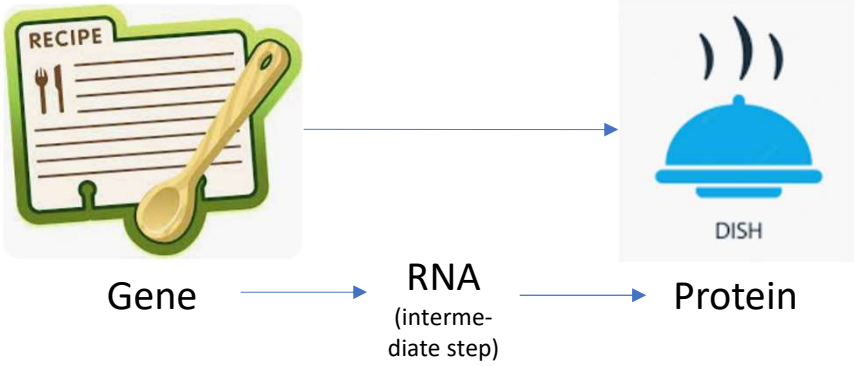
Transfer learning enables predictions in network biology (Geneformer)

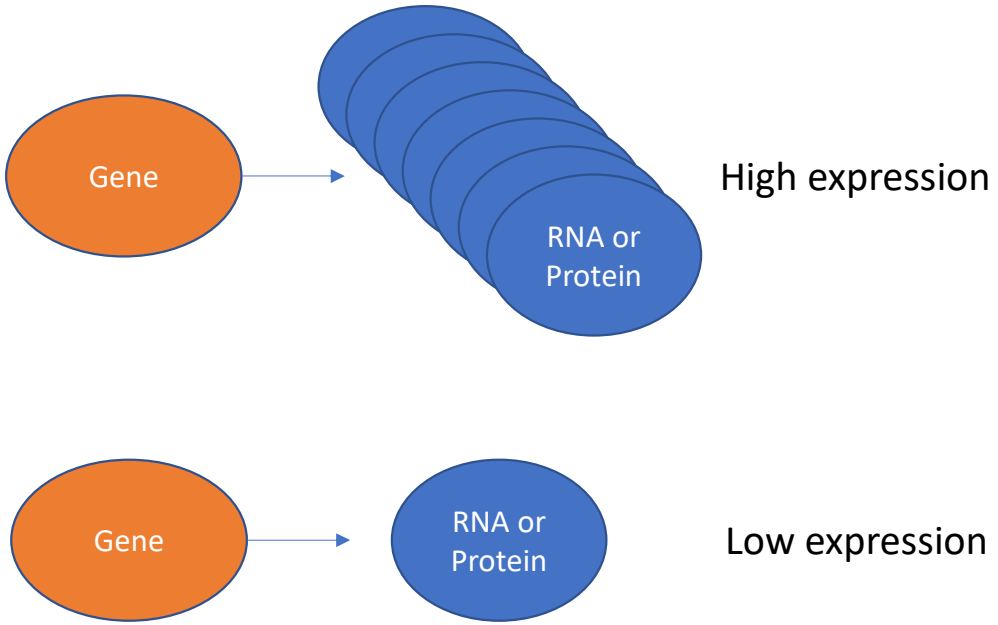
Nature 2023

Outline

- Background
- Pre-training fine-tuning framework
- Dataset
- Tasks

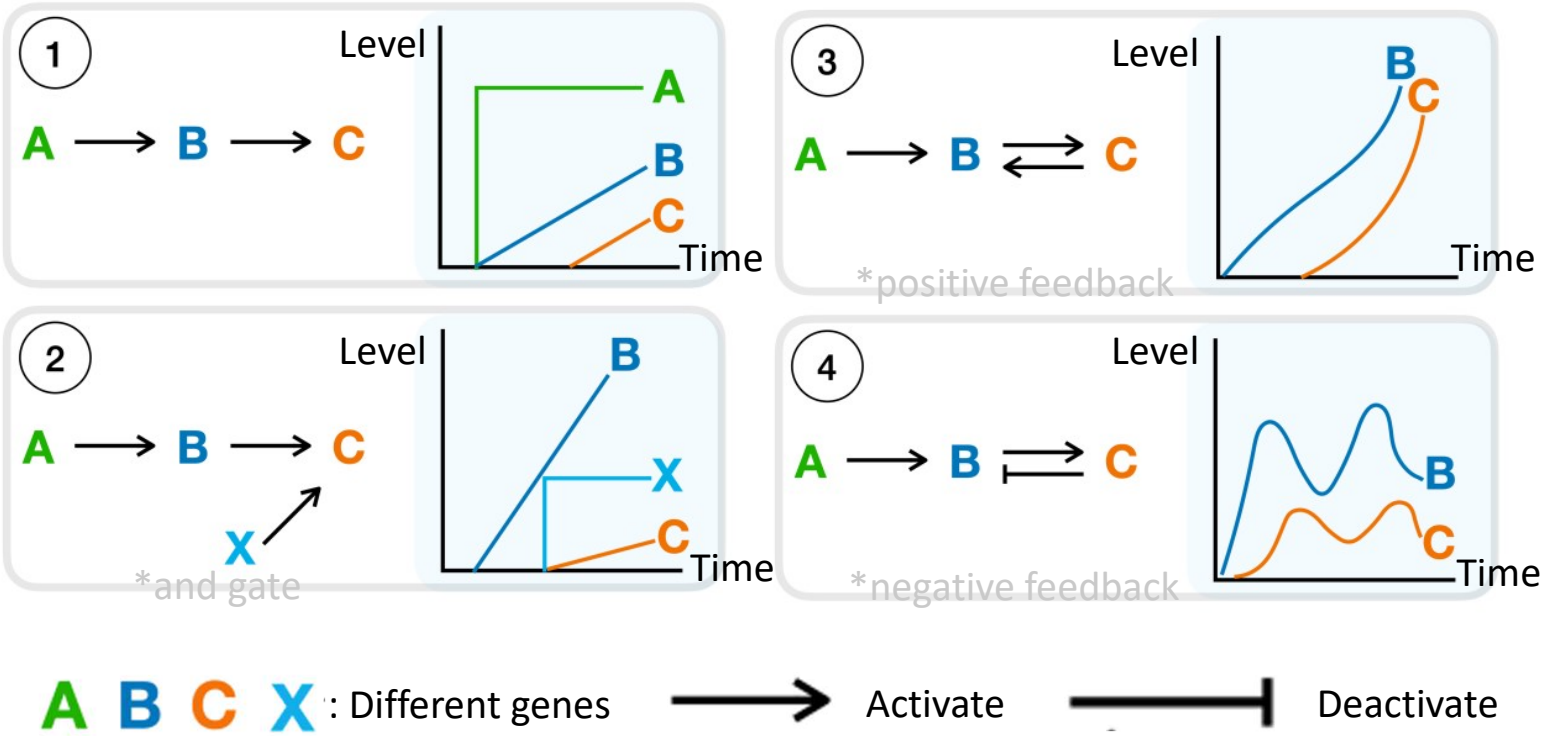
Gene expression





Gene regulatory network

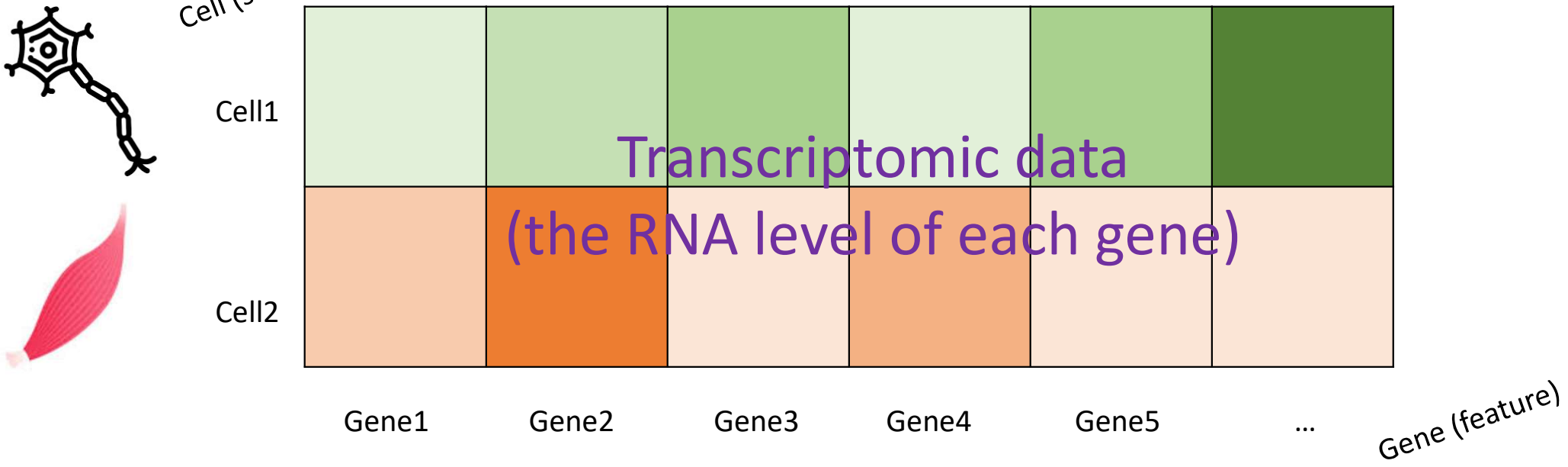
The expression of genes are controlled by other genes.



https://sitn.hms.harvard.edu/flash/2021/gene-regulatory-networks-from-dna-to-development/

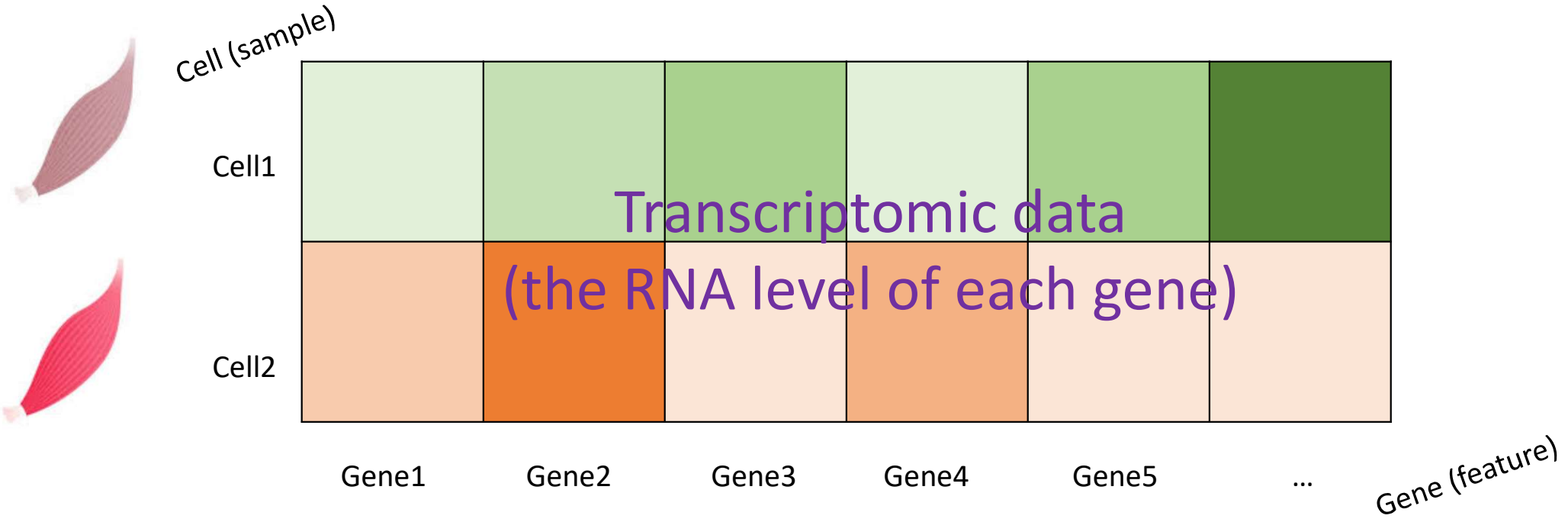
Different cell types have different gene expression

Gene expression can be viewed as a table



Cells in different states have different gene expression

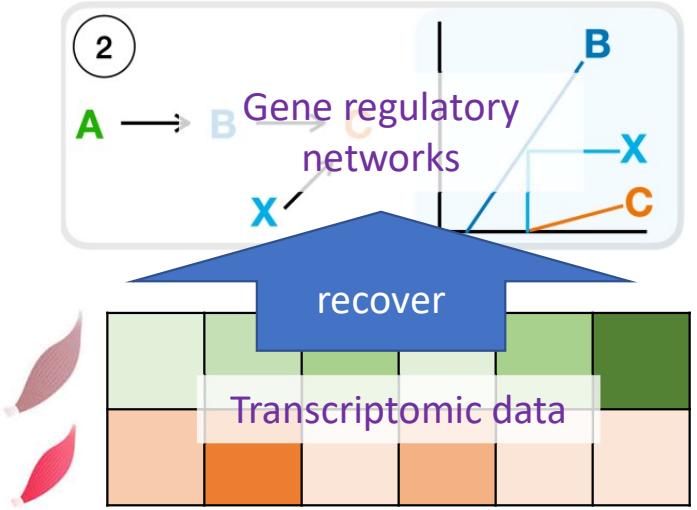
Gene expression can be viewed as a table



Motivation

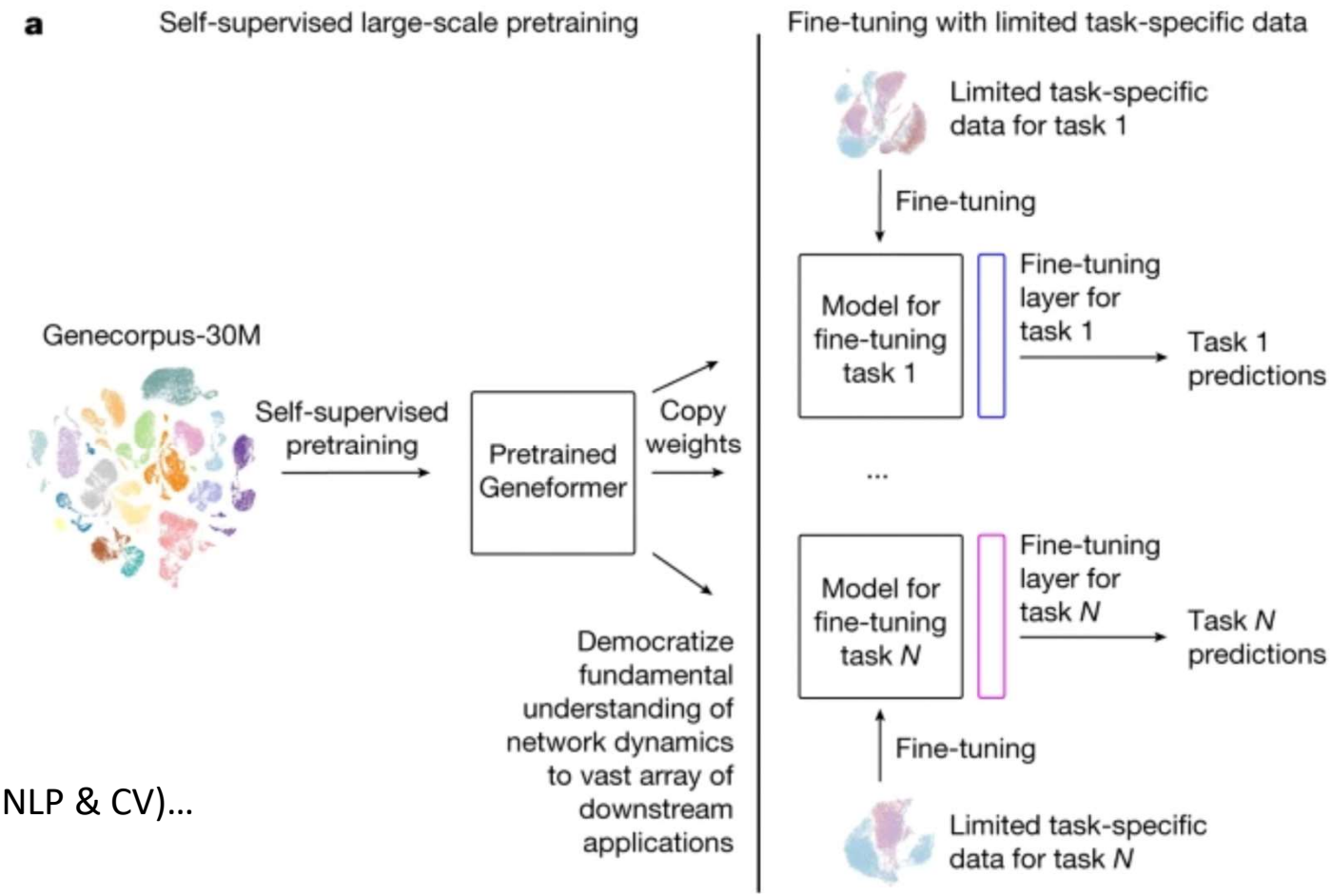
- Understanding the gene regulatory network helps us know what is the cause of the disease, so we can focus on correcting the cause instead of the effect

- However, discovering the gene regulatory networks needs a lot of transcriptomic data, which prohibits the network discovery when the data is limited (e.g., rare diseases).



Idea

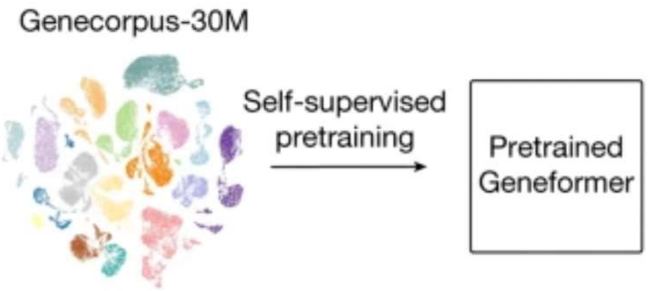
- The idea is to first self-supervisedly pre-train a model on rich transcriptomic data, and then fine-tune on the limited transcriptomic data.

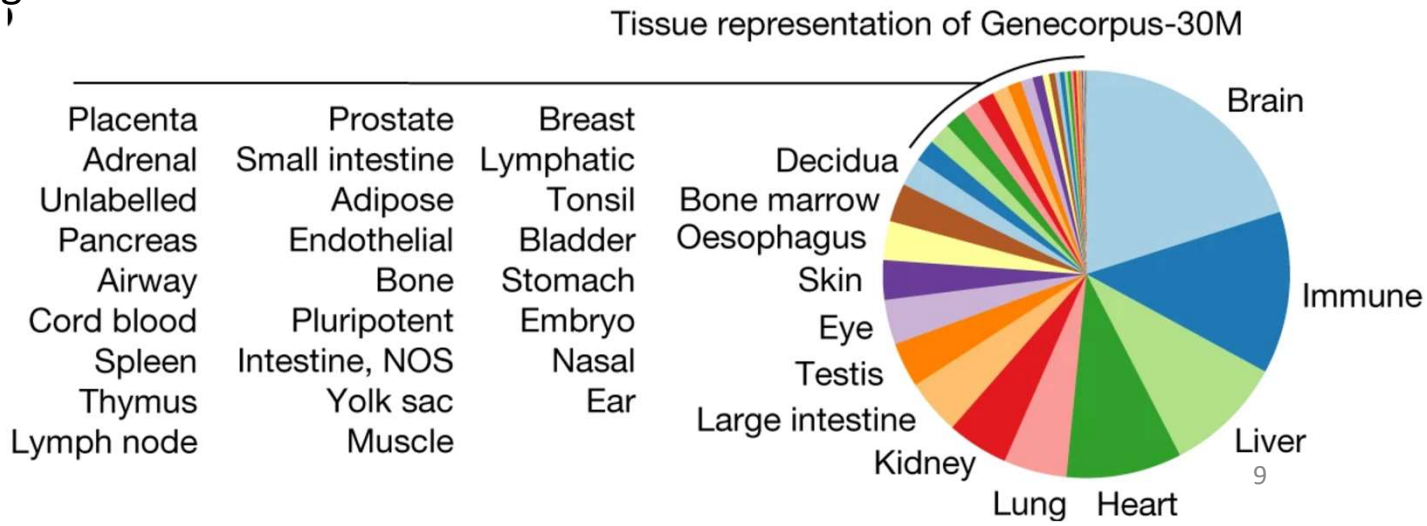


This idea is quite common these days (NLP & CV)…

Large-scale dataset: Genecorpus-30M

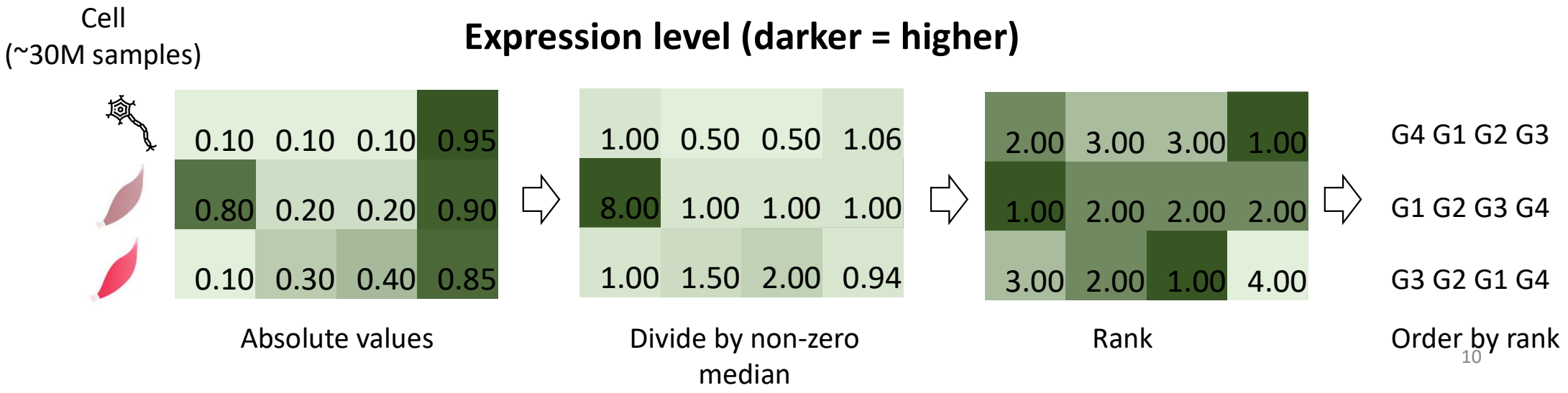
- 29,900,531 human single-cell transcriptomes
  - Collected from a broad range of tissues from publicly available data
  - 561 sub-datasets from 112 sources (after 2016)
  - excluded cells with high mutational burdens (for example, malignant cells and immortalized cell lines)





Rank value encoding

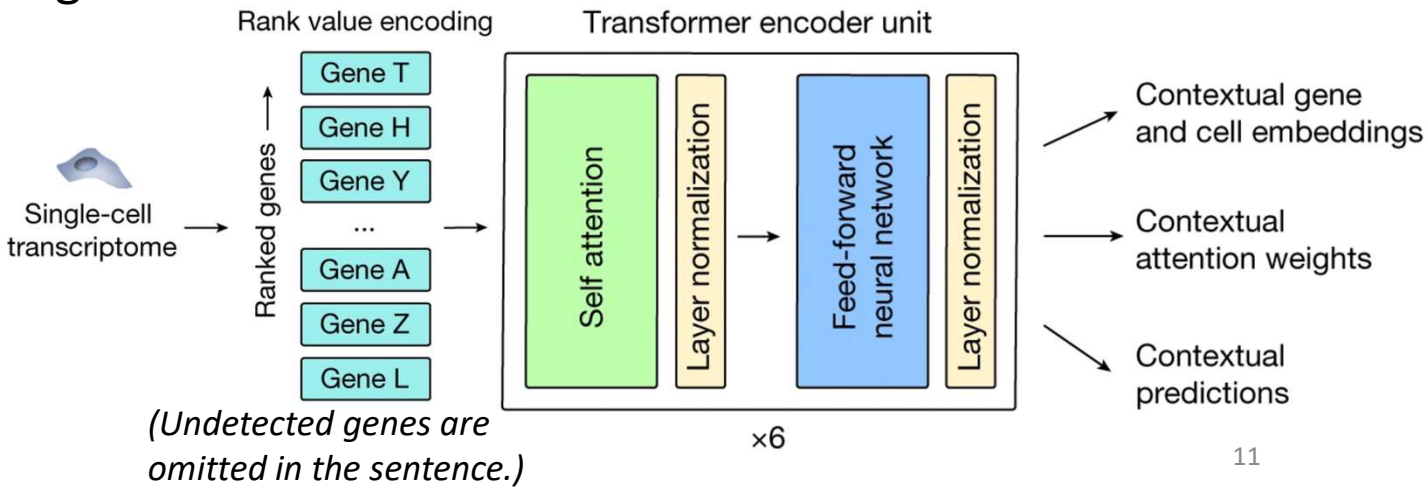
- Represented as the rank of normalized expression level.
  - The normalization is shown below
  - Normalization priortizes the genes that distinguish cell state
  - Normalization deprioritizes ubiquitously highly expressed housekeeping genes (e.g., right most gene)
  - Normalization remove technical artefacts that cause systematical biases



(output word emb)

Geneformer architecture

- A regular MaskLM transformer based on open-source library (huggingface)
- Each gene is treated as a token
  - 25,424 tokens for protein-coding or miRNA genes,
    which are detected in a median of 173,152 cells within Genecorpus-30M
  - 2 special tokens for masking and padding
- The input is a “sentence” of genes
  - 6 layers
  - Max length: 2048
  - Embed dim: 256
  - Attn head: 4
  - FF dim: 512
  - Mask rate: 15%

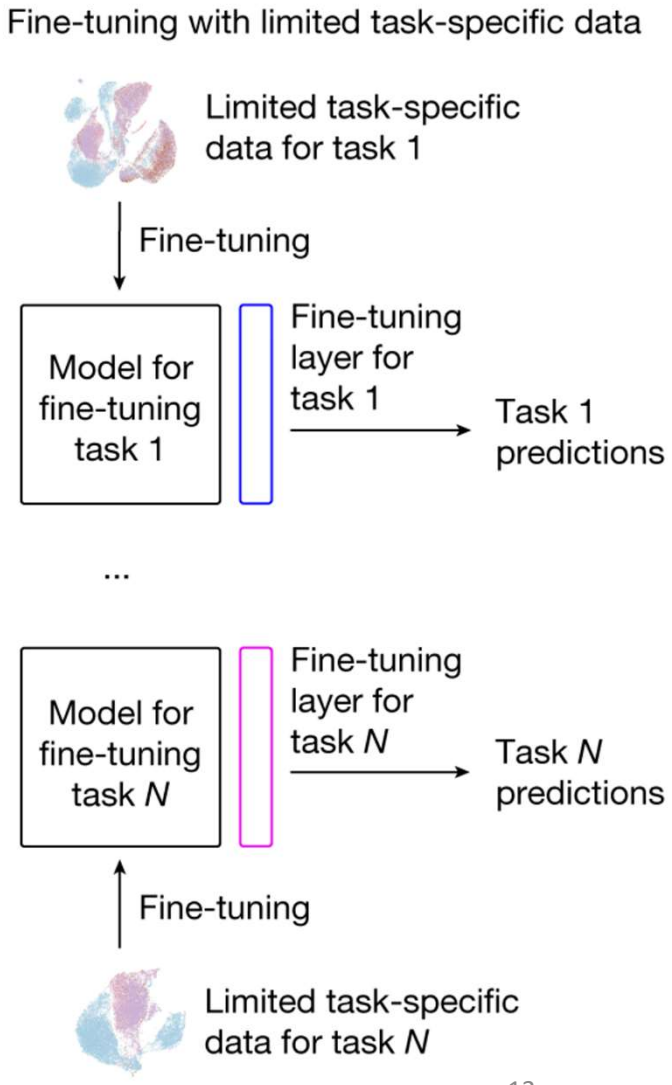


Training details

- Effieciency:
  - Length grouping (each minibatch contains the sentences with similar length)
  - Deepspeed distributed training (12 V100s on 3 nodes, 3 days)
    - seems to use model partition and CPU offloading

Fine-tuning

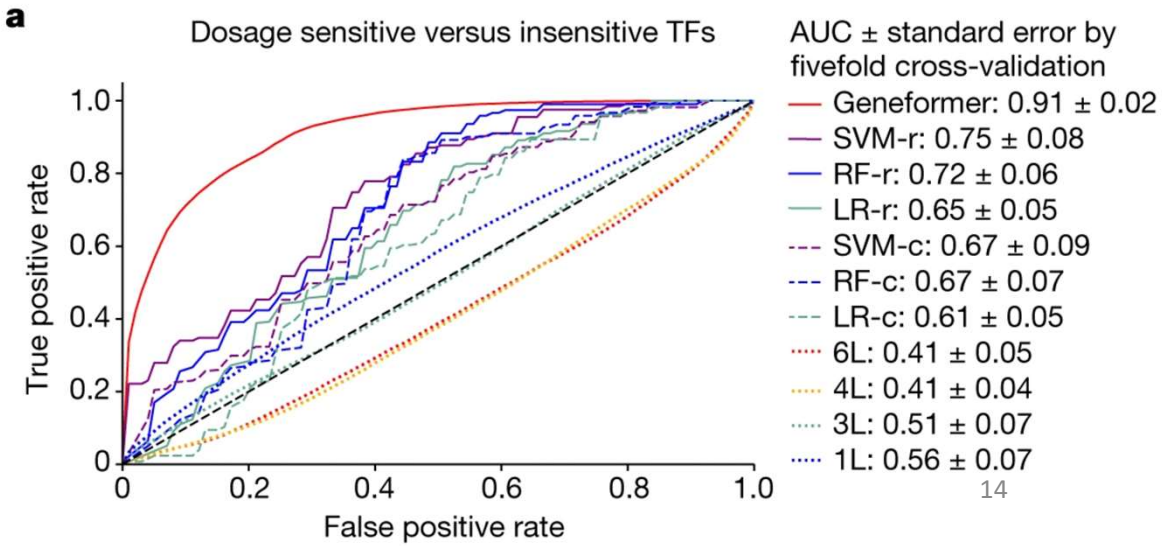
- Fine-tune an additional task specific transformer layer
- Only a single training epoch to avoid overfitting



Fine-tuning task: dosage sensitivity prediction

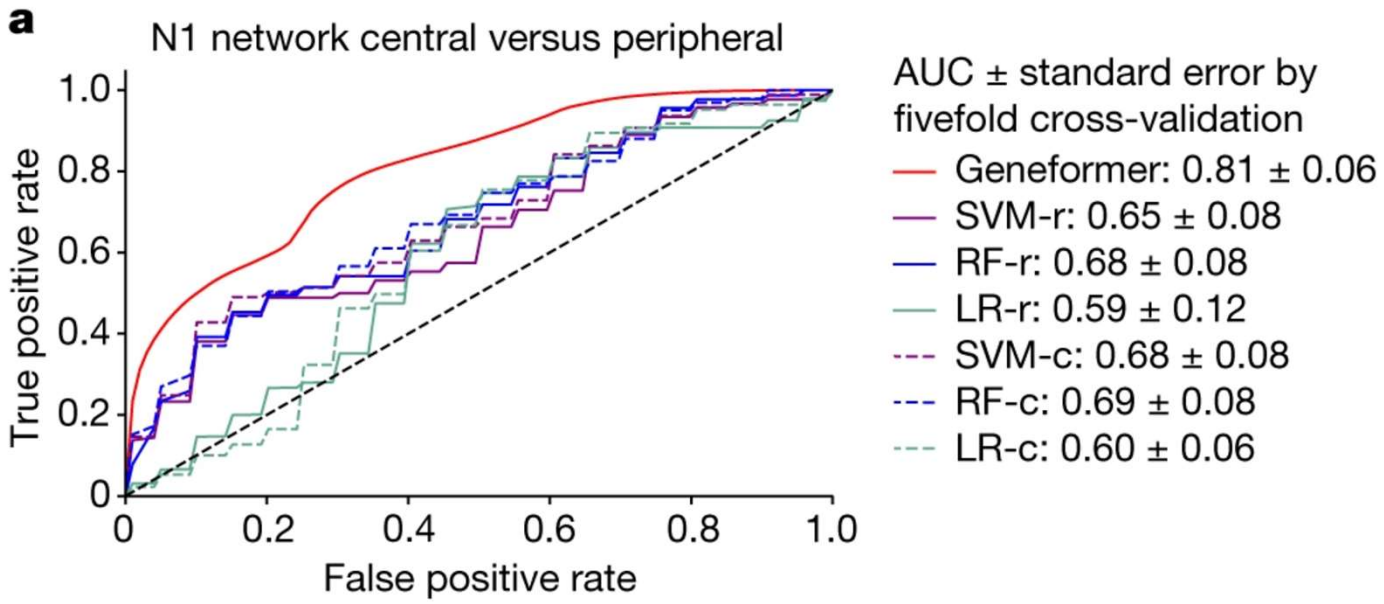
- The gene can duplicate. The same gene can occur more than once in the chromosome.
- Some genes are dosage sensitive, the change of their copy number has larger effect.
- The task is to classify whether the gene is dosage sensitive (binary classification).

- Only 10,000 cells are used to finetune Geneformer.
- Geneformer achieves best AUC compared with SVM, random forest (RF), and logistic regression (LR).
- 6L, 4L, 3L, 1L: Geneformer without pre-training



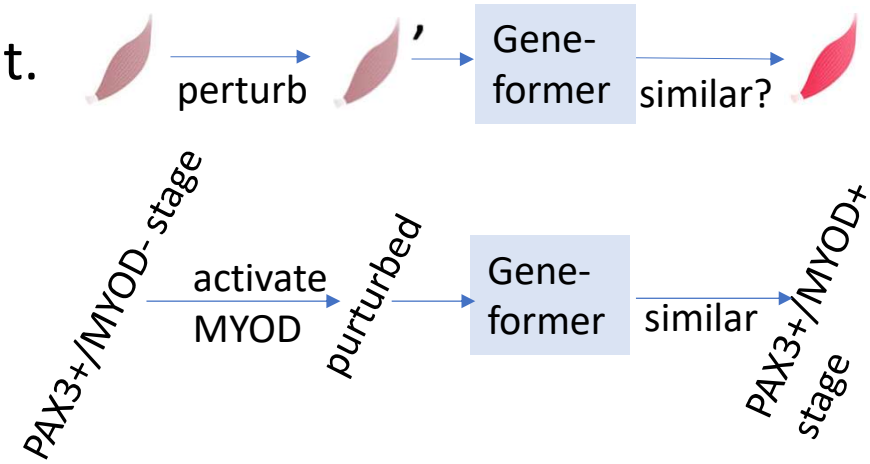
Fine-tuning task: Network dynamics predictions

- We want to predict whether a gene is in the central of the regulatory network (binary classification).
- Fine-tuned on 30,000 normal endothelial cells without perturbation data.
- Tested on NOTCH1 (N1)-dependent gene network



Zero-shot task: In silico gene perturbation

- We want to manually activate a gene.
- Assume = G1 G2 G3 G4
- We want to activate G3, so move G3 to the front.
  - ’ = G3 G1 G2 G4
- Given G3 G1 G2 G4, Geneformer outputs
  - The gene emb (word emb) of G1 to G4
  - The cell emb (sentence emb) by averaging the word emb of second last layer
- The cosine similarity to the other cell is calcuated to determine the effect of the purtabation



Attention analysis

- 20% of attention heads significantly attended transcription factors more than other genes

- Attention heads in the earliest layers were consistently the most diverse in terms of gene ranks they attended, suggesting that the model initially orients to the observed cell state through a joint survey of distinct portions of the input space.
- The middle layers were most broad in terms of gene ranks they attended
- The final layers were dominated by centrality-driven attention heads that focused on the highest ranked genes that uniquely define each cell state

Thanks!