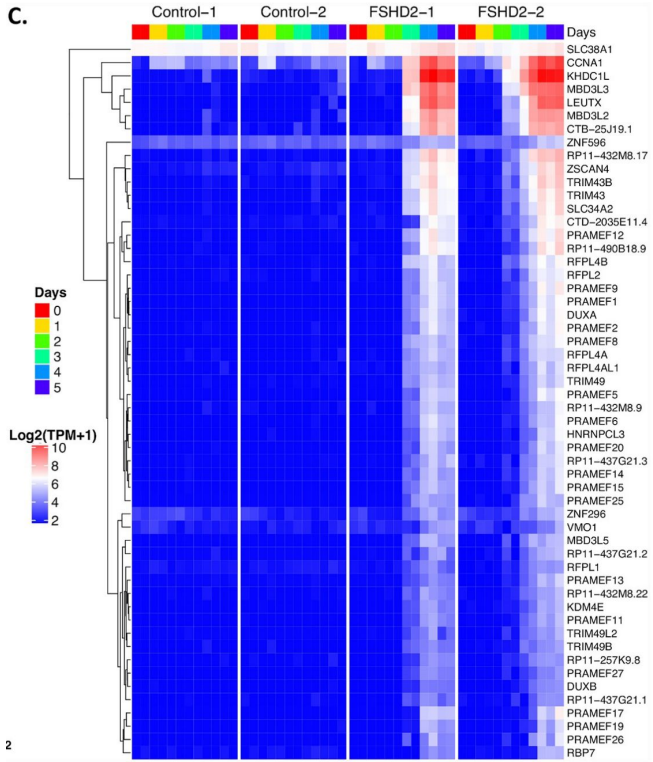
Intro to genomics visualizations

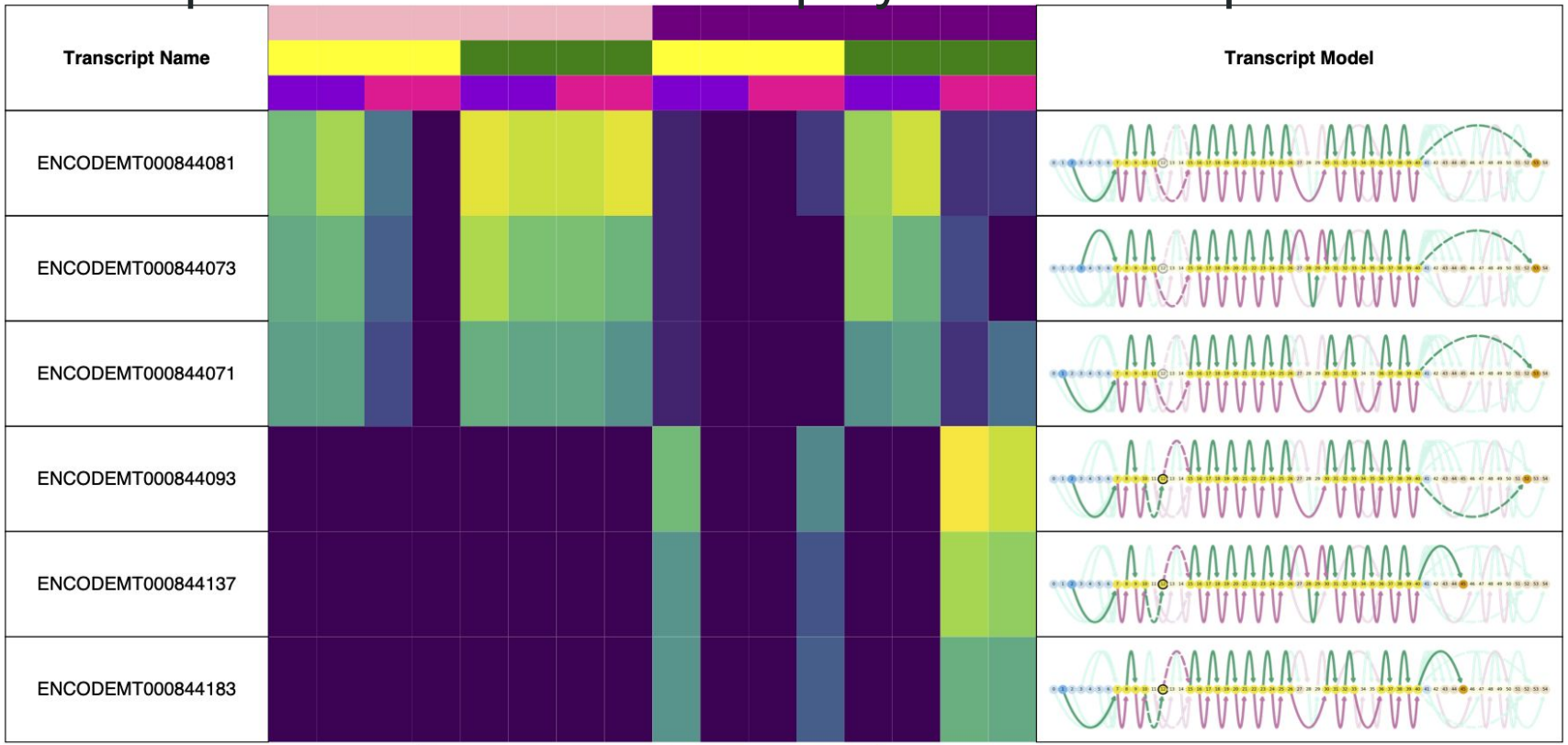
COSMOS 2021
7/13/21

Heatmaps are often used to display trends in expression or signal data

- Displays normalized expression or signal information (ie. TPM, log2(TPM), accessibility signal etc.)
- Often row or column-normalized (minimum and maximum color in each sample are used to normalize values in that row)
- Expression profiles often hierarchically clustered and displayed with dendrogram to order genes in an informative way



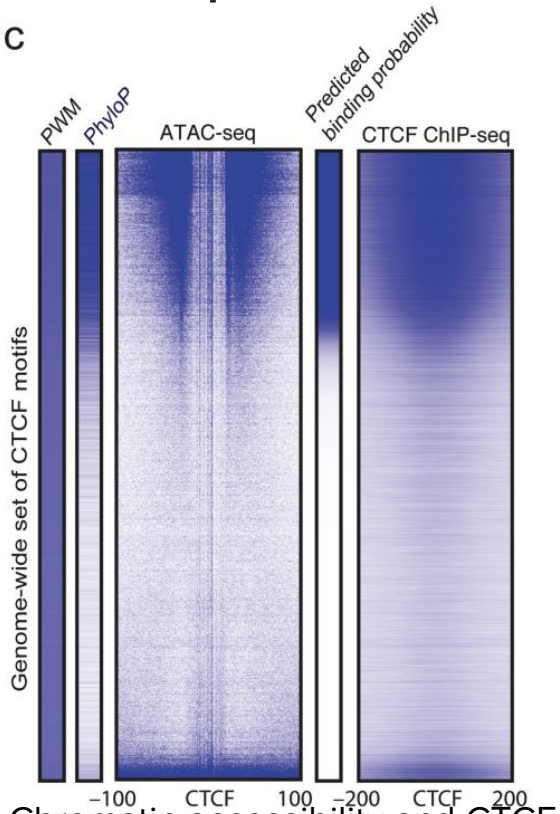
Heatmaps are often used to display trends in expression data



5xFAD
WT
Female
Male
Hippocampus
Cortex

Heatmaps are often used to display trends in expression or signal data

- Displays normalized expression information (ie. TPM, log2(TPM), accessibility signal etc.)
- Often row or column-normalized (minimum and maximum color in each sample are used to normalize values in that row)
- Expression profiles often hierarchically clustered and displayed with dendrogram to order genes in an informative way

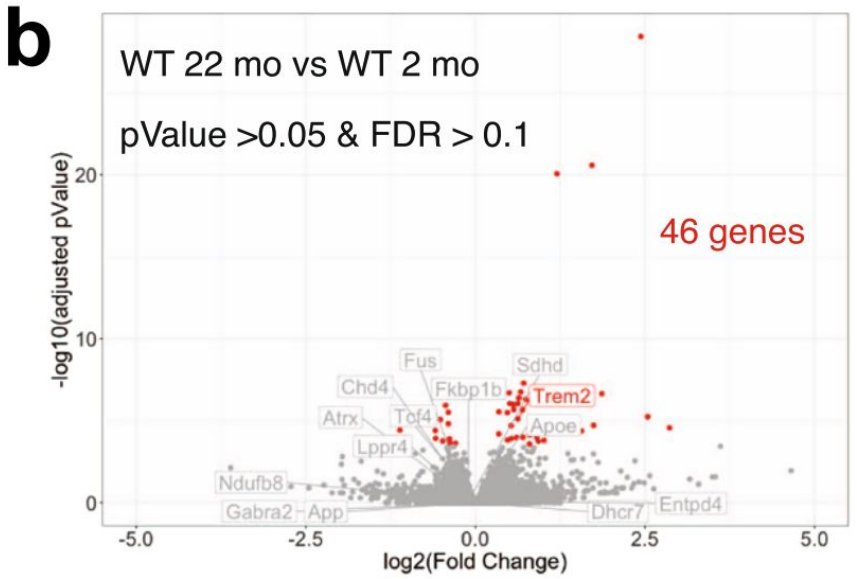


Chromatin accessibility and CTCF ChIP-seq signal genome-wide

Buenrostro et. al., 2013 (https://doi.org/10.1038/nmeth.2688)

Volcano plots are used to visualize differences in signal or expression between 2 conditions

- Used to visualize significance and fold change of signal or expression
- Often used for visualizing the results of differential gene expression testing



Volcano plot of differentially-expressed genes between 2 month and 22 month old mouse

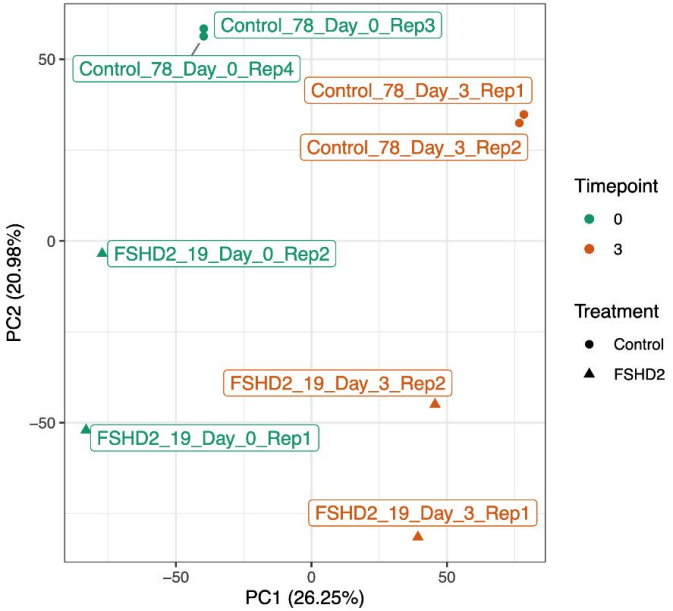
Baglietto-Vargas et al., 2021 (https://doi.org/10.1038/s41467-021-22624-z)

Lower dimensional representations of expression or signal data can reveal high dimensional similarity between datasets

- High dimensional data can include expression data or signal data
- PCA - principal component analysis
- tSNE - t-distributed stochastic neighbor embedding
- UMAP - uniform manifold approximation and projection

Lower dimensional representations of expression or signal data can reveal high dimensional similarity between datasets

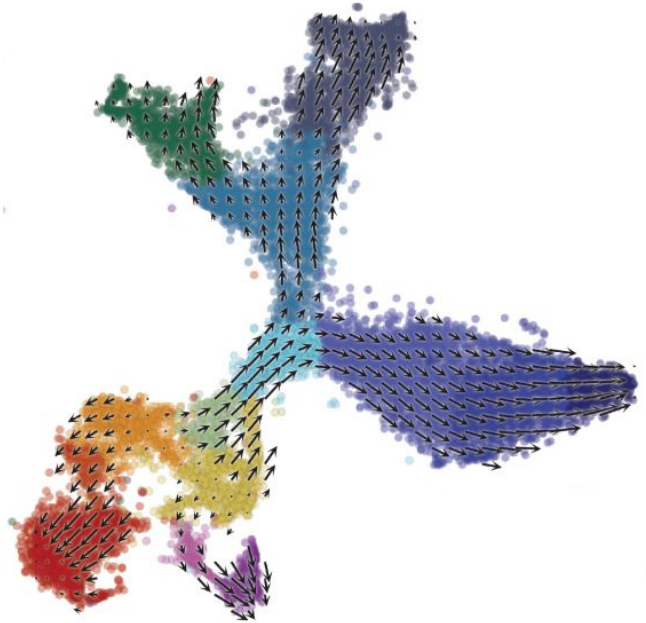
- High dimensional data can include expression data or signal data
- **PCA - principal component analysis**
- tSNE - t-distributed stochastic neighbor embedding
- UMAP - uniform manifold approximation and projection



PCA representation of multiple bulk RNA-seq gene expression datasets

Lower dimensional representations of expression or signal data can reveal high dimensional similarity between datasets

- High dimensional data can include expression data or signal data
- PCA - principal component analysis
- **tSNE - t-distributed stochastic neighbor embedding**
- UMAP - uniform manifold approximation and projection

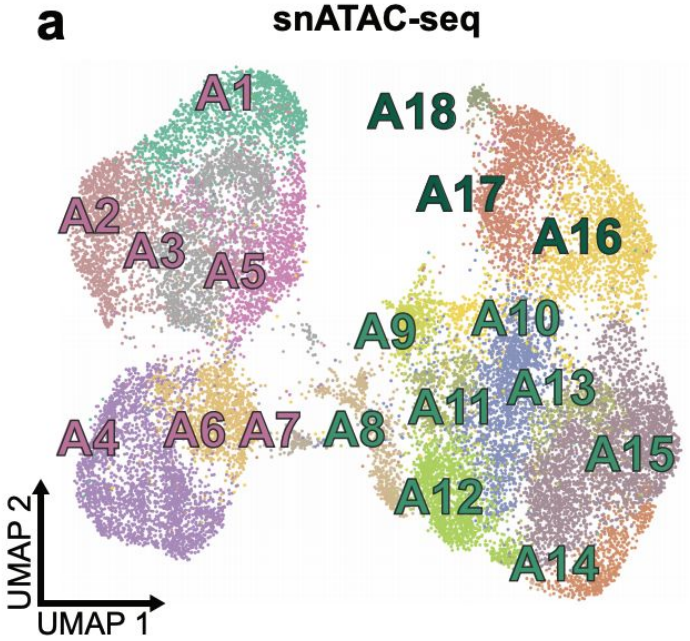


tSNE representation gene expression profiles in single-cell RNA-seq in developing brain

La Manno et. al., 2018 (https://doi.org/10.1038/s41586-018-0414-6)

Lower dimensional representations of expression or signal data can reveal high dimensional similarity between datasets

- High dimensional data can include expression data or signal data
- PCA - principal component analysis
- tSNE - t-distributed stochastic neighbor embedding
- **UMAP - uniform manifold approximation and projection**

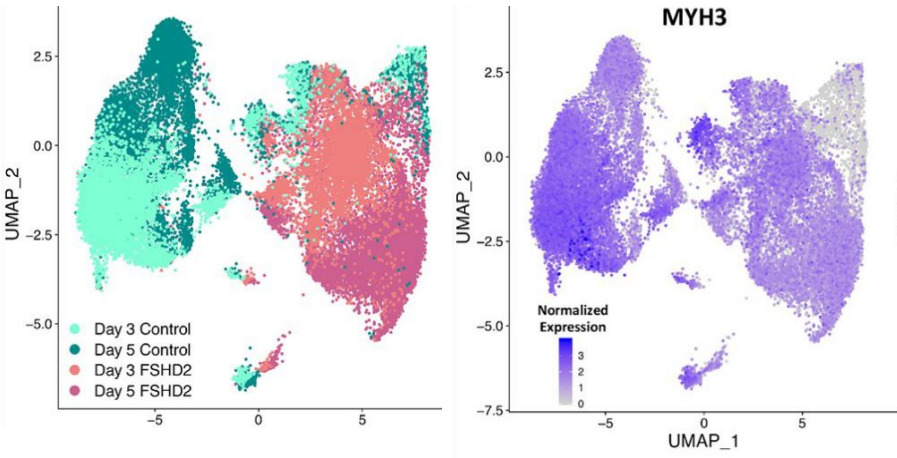


UMAP of chromatin accessibility signal in single-nucleus ATAC-seq of differentiating skeletal muscle

Rebboah and Reese et. al., 2021 (in review)

Lower dimensional representations of expression or signal data can reveal high dimensional similarity between datasets

- High dimensional data can include expression data or signal data
- PCA - principal component analysis
- tSNE - t-distributed stochastic neighbor embedding
- **UMAP - uniform manifold approximation and projection**

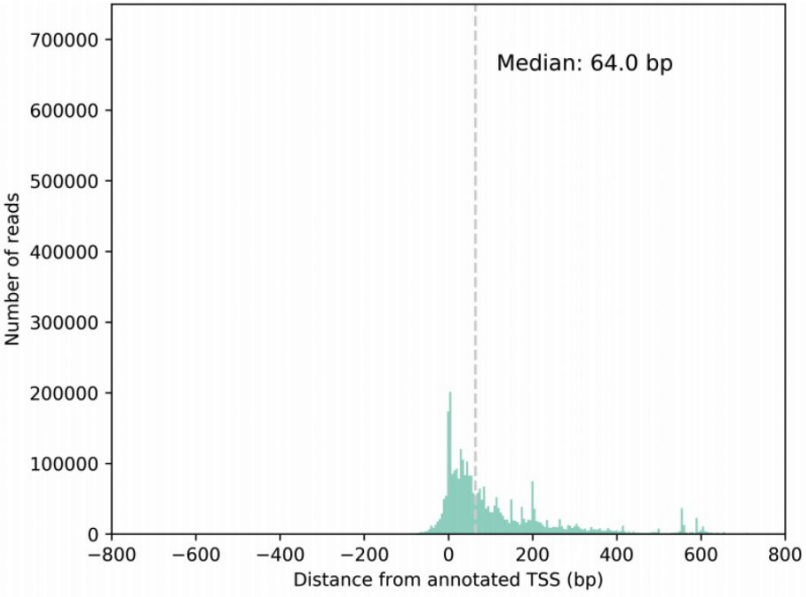


UMAP of gene expression in FSHD vs. control single-nucleus RNA-seq colored by sample (left), expression of *MYH3* (right).

Jiang and Williams et. al., 2020 (https://doi.org/10.1371/journal.pgen.1008754)

Histograms and density functions can be used to summarize distributions in data

- Plots the density or raw number of occurrences per binned or continuous measurements

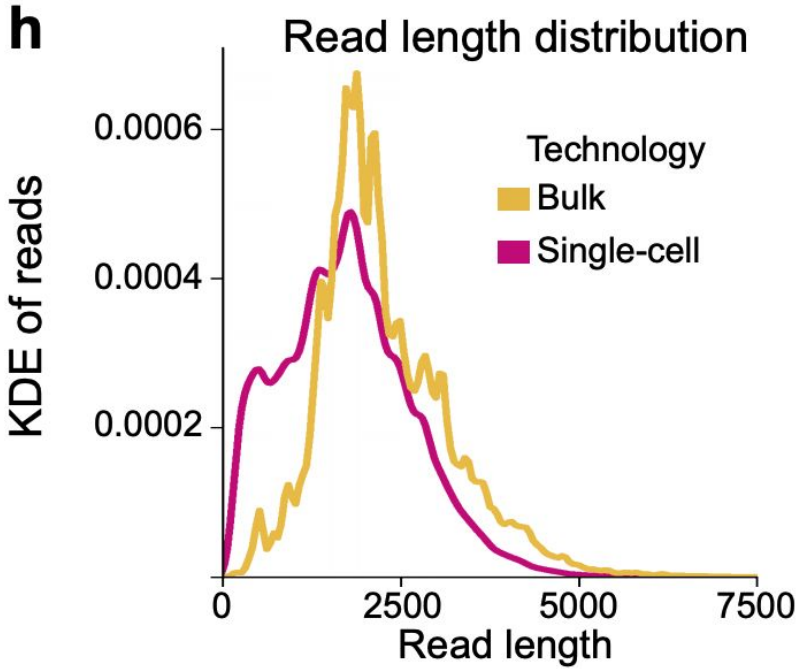


Histogram of distance from annotated TSS per read

Wyman, Balderrama-Gutierrez and Reese et. al. (in review)

Histograms and density functions can be used to summarize distributions in data

- Plots the density or raw number of occurrences per binned or continuous measurements

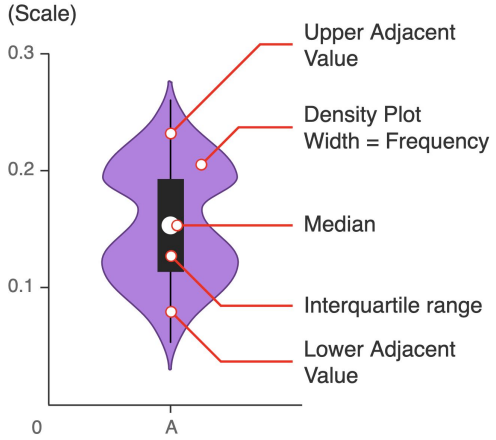


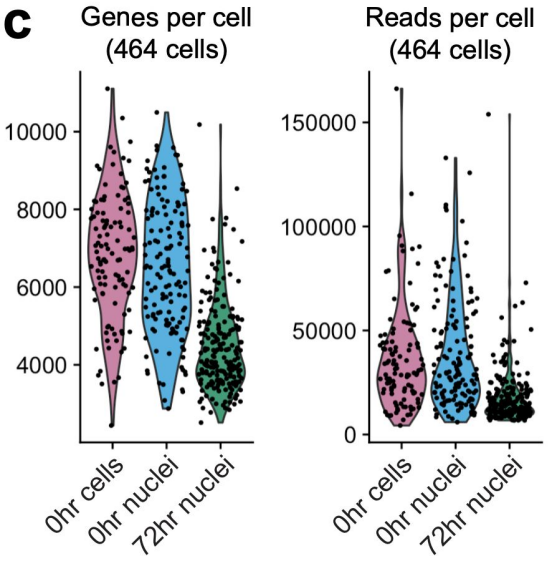
Kernel density estimate of distribution of read lengths between two experimental protocols

Rebboah and Reese et. al., 2021 (in review)

Violin plots are useful in single cell data analysis

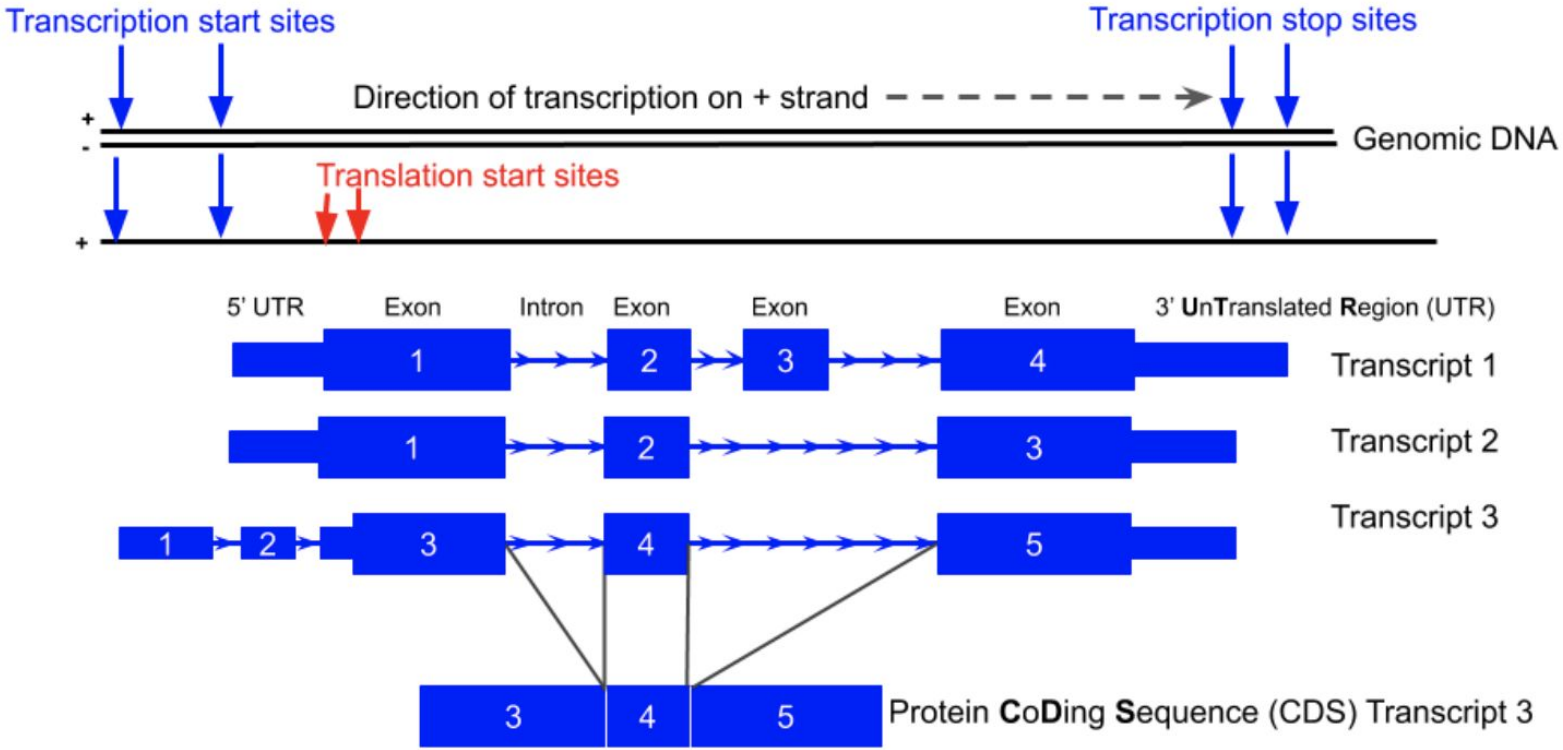
- Similar to box plot but with kernel density plot on each side
- Show full distribution of data
- Visualize expression levels of a gene across all cells, or quality control metrics (genes detected per cell, reads detected per cell)





Rebboah and Reese et al., 2021 (in review)

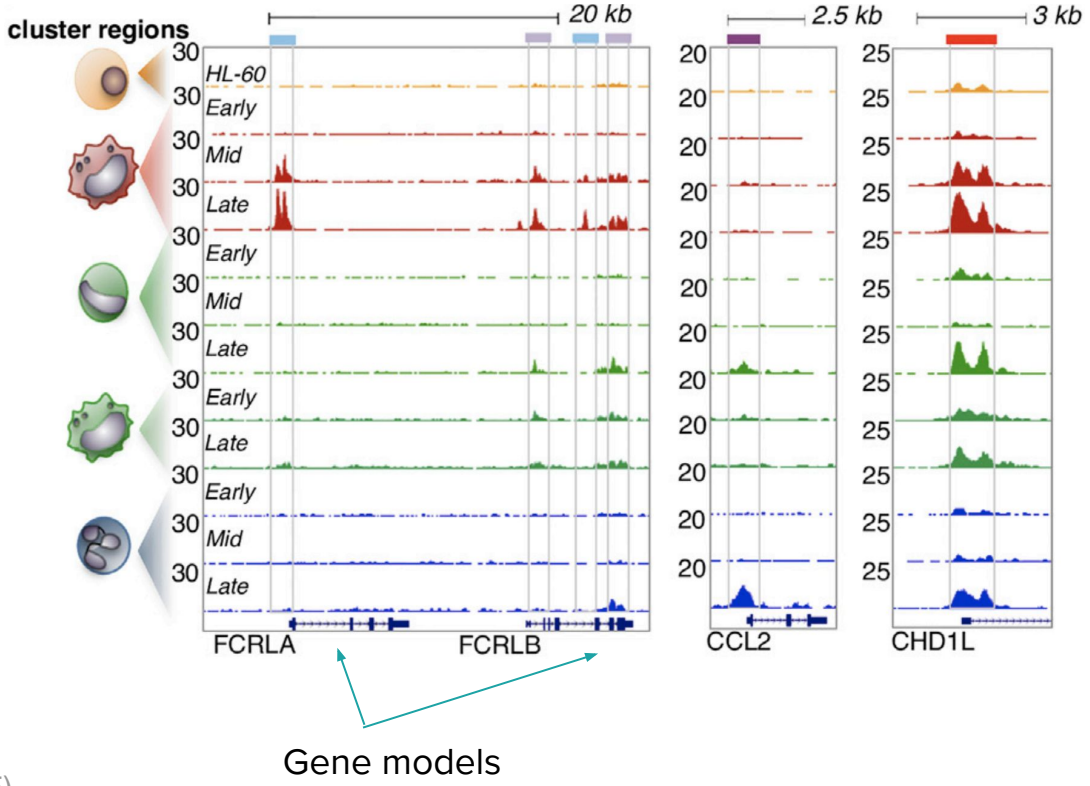
Traditional browser-style representation of transcript and gene models



Slide credit: https://www.melbournebioinformatics.org.au/tutorials/tutorials/Genome_browsers/GenomeBrowsers_Intro/

Signal such as chromatin accessibility can be visualized as “peaks” on the genome browser

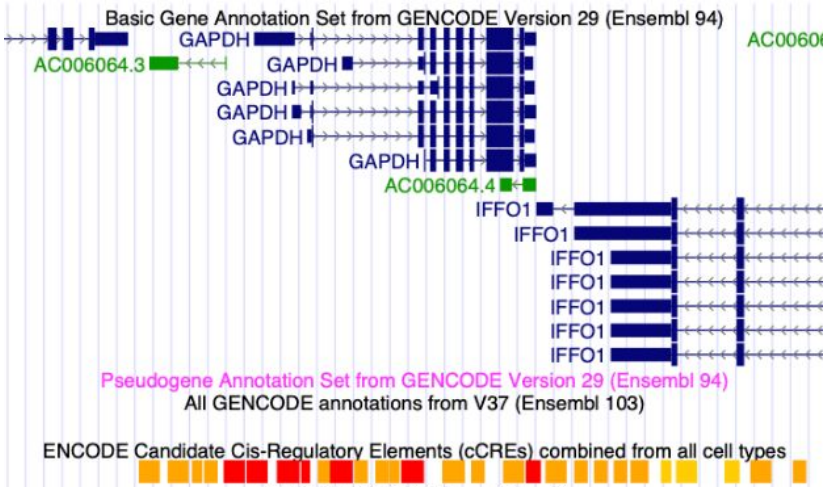
- Gene expression measured in discrete counts but chromatin data involves “peak calling” of signal across genomic location
- Zoom into interesting regions (like promoters) to visualize signal



Ramirez et al., 2017 (doi.org/10.1016/j.cels.2017.03.005)

Discrete regions without signal (BED regions) can also be visualized on the genome browser

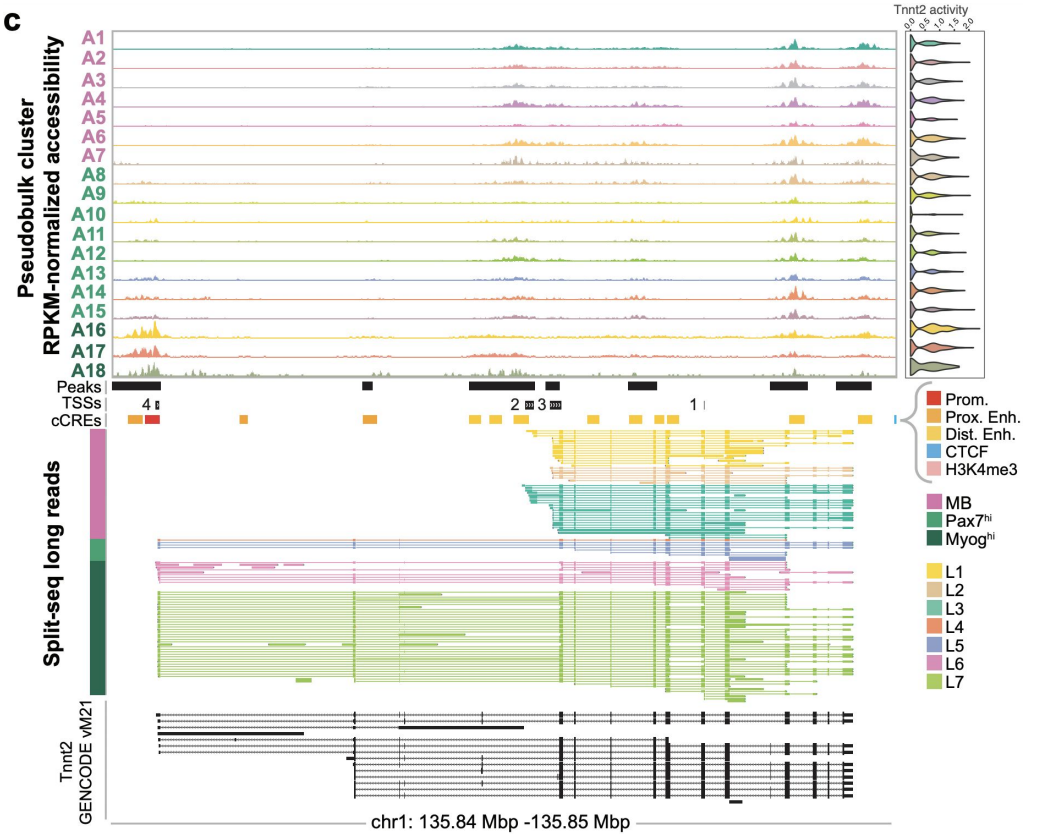
- Useful for regions with annotated function
- ENCODE cCREs (candidate cis-regulatory elements) - regions that are thought to have some regulatory impact on gene expresison



hg38 ENCODE cCREs
at the *GAPDH* locus

The genome browser facilitates visualization of multiple types of data at the same location

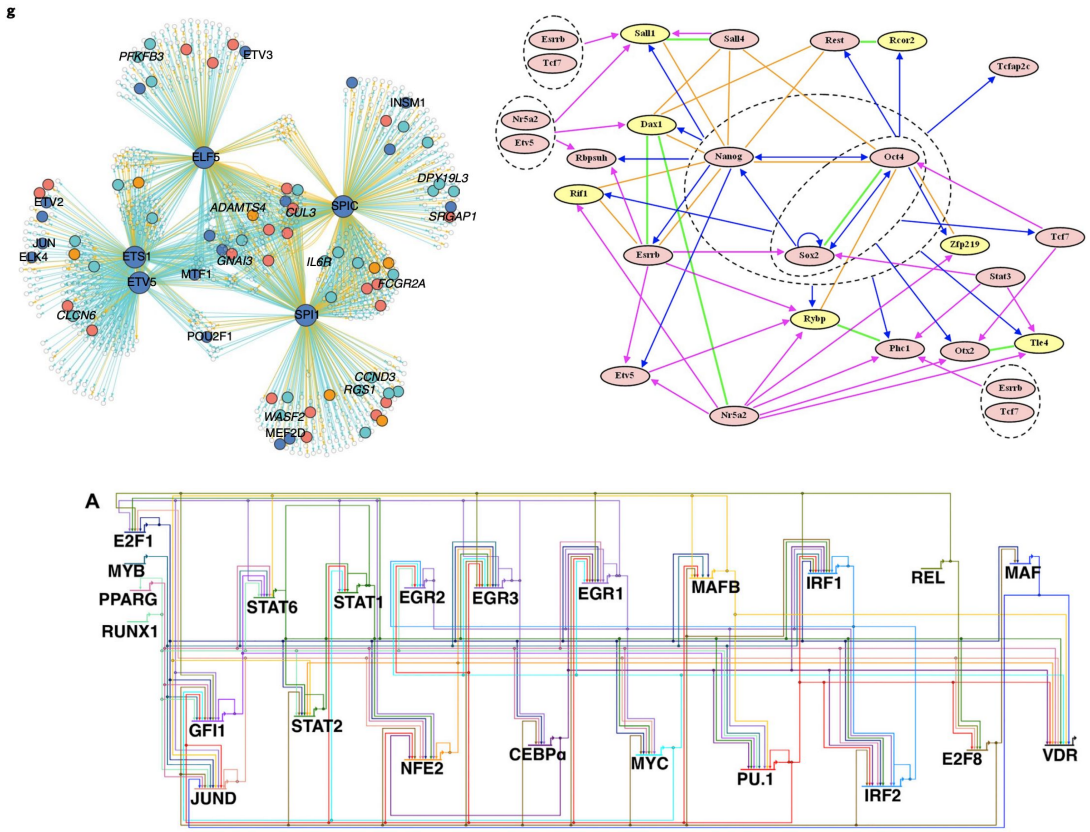
- Open chromatin (scATAC) signal
- gene “activity” violin plots
- transcript models
- TSS BED regions
- ENCODE cCRES
- Integrate ATAC and RNA data modalities in one figure



Rebboah and Reese et al., 2021 (in review)

Network diagrams show relationships between genes

- Visualize system-level gene regulation
- Genes are represented by nodes connected by edges
- Difficult to construct and often require experimental validation



Zhou et al., 2007 (doi.org/10.1073/pnas.0701014104); Ramirez et al., 2017 (doi.org/10.1016/j.cels.2017.03.005); Morabito and Miyoshi et. al., 2021 (https://doi.org/10.1038/s41588-021-00894-z)