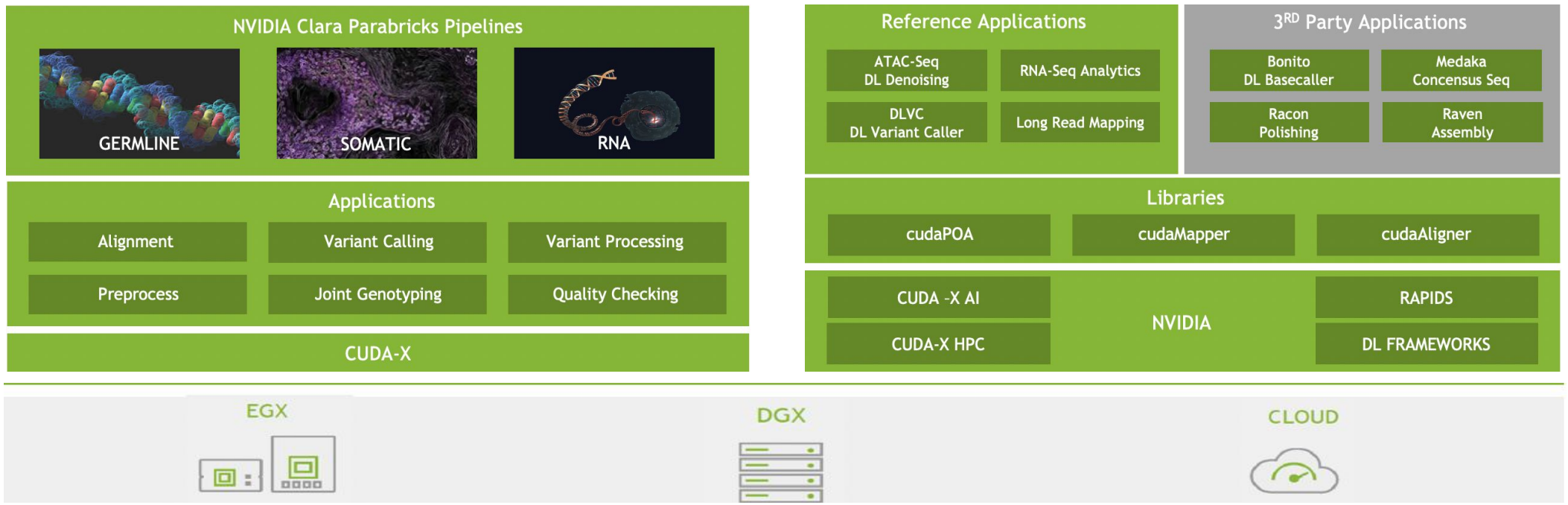
Machine Learning Tools to Analyze Gene Expression and Regulation

Avantika Lal, 07/15/2021

Outline

- GPU-accelerated genomic analysis at NVIDIA
- Gene expression and multimodal data integration
- Single-cell sequencing
- Variational autoencoders
- Predicting gene expression from sequence
- Deep learning to improve data quality

GPU-accelerated genomics at NVIDIA



Parabricks v3.6 Release- July 2021

WGS Pipeline for 30x Human Genome in 22 Minutes on an DGX A100

Update referenced mapped data / re-analysis of legacy data

- > 10x acceleration of BAM2FastQ

More Comprehensive Somatic Calling

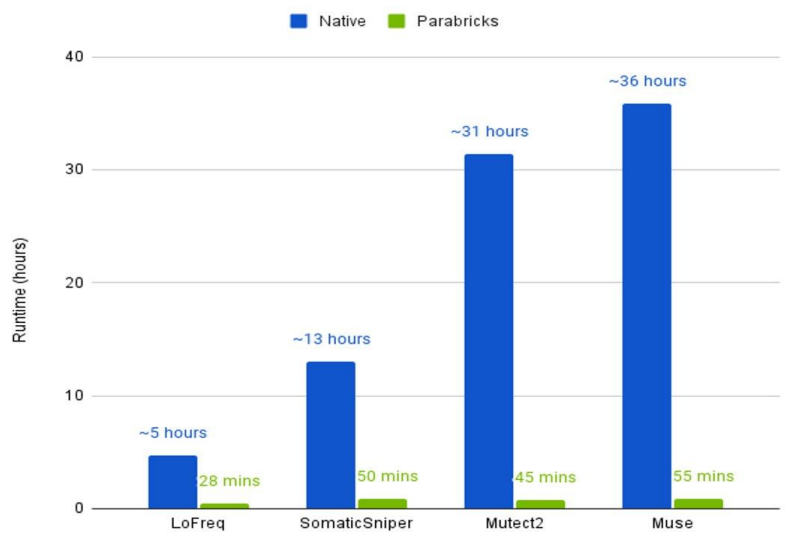
- LoFreq addition, expanding to 4 callers
- Vote-based merge tool for fast variant filtering (e.g. allele frequenc
- VCF Annotation Tool for better quality variants (+ BAM QC)

De novo germline mutation pipeline

- Supports trio sequencing and uses Google's DeepVariant1.0

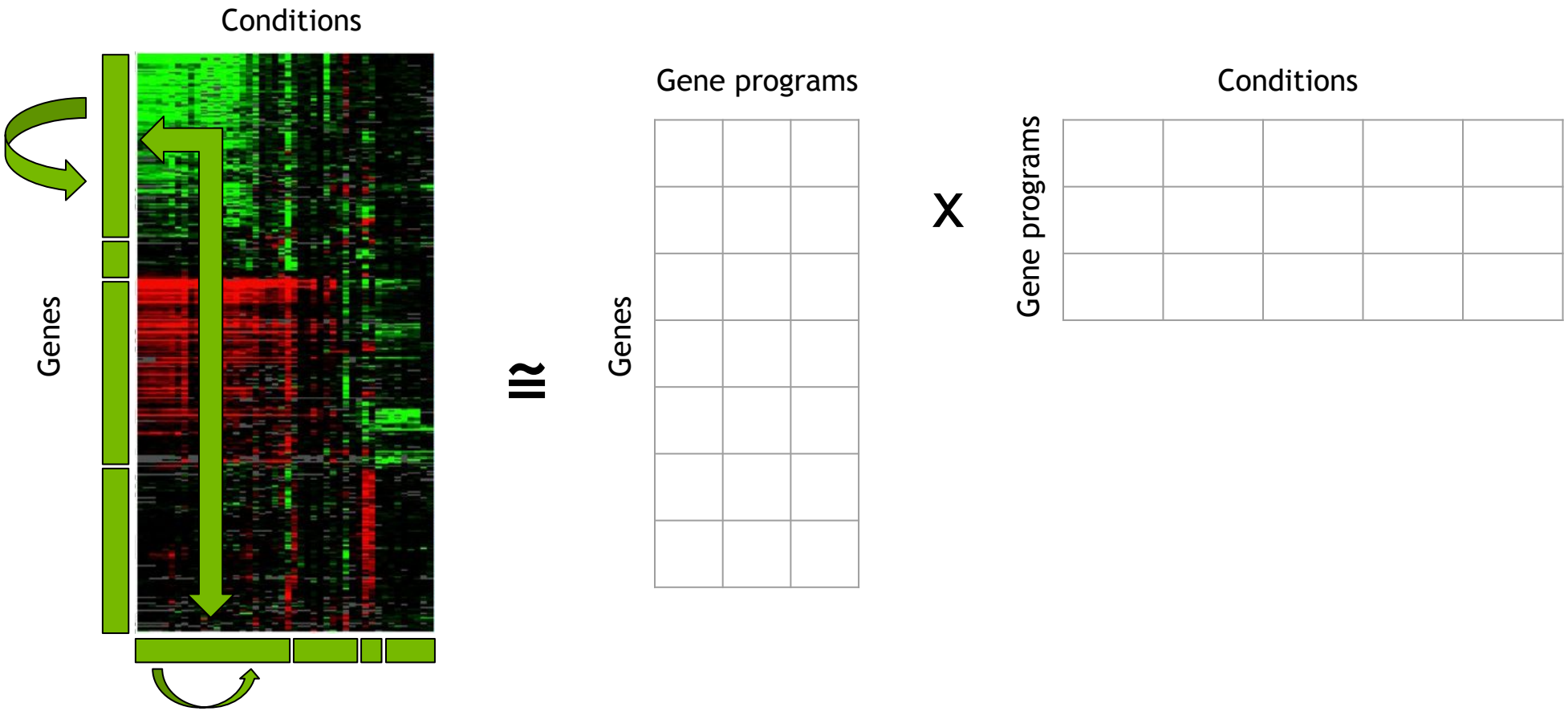
Two Structural Variant Callers

- Manta and Smoove (Lumpy)

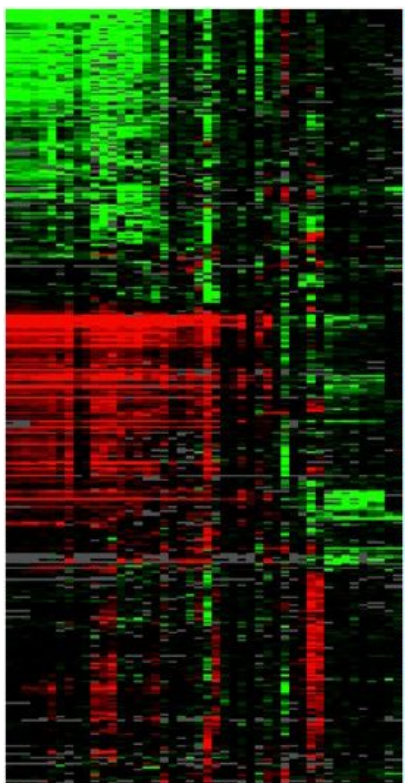


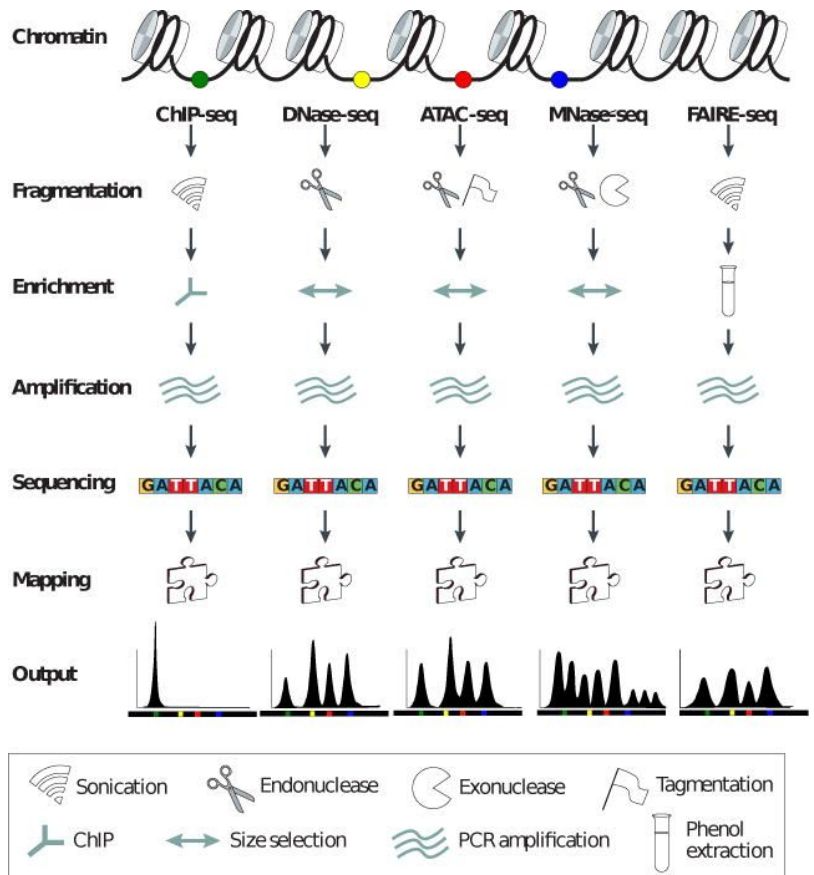
Free 90 day trial license
https://www.nvidia.com/en-us/docs/nvidia-parabricks-general/

Gene expression profiling and multi-omic integration



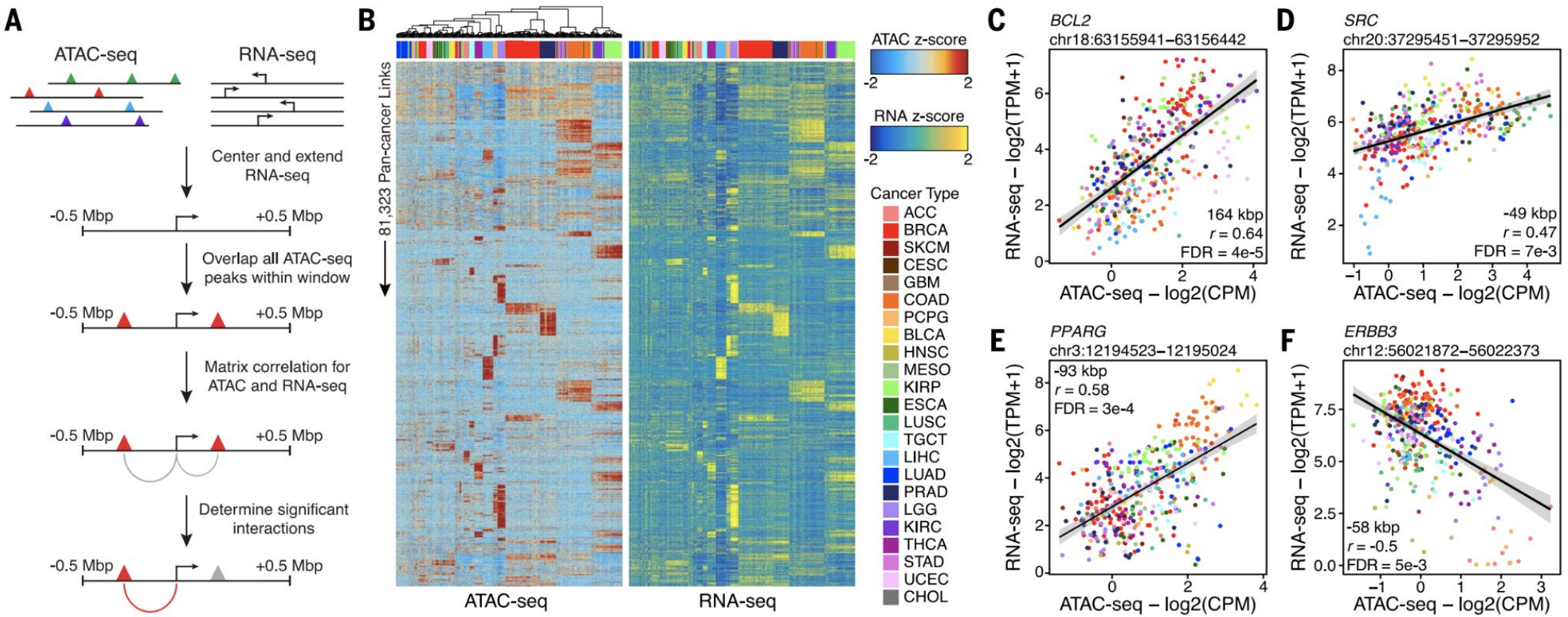
Gene expression profiling and multi-omic integration





Meyer, C., Liu, X. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* 15, 709–721 (2014).

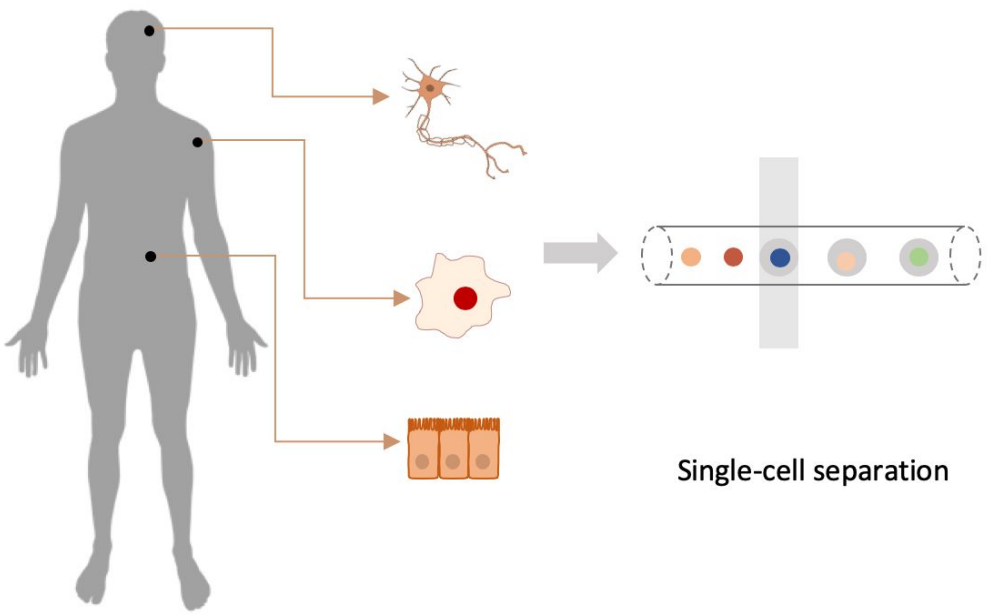
Multi-omic data integration

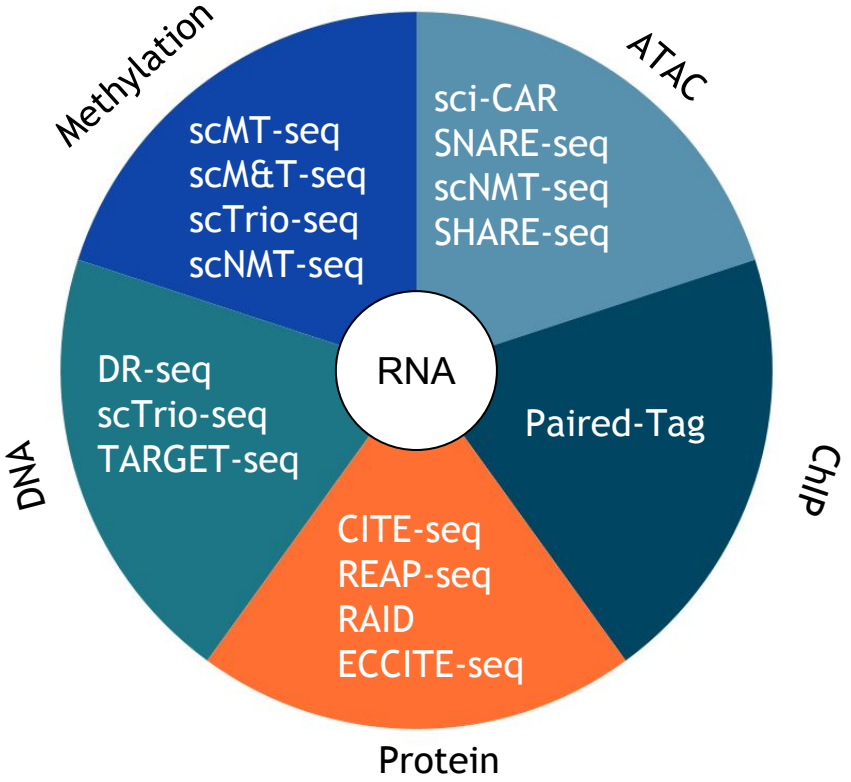


M. R. Corces et al., Science 362, eaav1898 (2018). DOI: 10.1126/science.aav1898

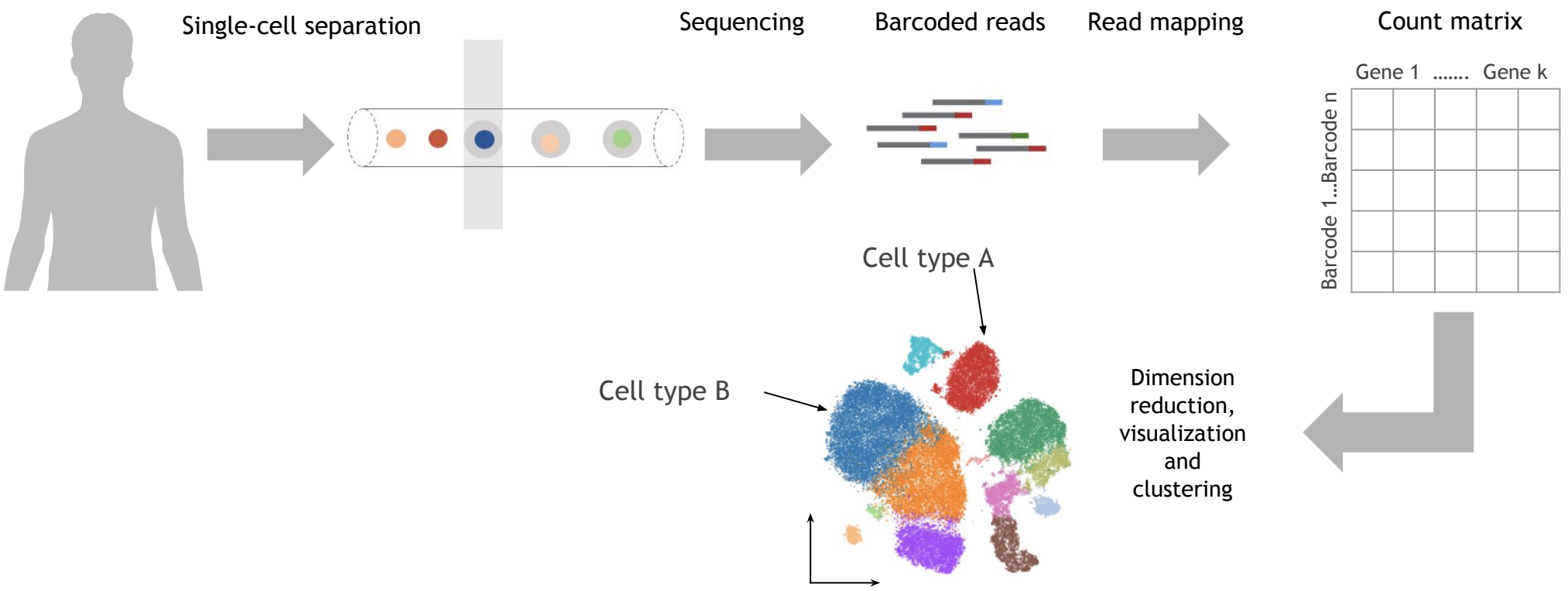
NVIDIA

Single-cell sequencing





Single-cell data analysis



Analytical challenges in single-cell data

Doublets & multiplets

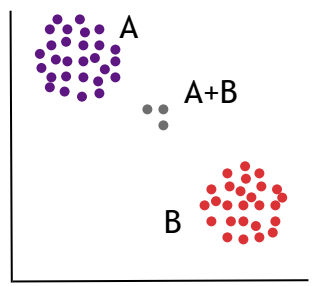
Library size

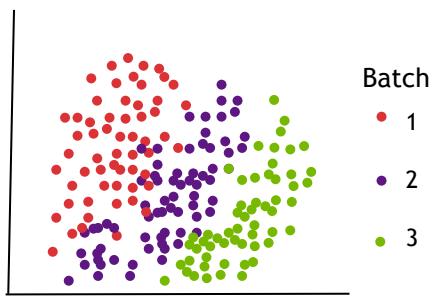
Batch effects

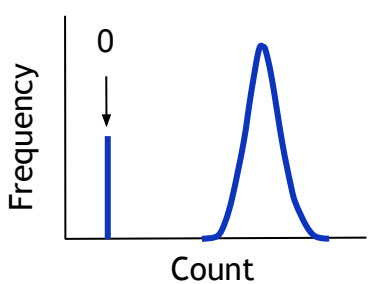
Noise

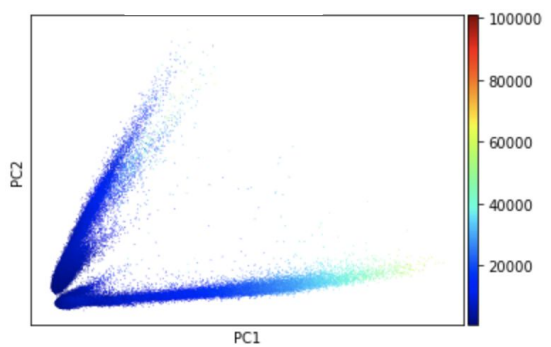
Dropout

Dimension reduction, clustering and differential expression

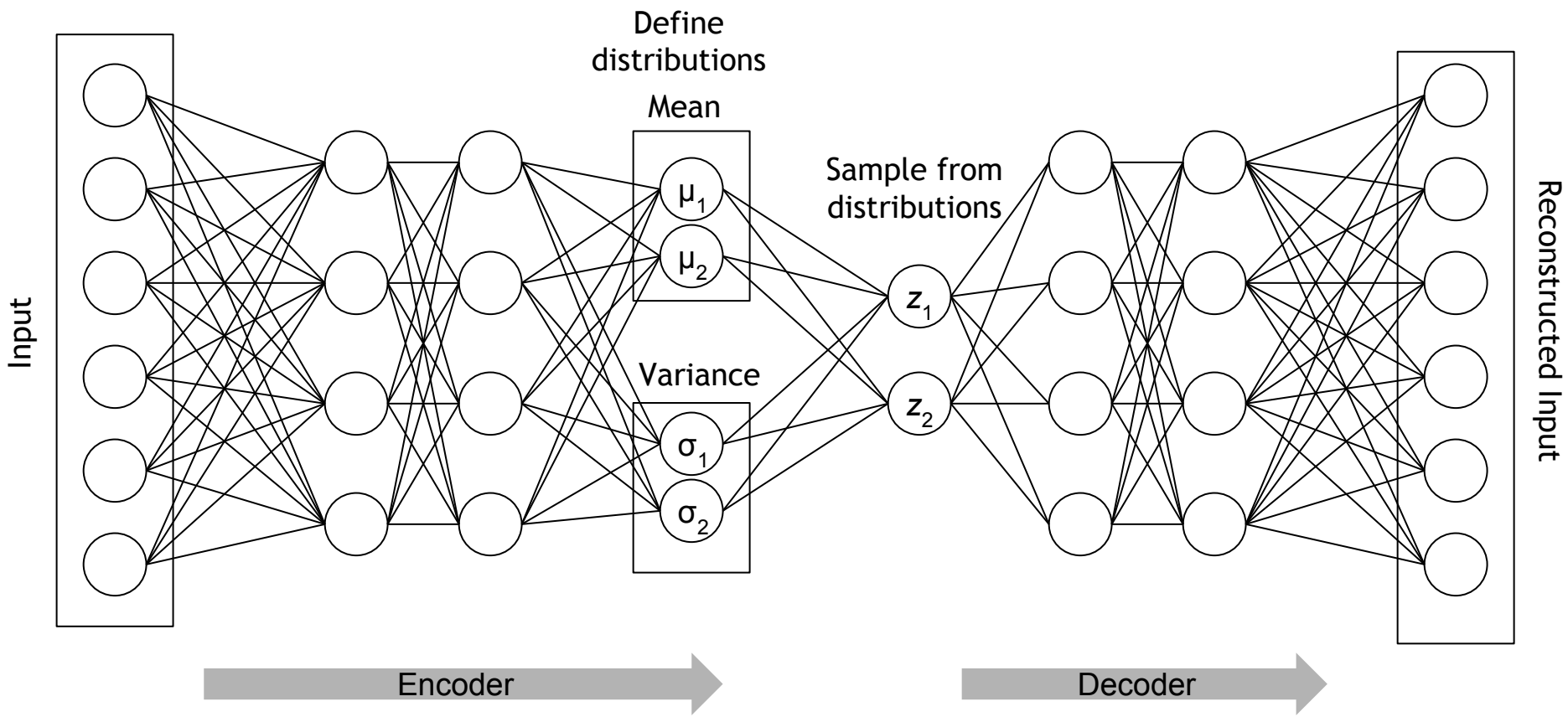




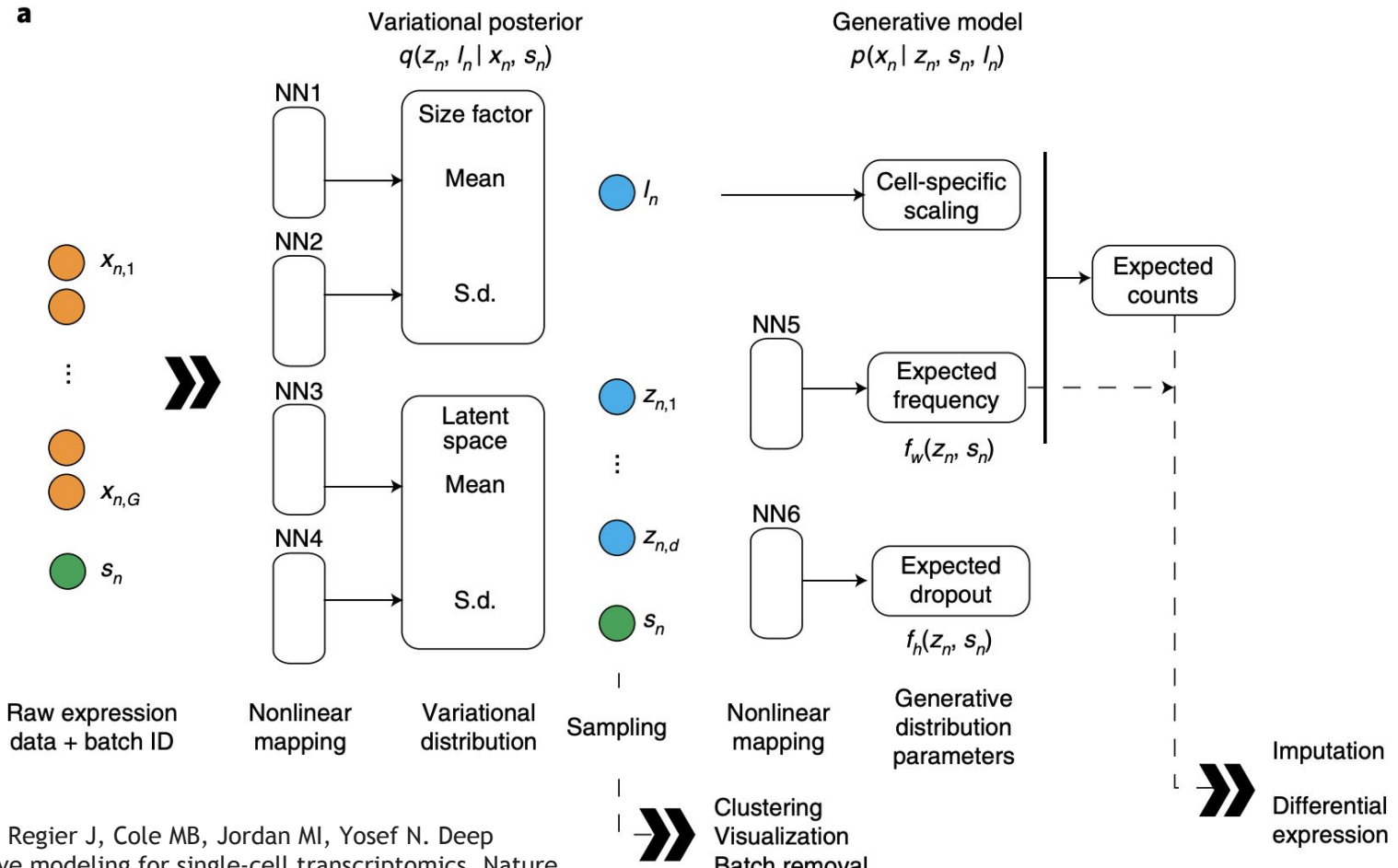


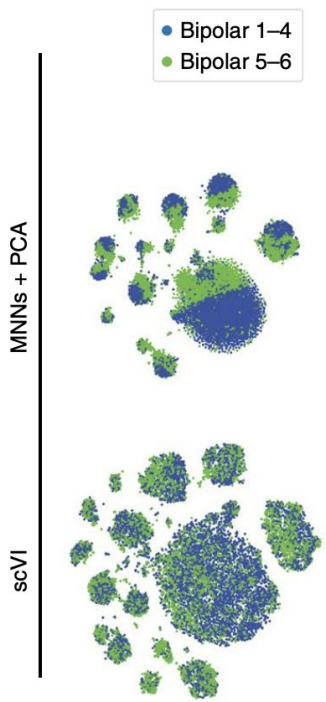


Variational Autoencoder (VAE)



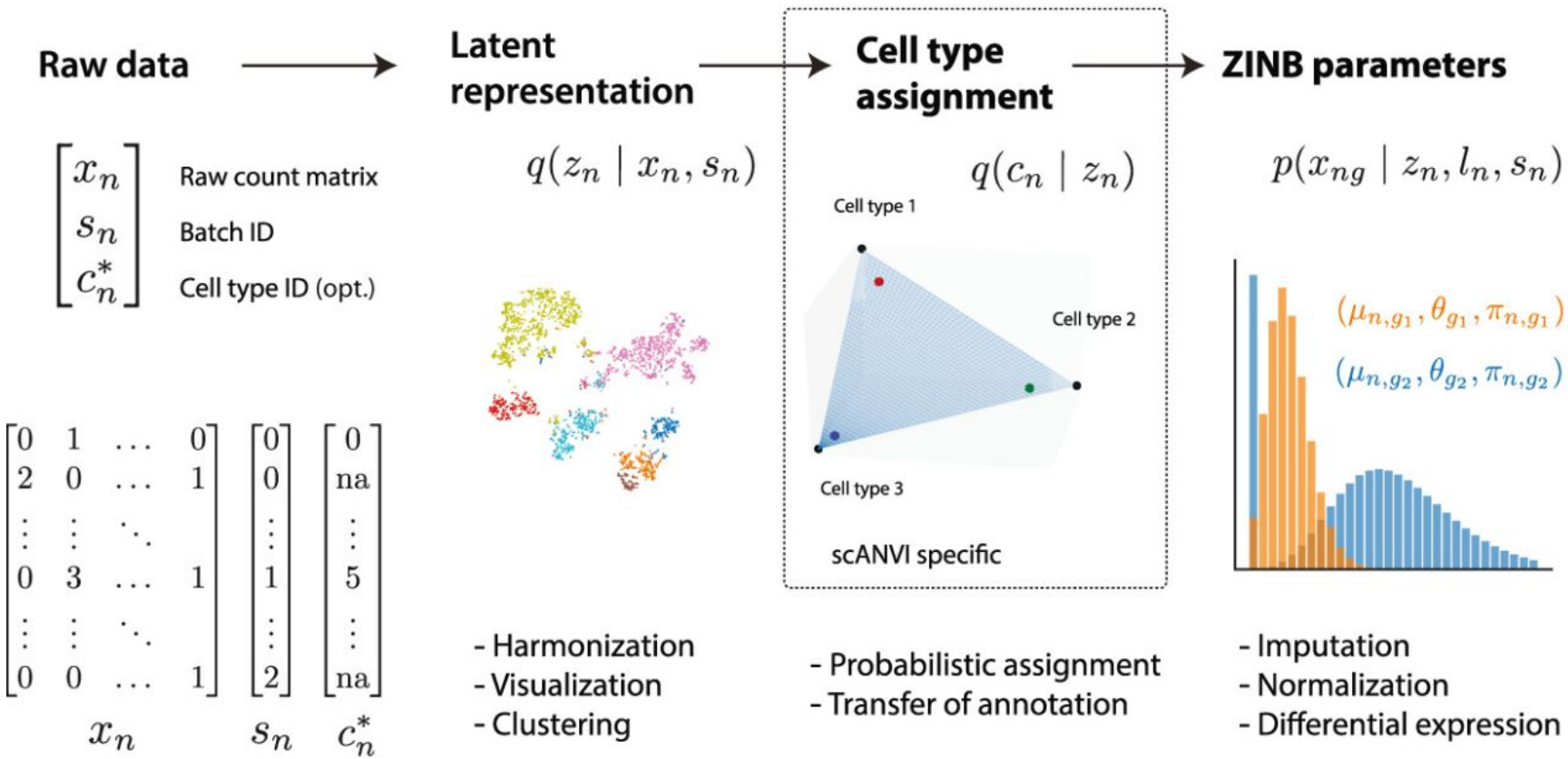
scVI





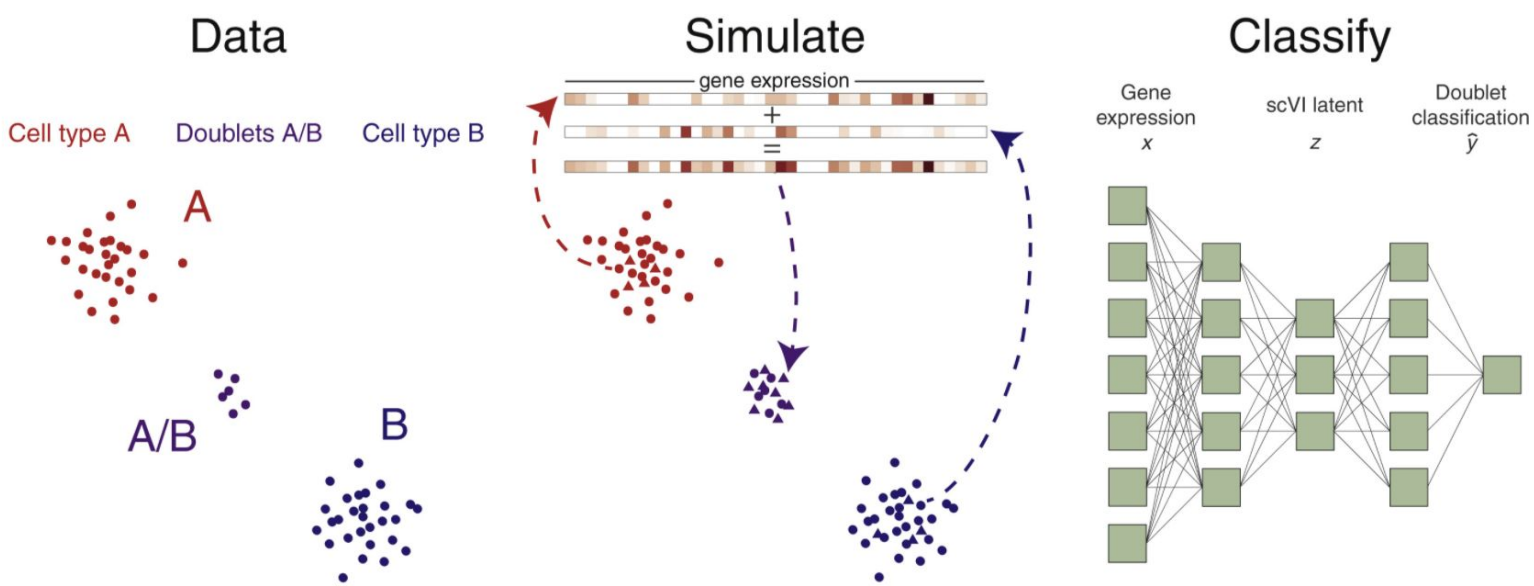
Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature methods. 2018 Dec;15(12):1053-8.

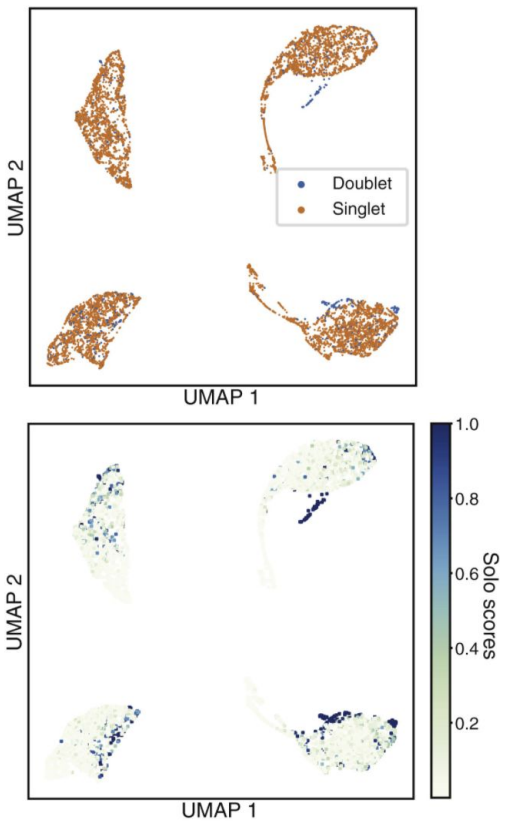
Cell type annotation



Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Molecular systems biology. 2021 Jan;17(1):e9620.

Doublet identification

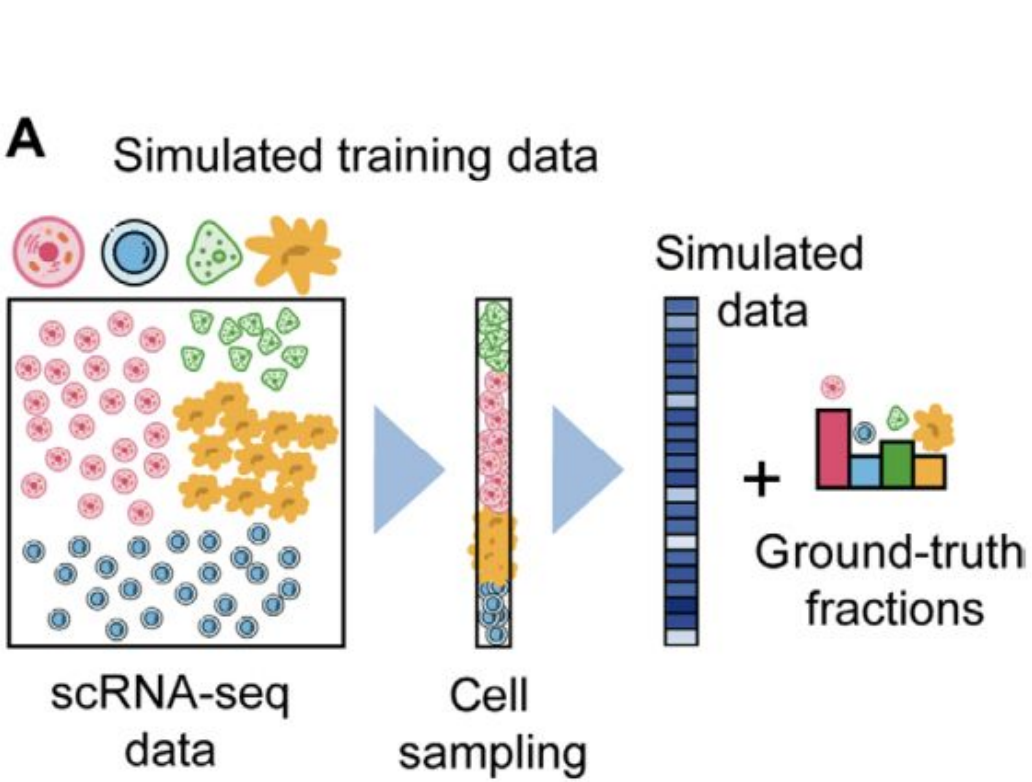


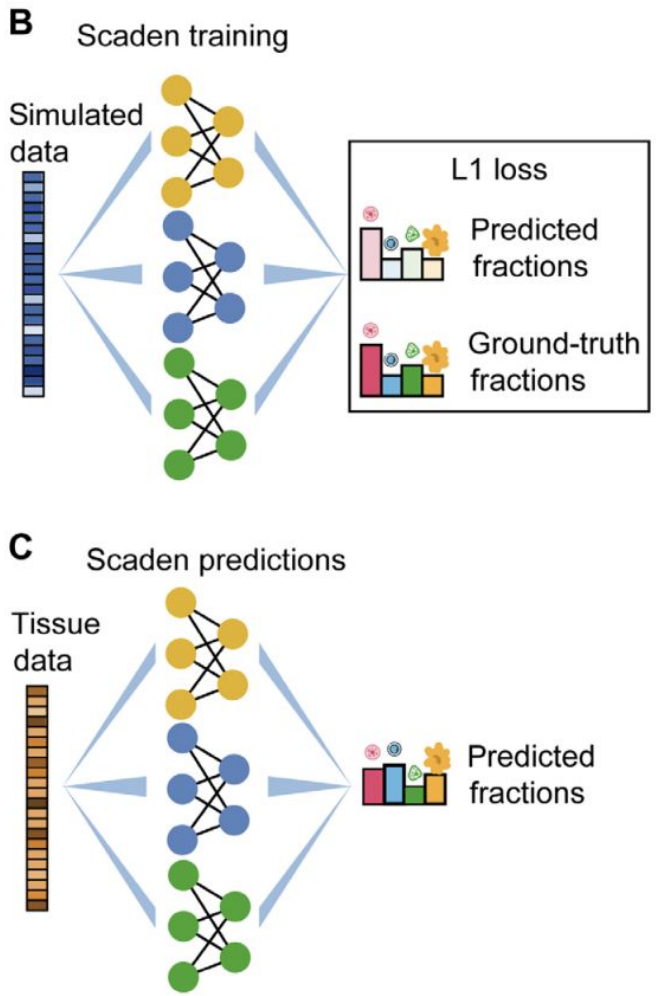


Bernstein NJ, Fong NL, Lam I, Roy MA, Hendrickson DG, Kelley DR. Solo: doublet identification in single-cell RNA-Seq via semi-supervised deep learning. Cell Systems. 2020 Jul 22;11(1):95-101.

NVIDIA.

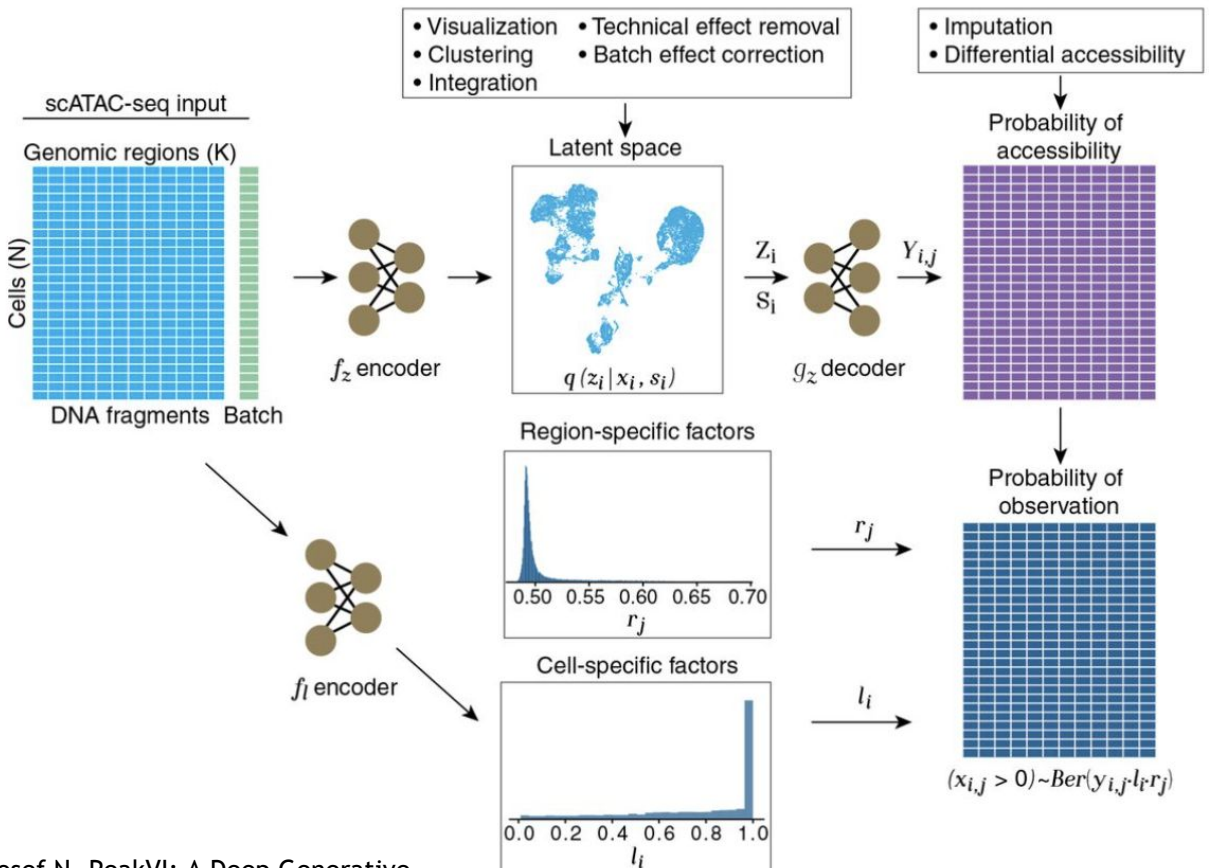
Bulk deconvolution





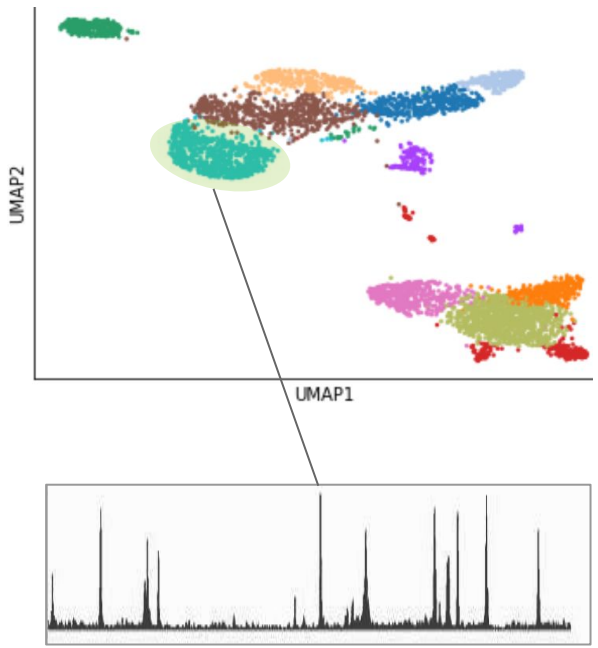
Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, Heutink P, Bonn S. Deep learning–based cell composition analysis from tissue expression profiles. Science advances. 2020 Jul 1;6(30):eaba2619.

VAEs for other single-cell modalities



Ashuach T, Reidenbach DA, Gayoso A, Yosef N. PeakVI: A Deep Generative Model for Single Cell Chromatin Accessibility Analysis. bioRxiv. 2021 Jan 1.

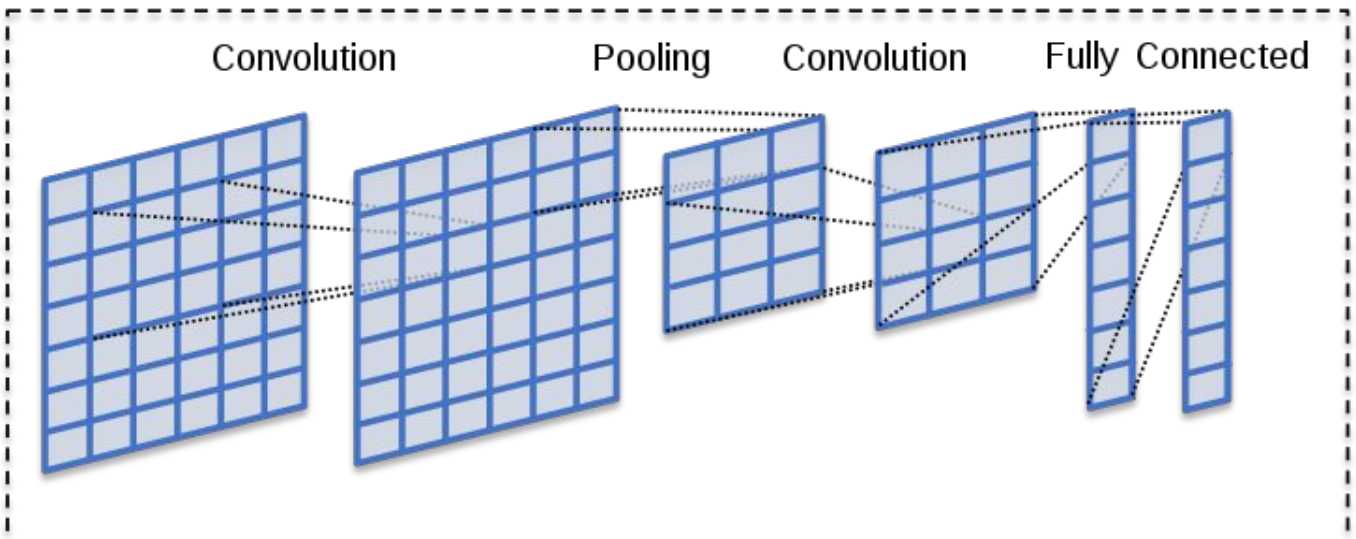
Modeling cell-type specific profiles

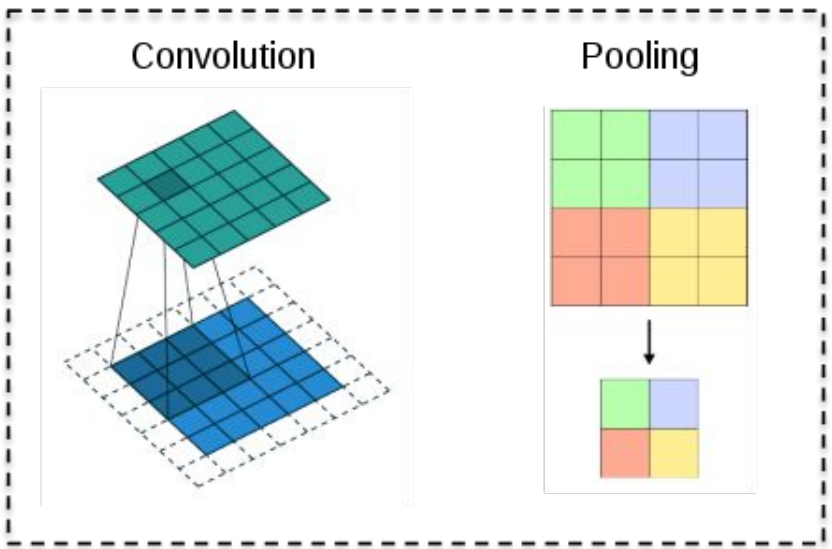


Given DNA sequence, can we predict gene expression and/or functional profiles in different cell types?

Based on these learned rules, can we predict the cell-type specific effect of sequence variation and mutations?

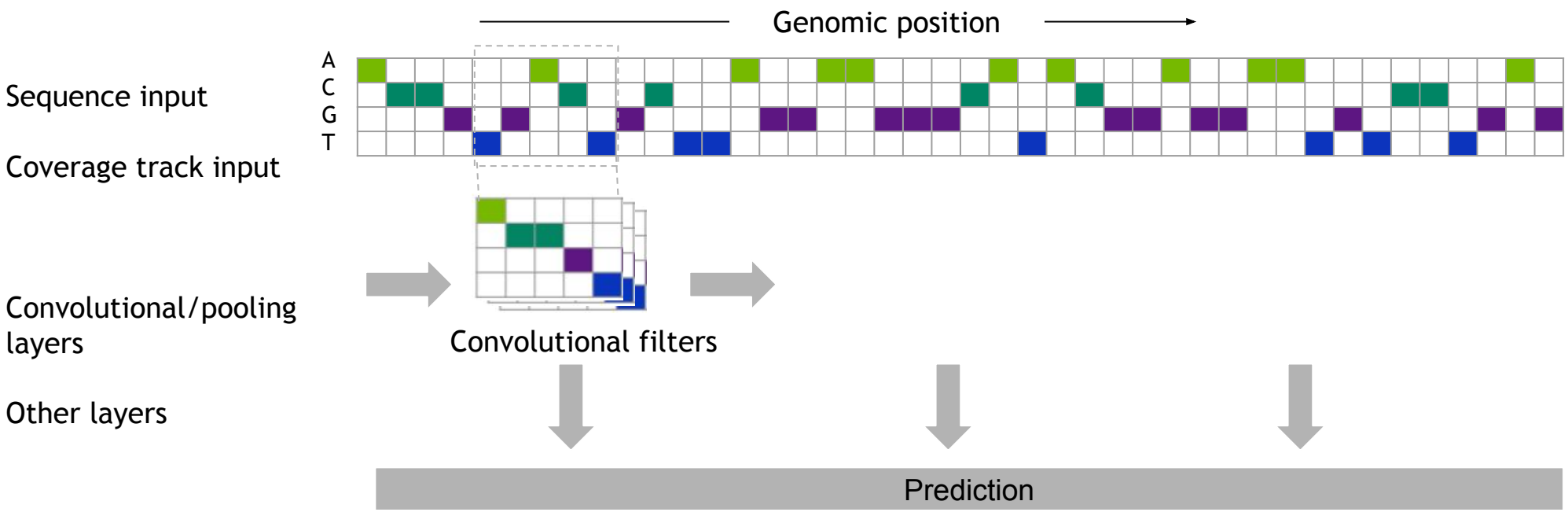
Convolutional neural networks



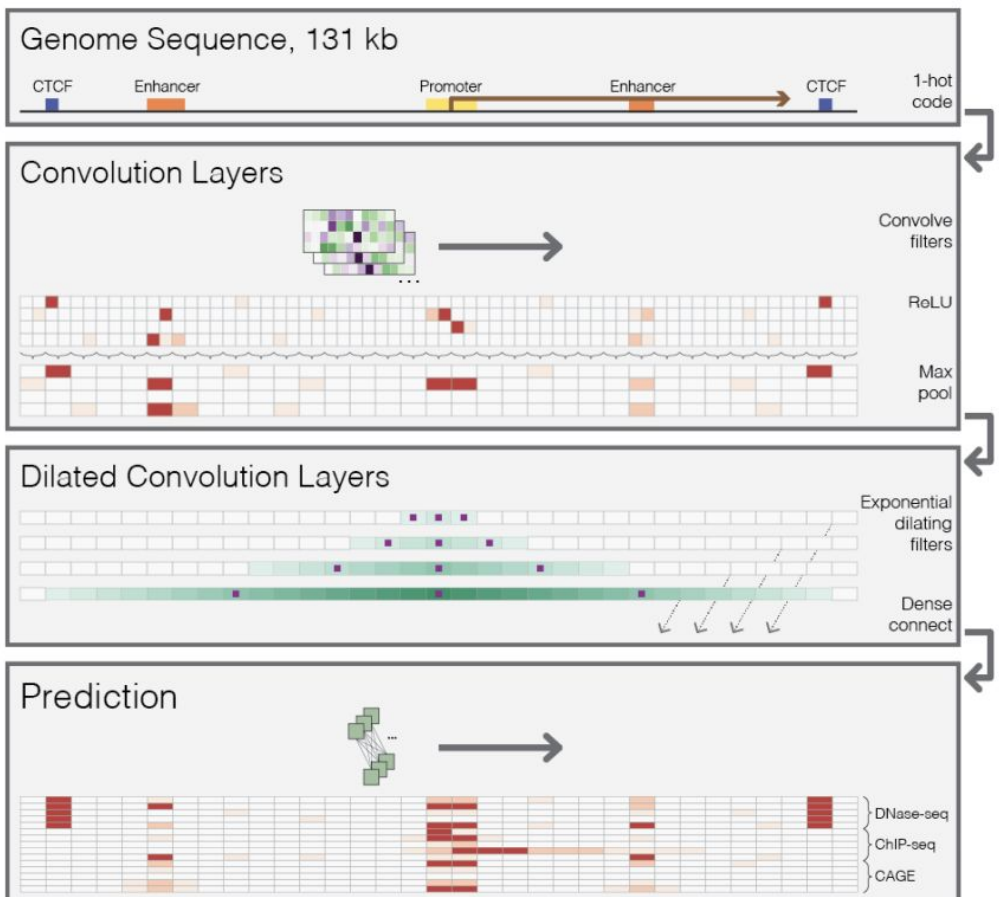


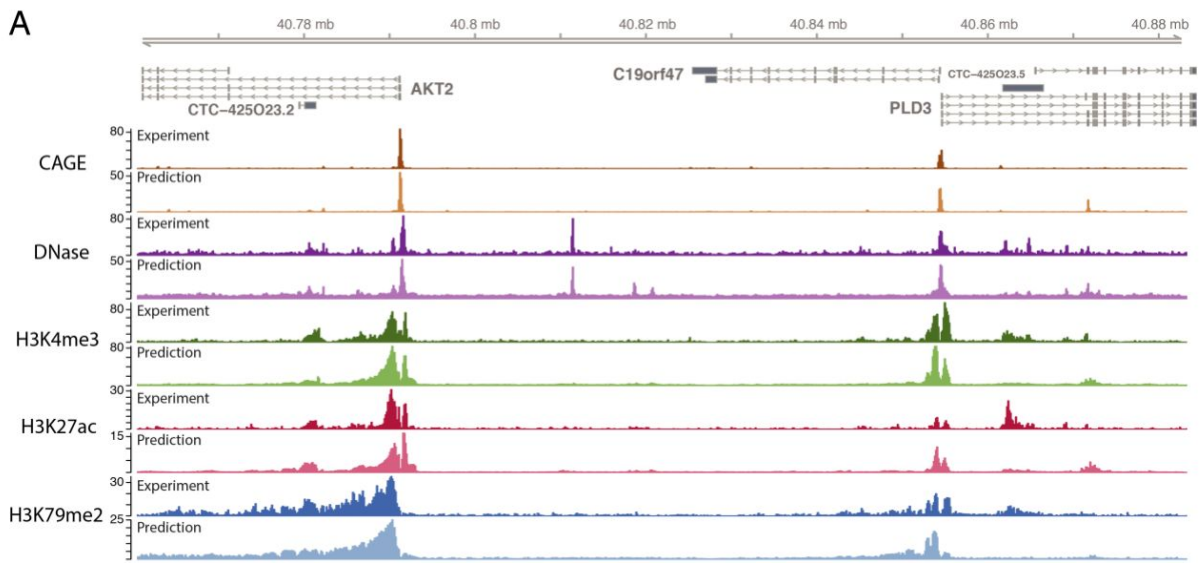
https://commons.wikimedia.org/wiki/File:ConvolutionAndPooling.svg

CNNs for genomics



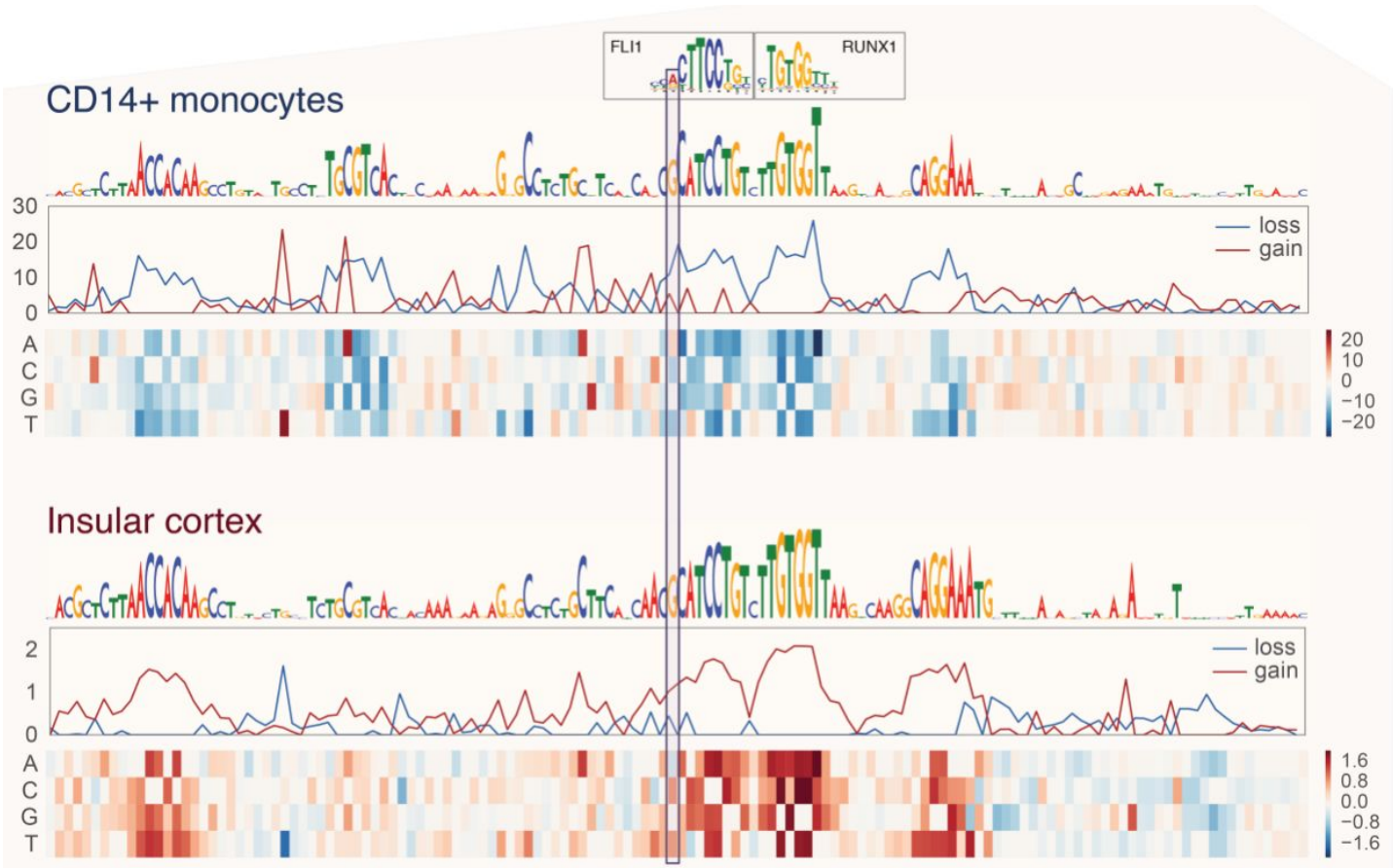
Basenji





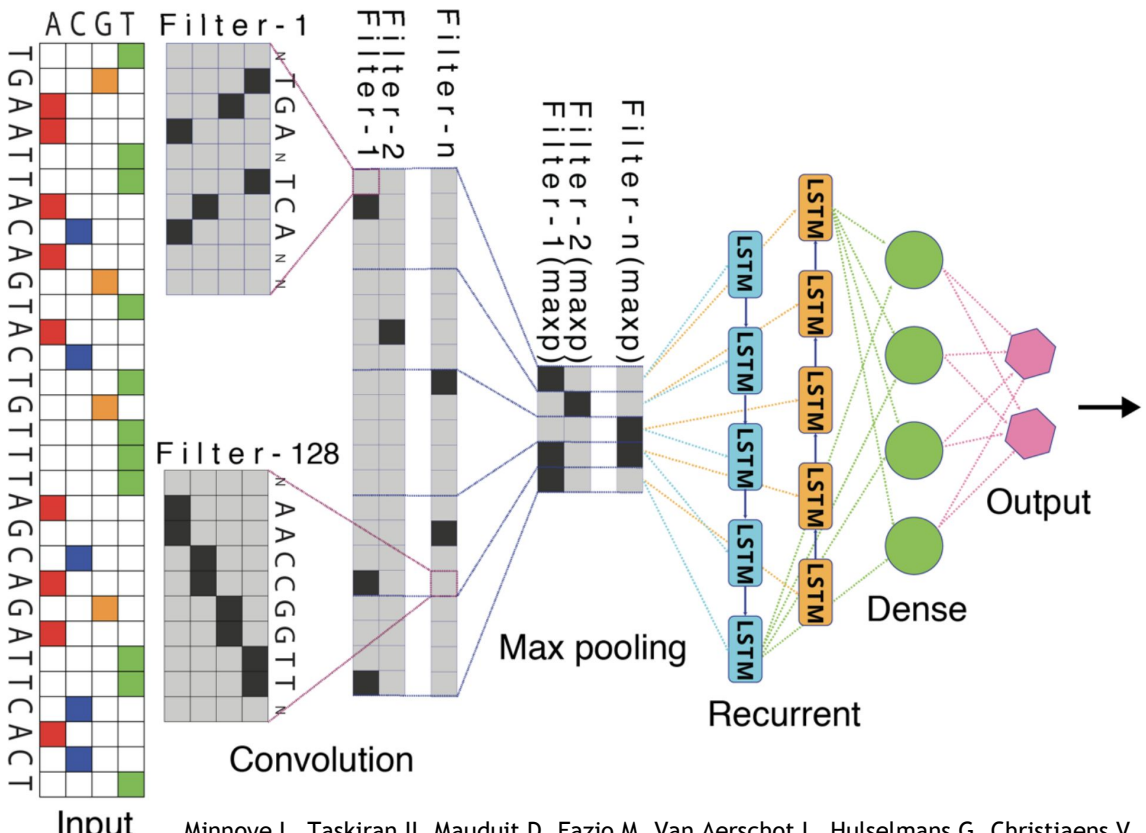
Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome research. 2018 May 1;28(5):739-50.

Predicting the impact of genetic variation

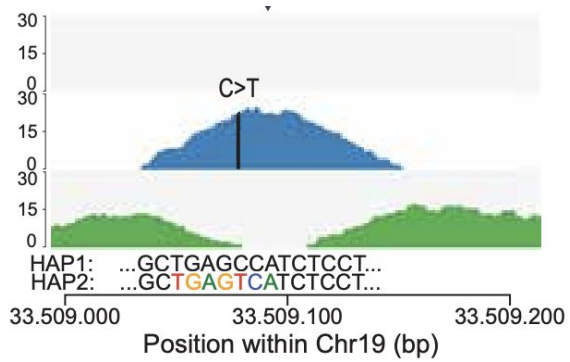


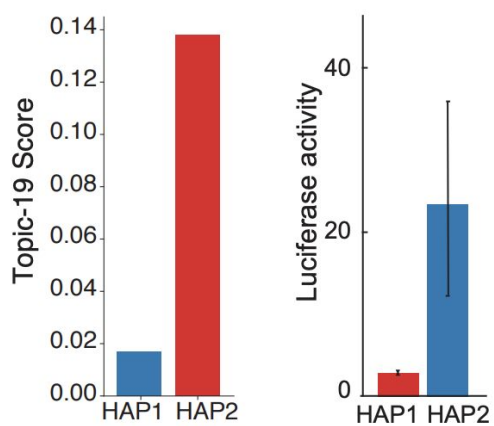
Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome research. 2018 May 1;28(5):739-50.

Predicting somatic mutation effects



Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, Christiaens V, Makhzami S, Seltenhammer M, Karras P, Primot A. Cross-species analysis of enhancer logic using deep learning. Genome research. 2020 Dec 1;30(12):1815-34.

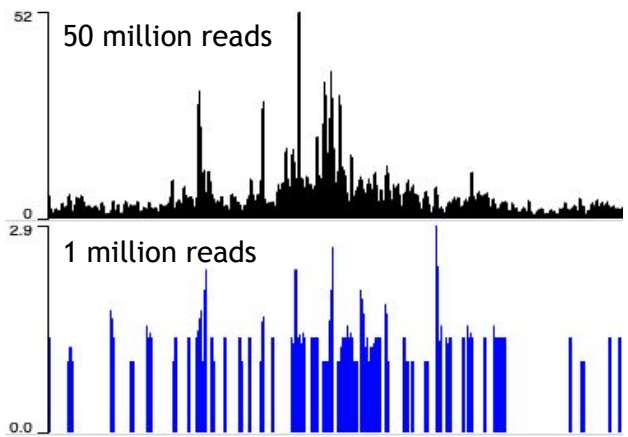




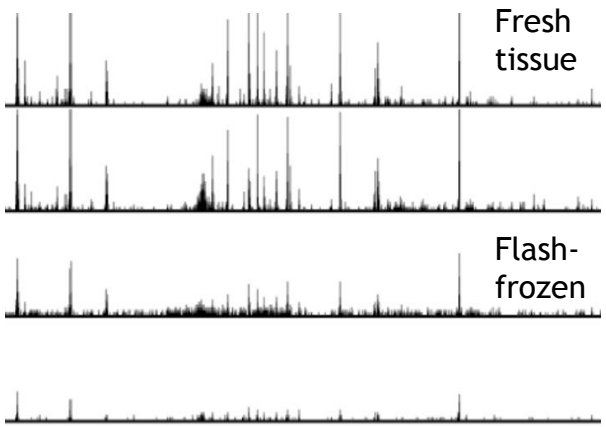
Atak ZK, Taskiran II, Demeulemeester J, Flerin C, Mauduit D, Minnoye L, Hulselmans G, Christiaens V, Ghanem GE, Wouters J, Aerts S. Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. Genome research. 2021 Jun 1;31(6):1082-96.

Epigenomic data quality

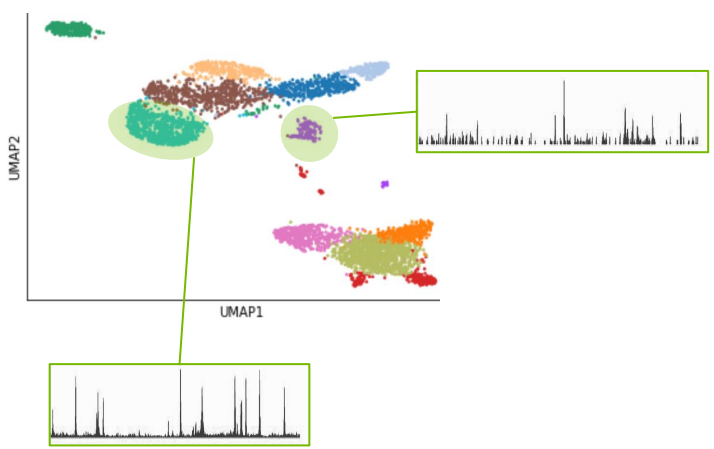
Low sequencing depth



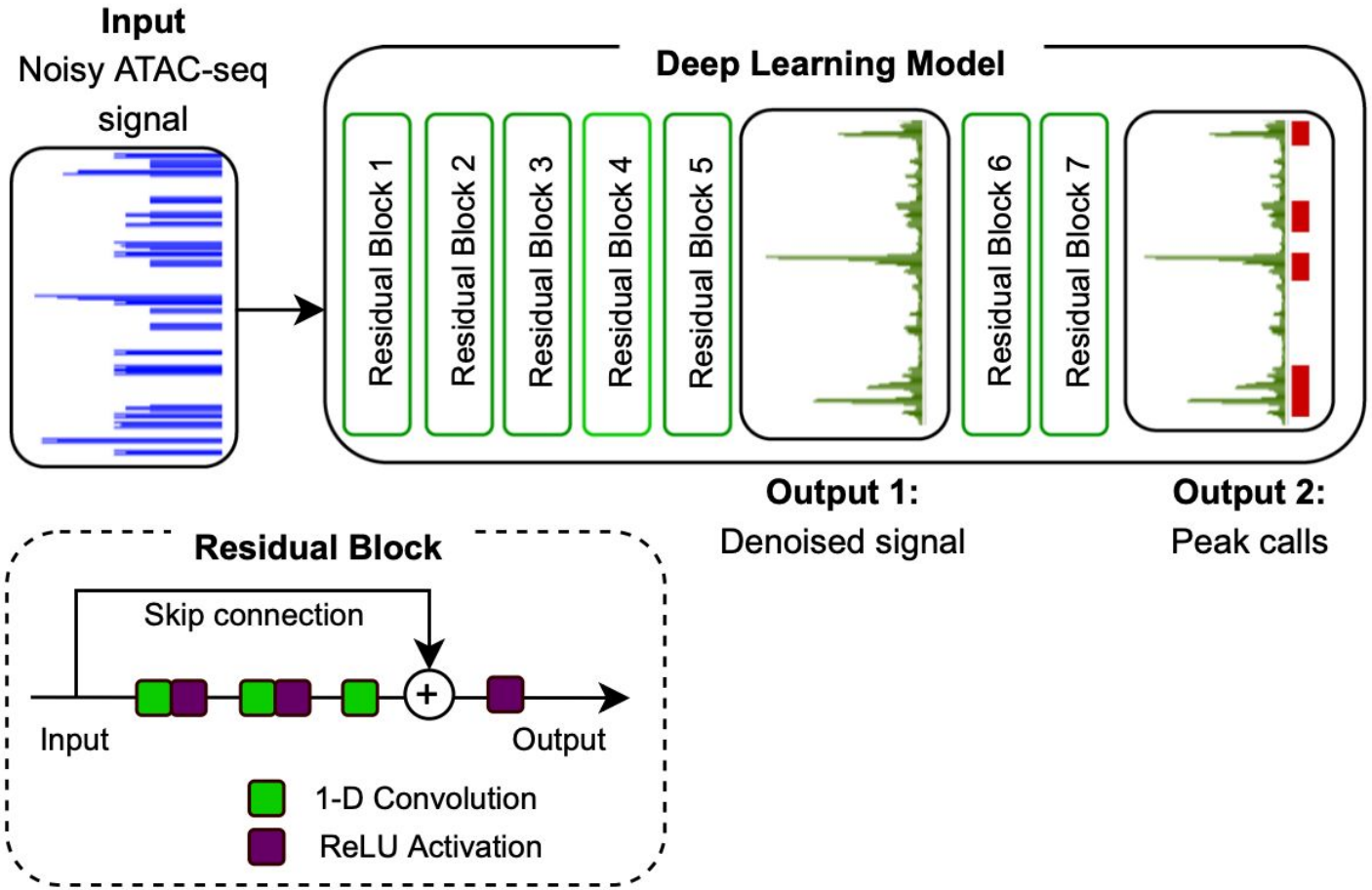
Sample/experimental factors



Low aggregate cell count

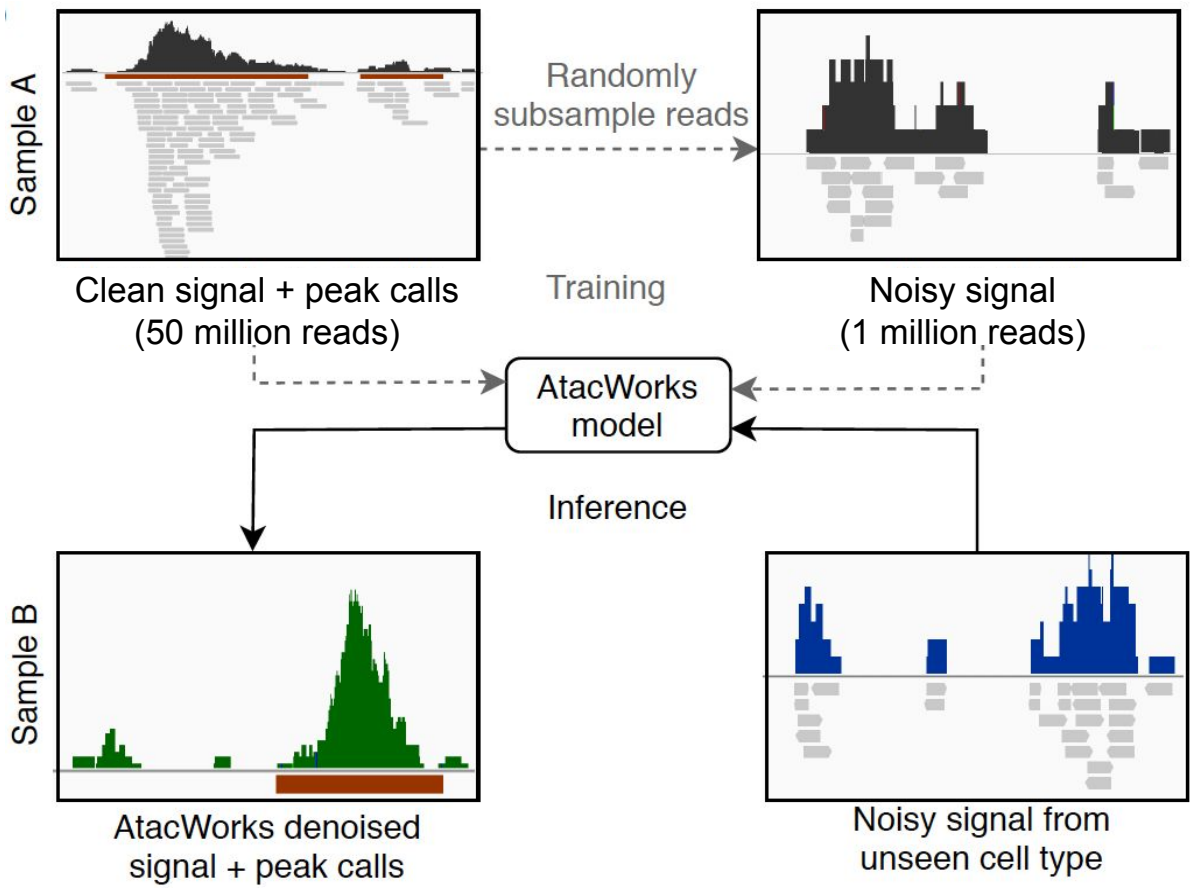


AtacWorks

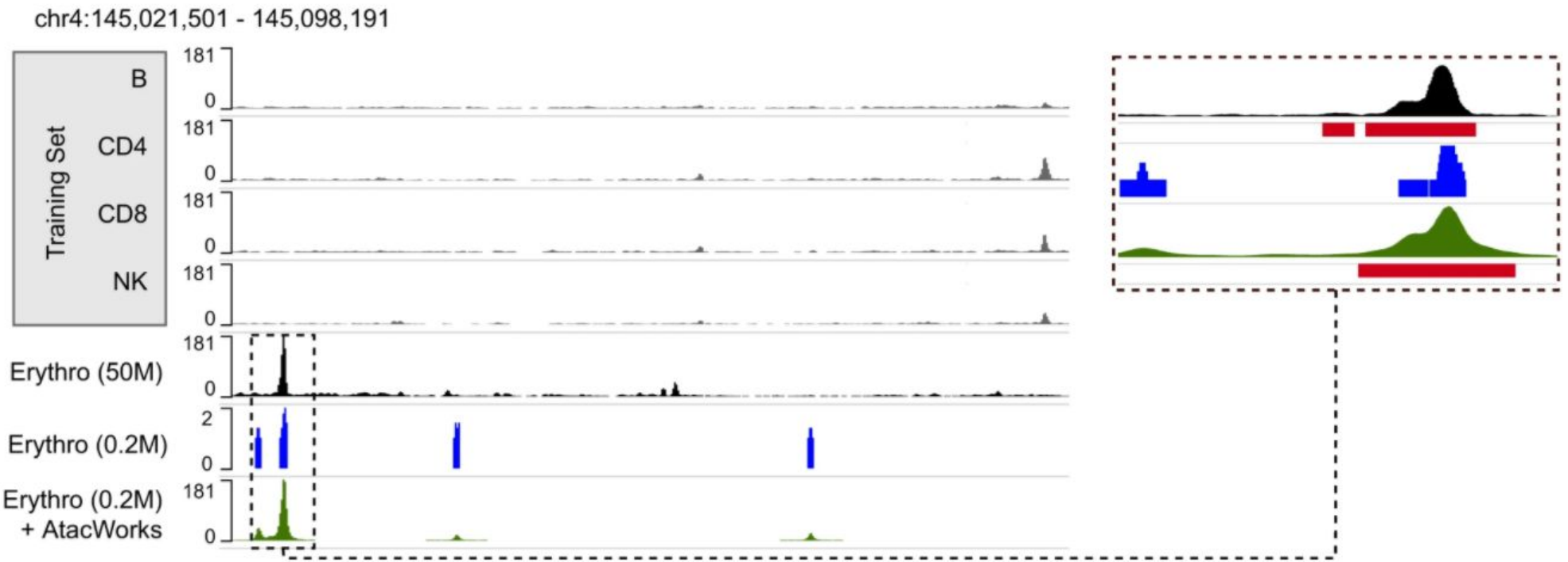


Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

Training models to enhance low-coverage ATAC-seq

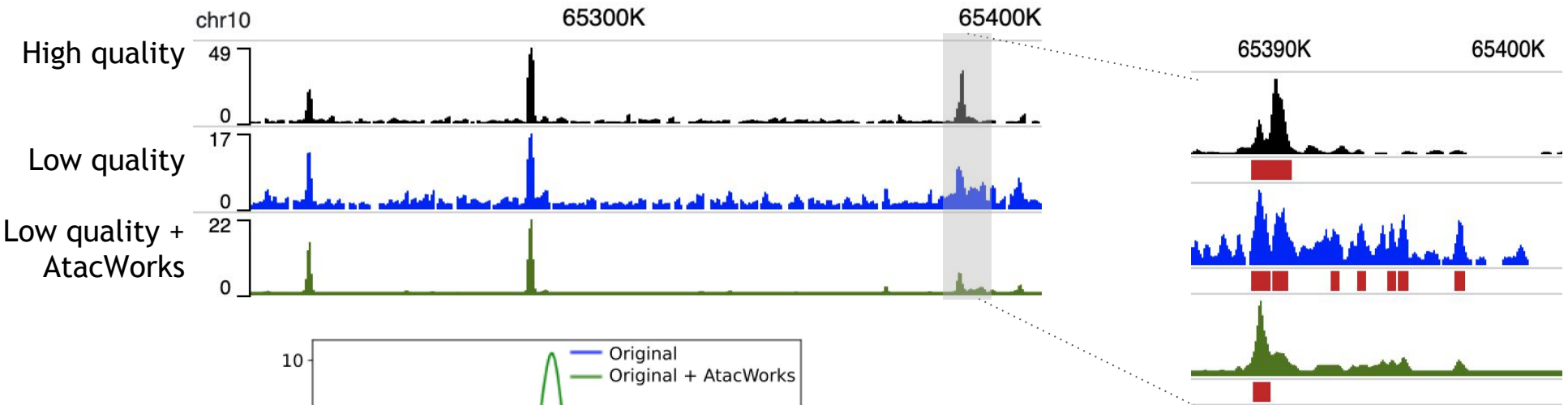


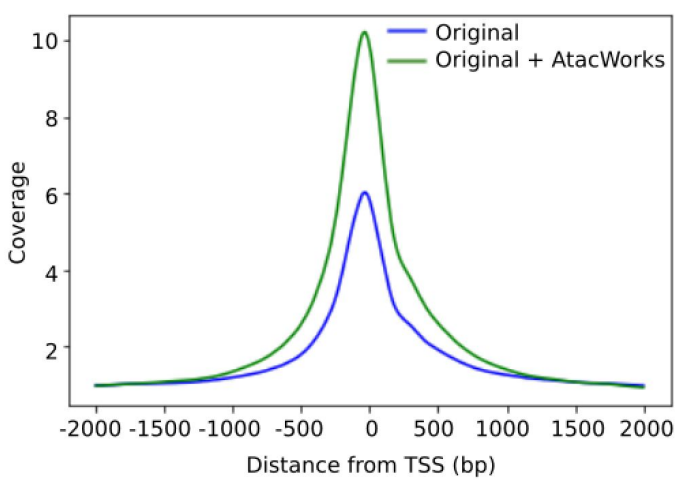
Denoising and peak calling from low-coverage ATAC-seq



Rescuing low-quality data

Bulk ATAC-seq data from human Erythroblasts

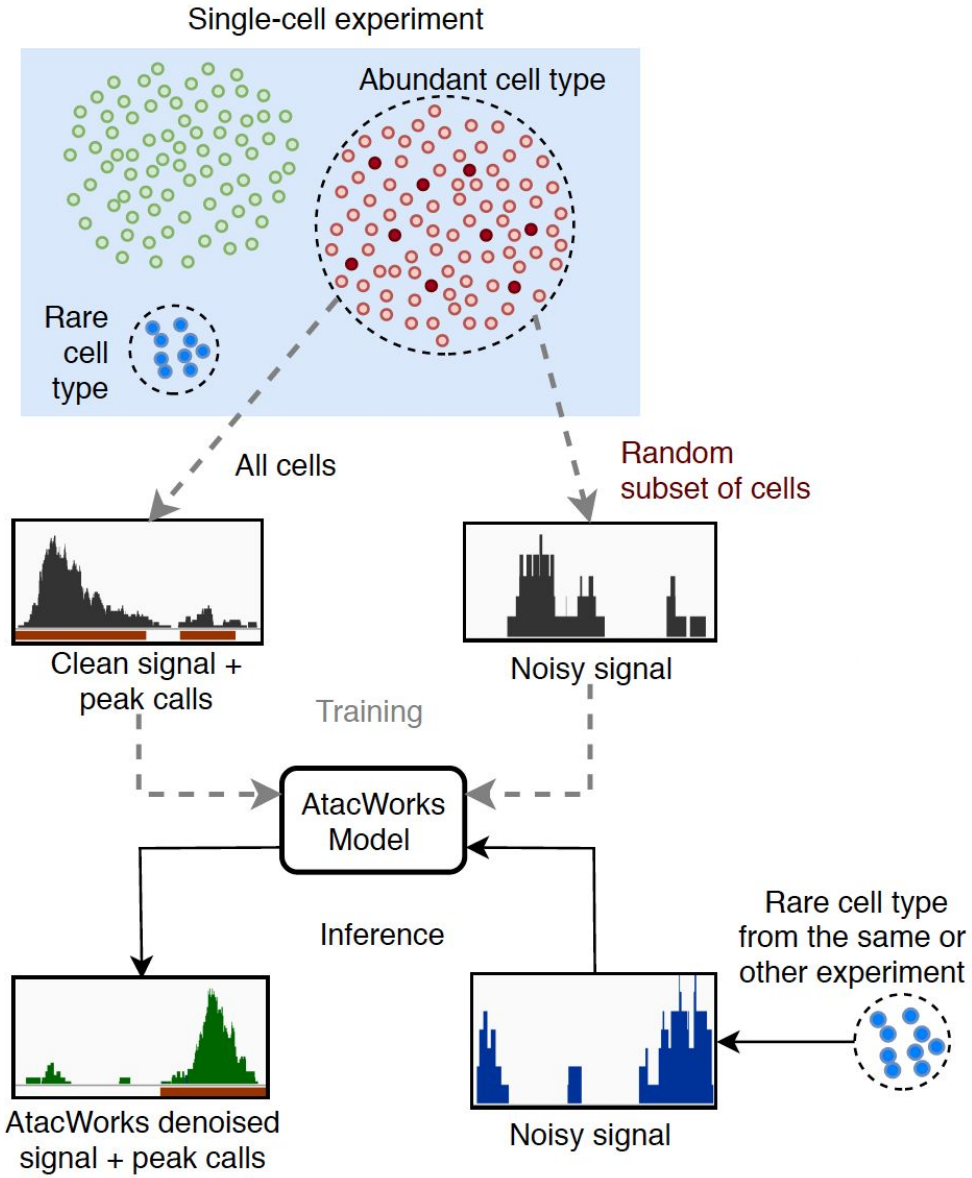




Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

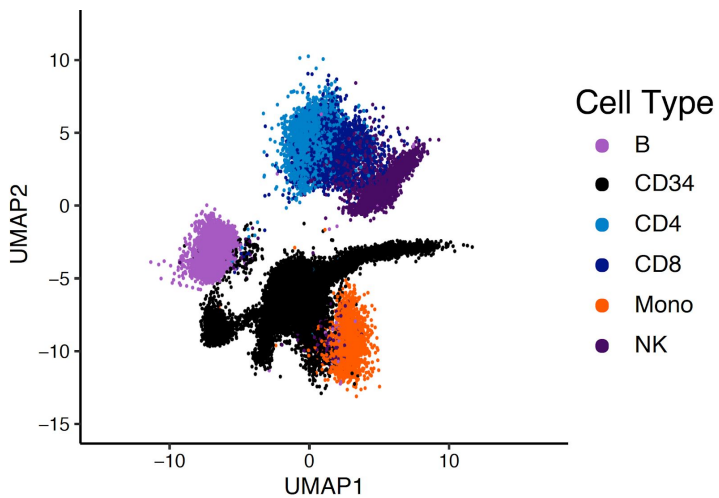
NVIDIA

Applying AtacWorks to scATAC-seq

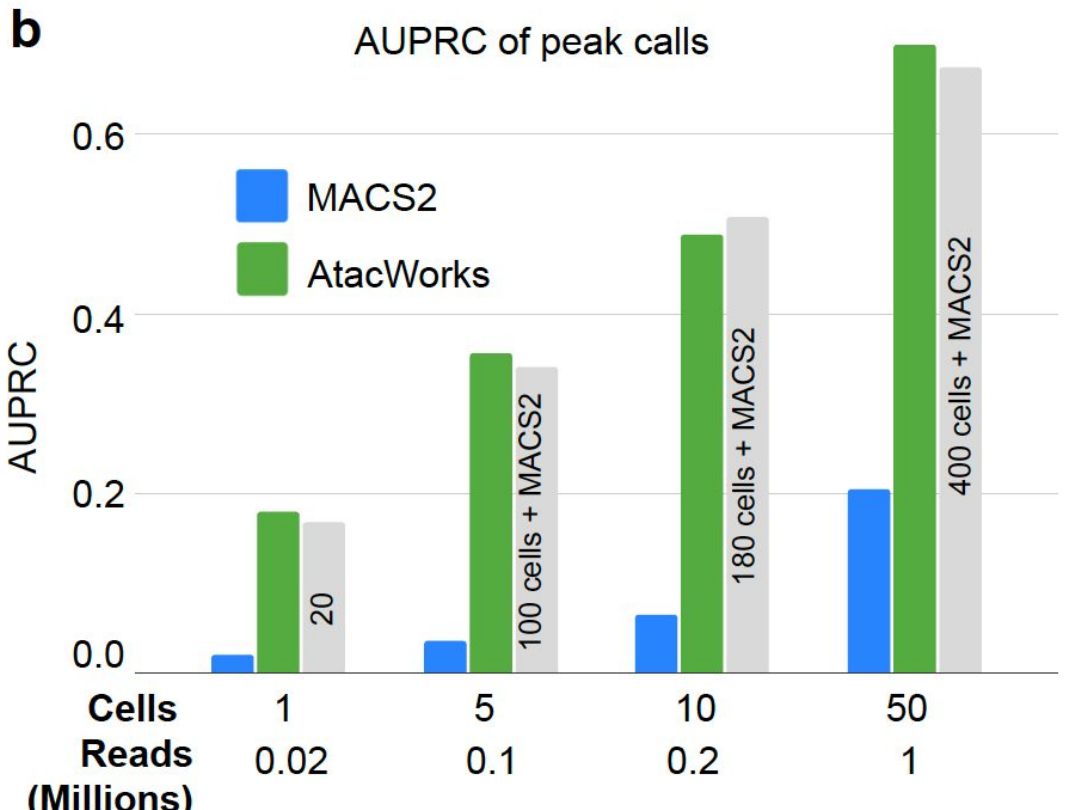


Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

Increasing single-cell resolution

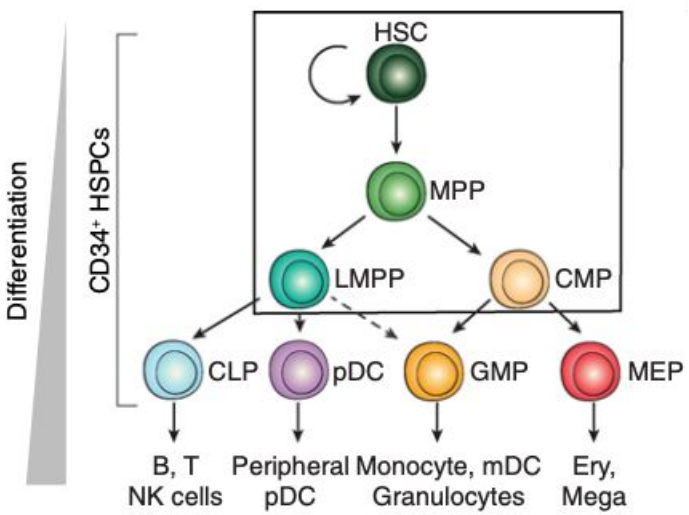


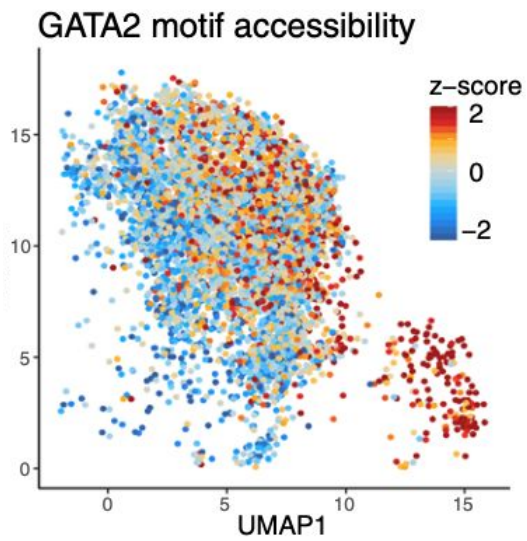
AtacWorks can obtain the same quality results from ~10x fewer cells, increasing the resolution of single-cell chromatin accessibility profiling by an order of magnitude.

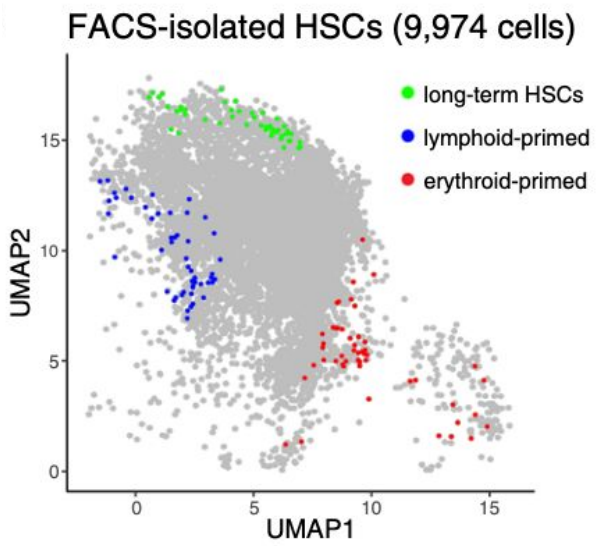


Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

Identifying regulatory DNA associated with lineage priming

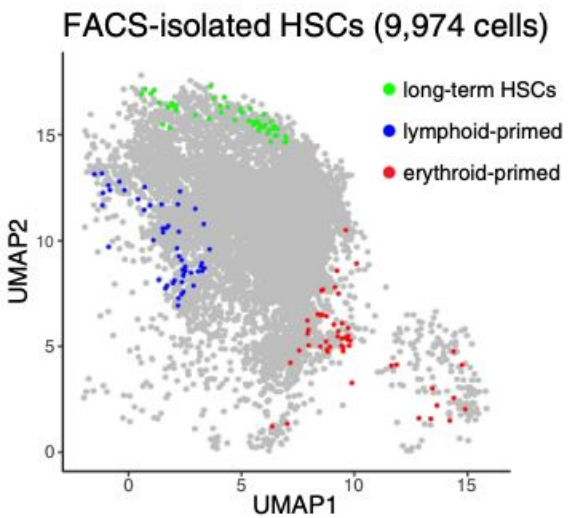


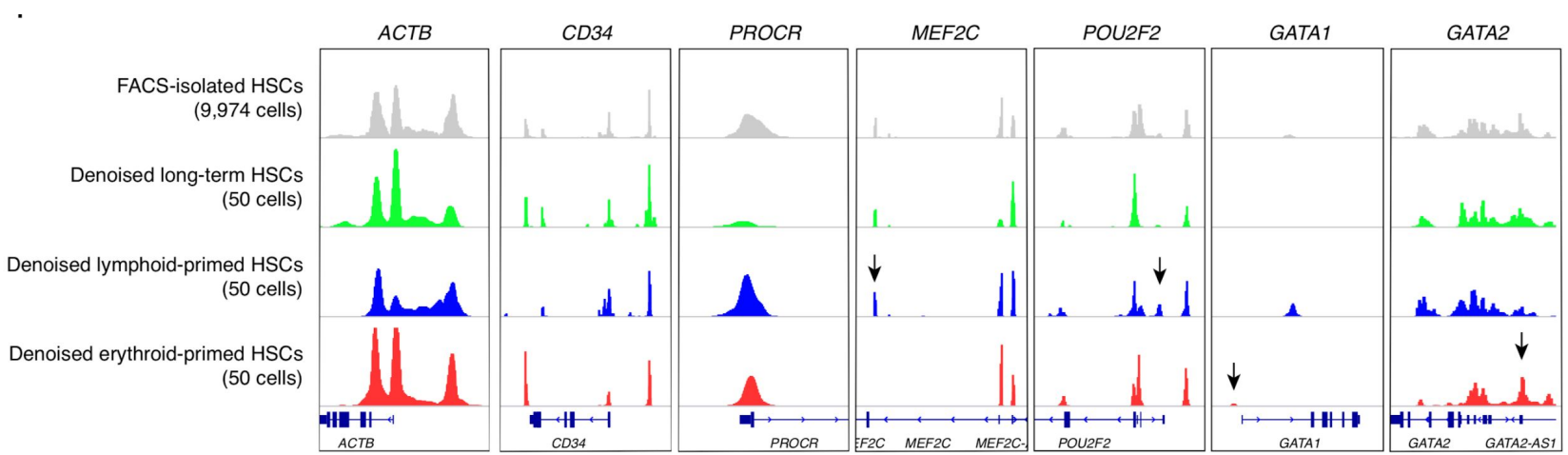




Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

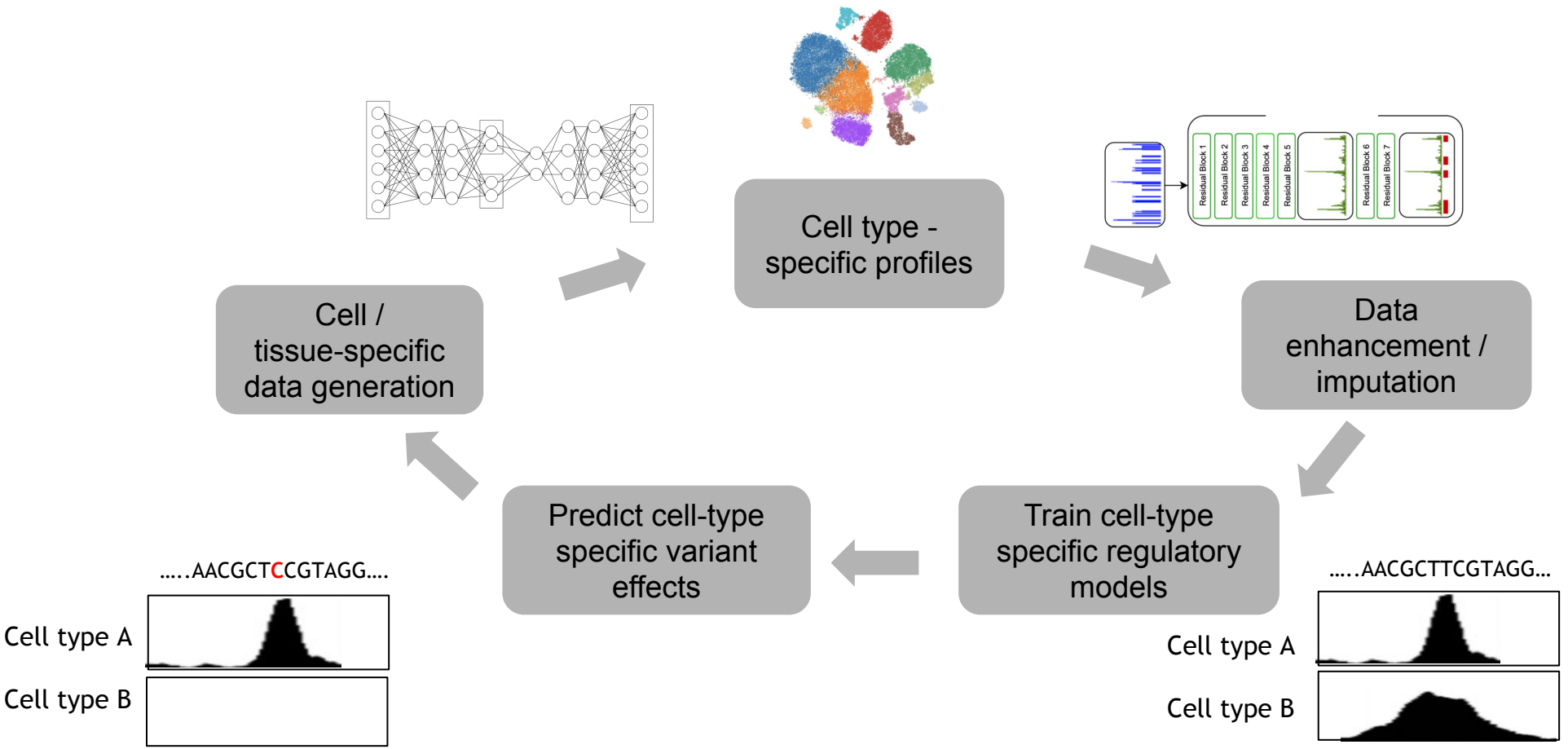
Identifying regulatory DNA associated with lineage priming





Lal, A., Chiang, Z.D., Yakovenko, N. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun 12, 1507 (2021).

Regulatory modeling with deep learning




NVIDIA