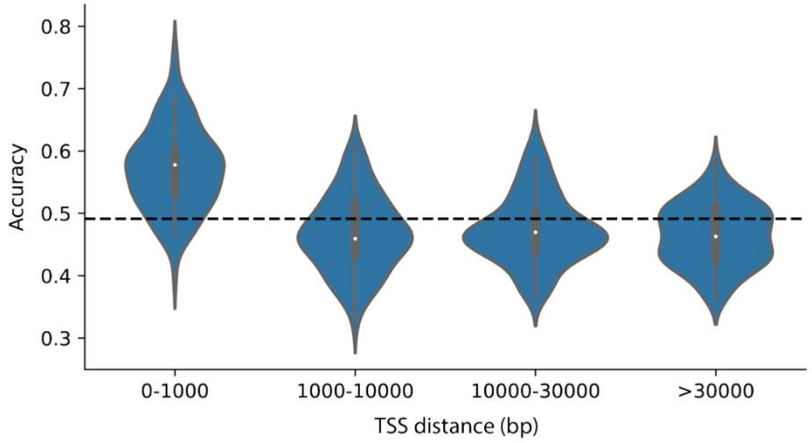


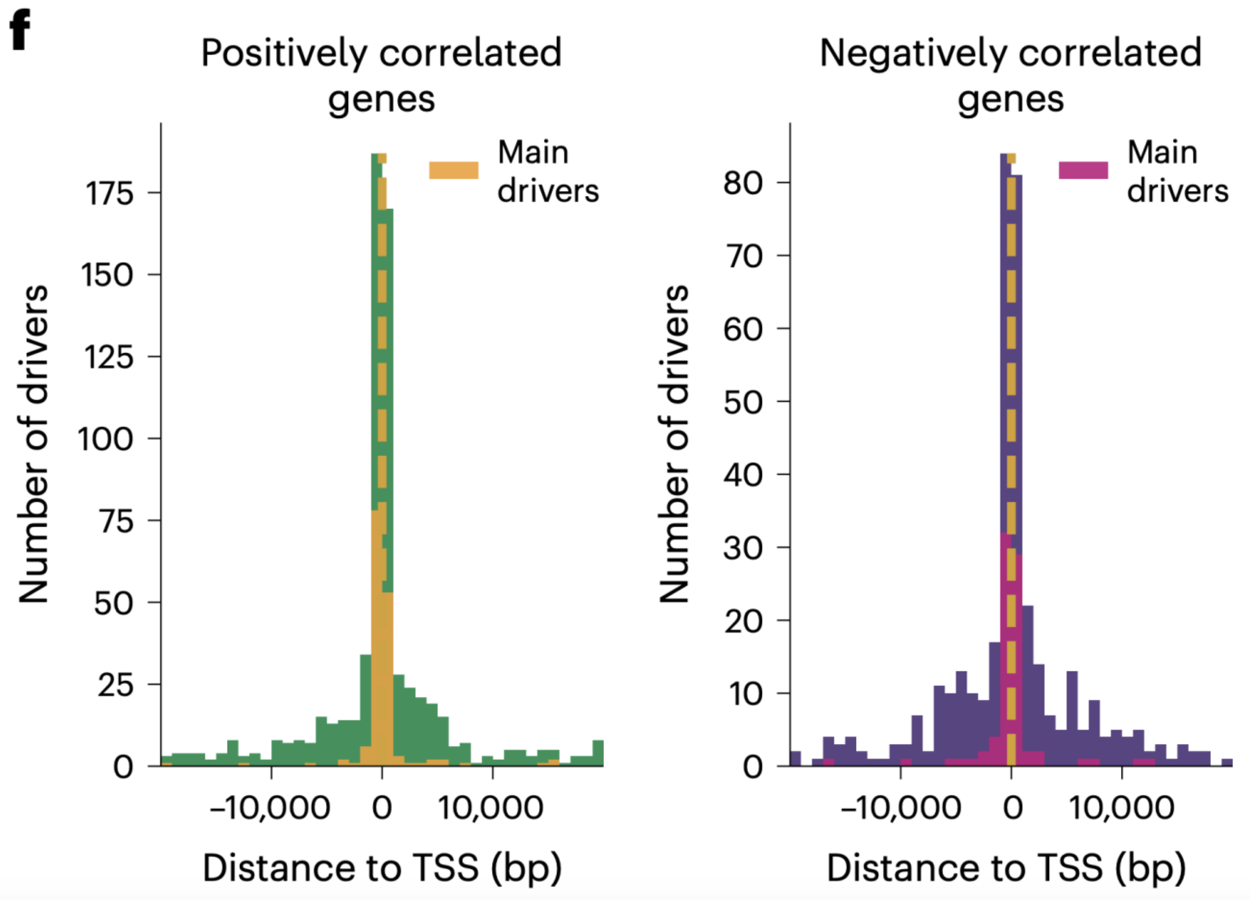
- Genes where Enformer does **poorly** (purple) have a **low fraction** of supported SNVs
    - Supported by eQTL analysis
- Genes where Enformer does **well** have **supported** SNVs
- Predicting gene expression depends on correctly identifying **driver variants**

Sasse et al. *Nature Genetics* (2023).

Enformer doesn’t use its full context window

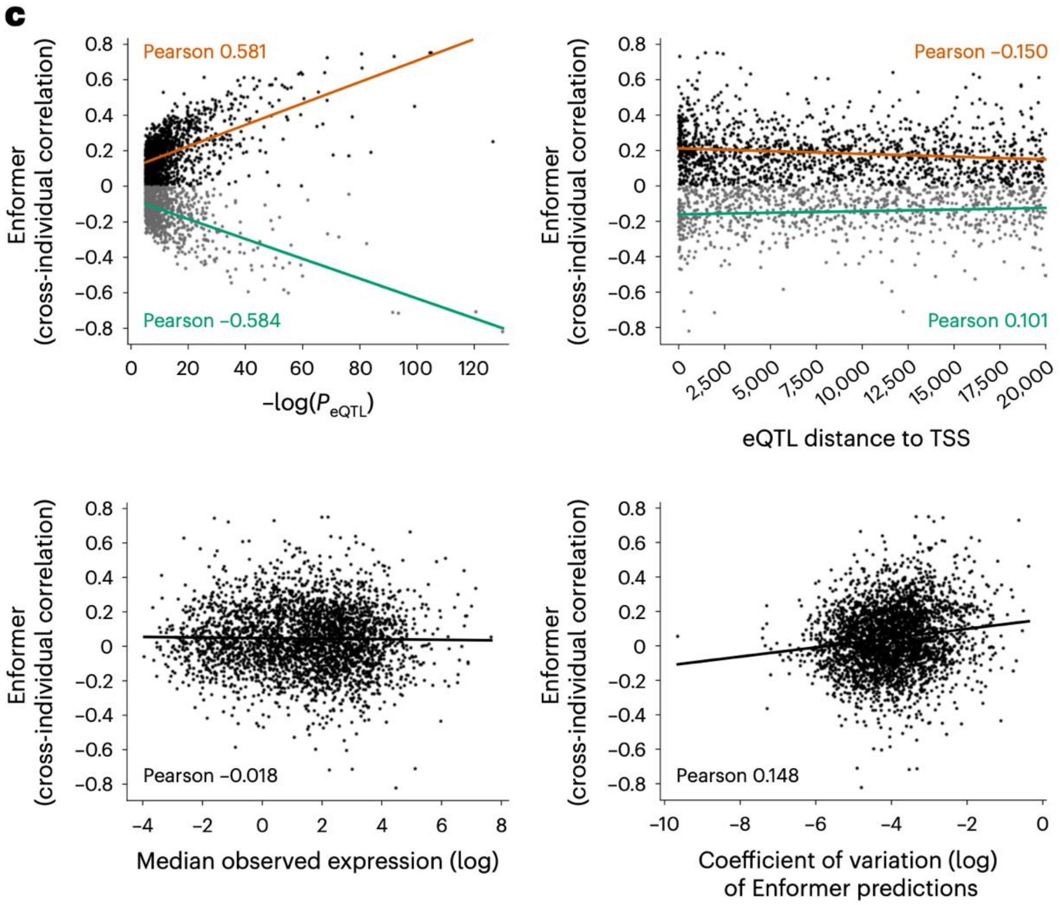
- Most drivers it finds are very **near** the **TSS**
- Model is most **accurate** on variants near TSS





Sasse et al. *Nature Genetics* (2023).
Huang et al. *Nature Genetics* (2023).

Hard to tell why models fail

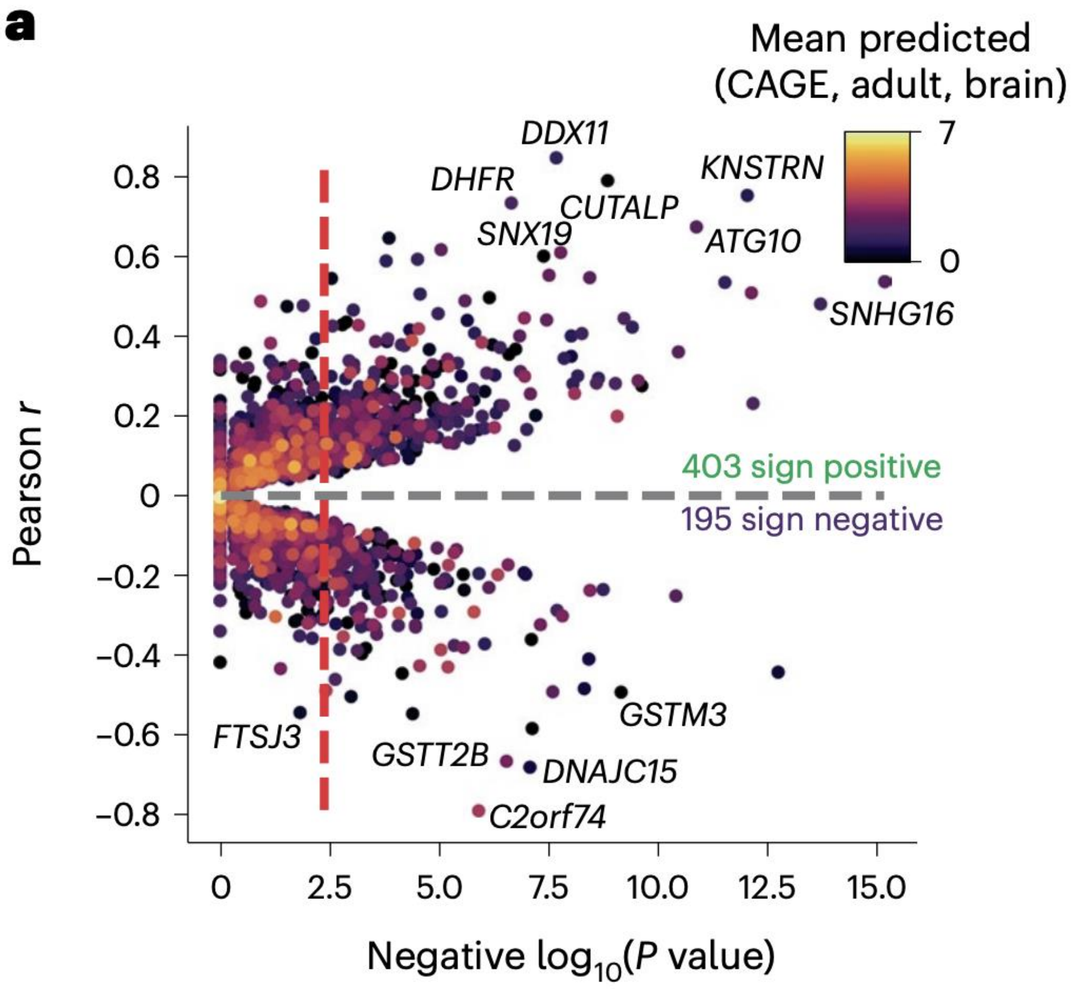


- Cross-individual correctness **does not correlate** with eQTL *p* value, distance to TSS, or other features
- It might just be **noise**

Huang et al. *Nature Genetics* (2023).

Conclusions

- Deep learning models do not understand whether a variant will make gene expression go **up** or **down**
- Can learn patterns across a **population** but not in **individuals**
- Models have "blurry vision"
- Understand whole **motifs**, not individual **bases**



Sasse et al. *Nature Genetics* (2023).

Future work

- Sasse et al.:
  - **Train** on **diverse genomes** and their corresponding gene expression
  - Figure out how to model additional data such as **post-transcriptional** RNA processing that impacts gene expression
  - **Assess** models on **direction** of effect of SNVs in individuals
- Huang et al.:
  - Determine if models can predict direction of effect of SNVs on **other data modalities** such as chromatin accessibility
    - If not, then they struggle to understand **regulatory grammar**
      - Incorporate hierarchical models of gene expression
    - If so, then they need to learn **local effects**
      - **Train** on more **diverse genomes**

Next meeting

- **Date and Time**: Tuesday, April 22nd, 12 - 1pm
- **Location**: Malone 228 and Zoom
- **Presenter**: Kuan-Hao Chao

Sign up to present this semester!!! →
We are on the second pass now