1. Lecture Title and Date

- Genome Annotation (Multi-omic Analyses) (25m7-part1)
- February 12

2. Objectives of the Lecture

- Understand the concept of genome annotation including regulatory and non-coding elements
- Understand how to generate RNA-seq and ChIP-seq data for multi-omic annotation
- Explain how to use multi-scale analyses for epigenomic data interpretation
- Understand the integration strategies for combining different omics data sets
- Describe how genome organization such as Hi-C and TADs complements 1D annotation

3. Key Concepts and Definitions

- **Genome Annotation**: Identify and label functional elements in the genome, including protein-coding genes, regulatory elements, non-coding RNAs, and other functional regions
- **Non-coding Features**: Regions of the genome that are critical for gene regulation but do not code for proteins.
- **ChIP-seq:** A way to profile DNA-protein interactions. Peak calling is used to find significant binding or modification regions.
- **Peak Calling:** A statistical method to identify intervals of sequencing signals. It may compare ChIP signal to a control to identify true enrichment above background
- **IDR:** A metric to calibrate peak calls across biological or technical replicates through comparing the rank of peaks in 2 replicate experiments
- **Multi-scale Analysis**: Exam the genomic signals at different window sizes to capture both narrow and broad features in the epigenome.
- **Hi-C**: A way to measure physical contacts between genomic regions. Data are represented as a matrix and each cell represent interaction frequency between 2 genomic loci.
- **TAD**: Self-interacting genomic regions identified by Hi-C and other 3D conformation assays.

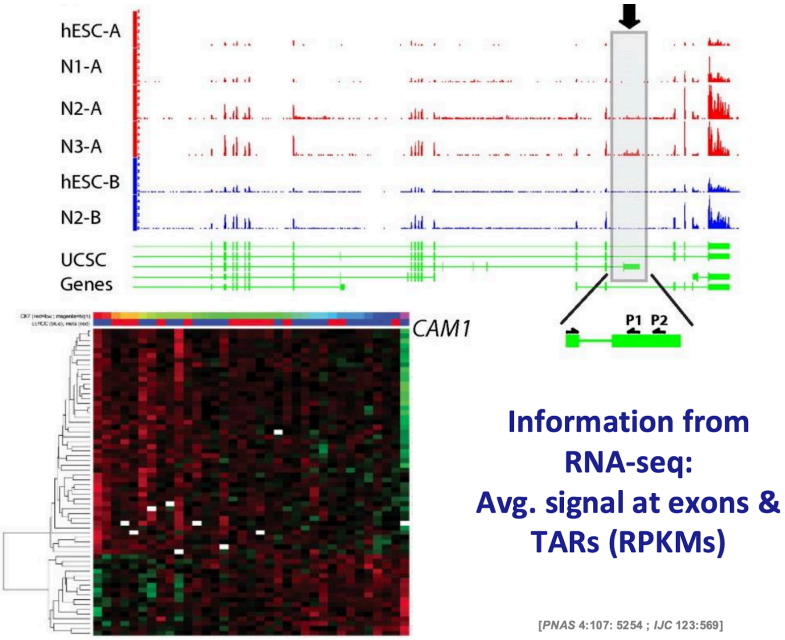
4. Main Content/Topics

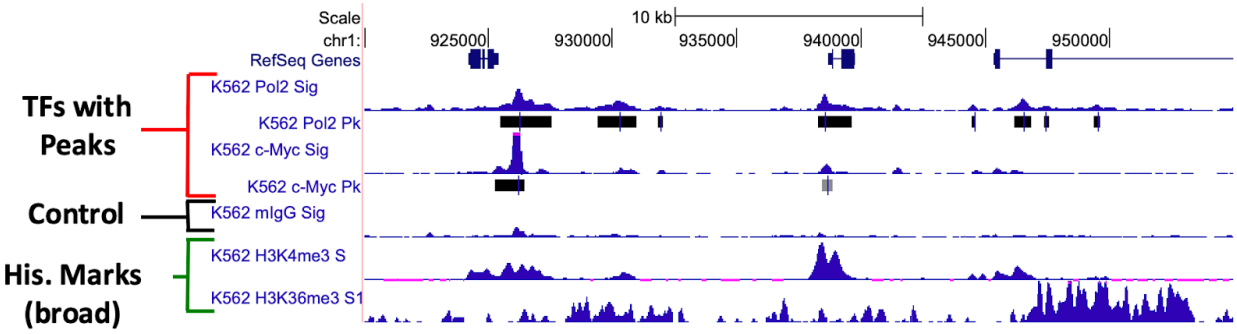
- **Defining Genome Annotation**

Annotation does more than identifying protein, it also involves labeling any functional element in the genome. This process could be considered as highlighting portions of a text. Since many functional elements do not encode proteins, a significant focus is placed on non-coding features.

- **RNA-seq and ChIP-seq for Functional Annotation**

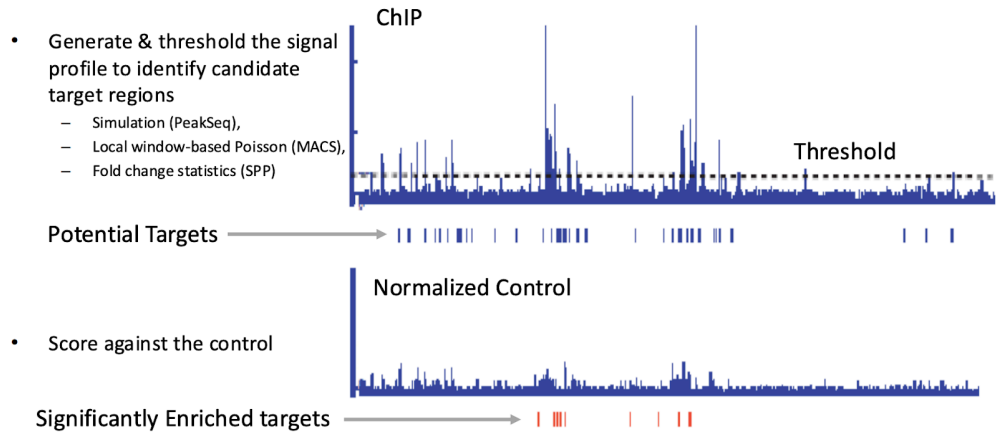
Researchers could take advantage of functional genomics assays like RNA-seq and ChIP-seq to capture regulatory information which can be mapped back onto the genome. **RNA-seq** measures transcriptional output and it reveals active exons, transcript boundaries, and novel transcriptionally active regions. The signal found is usually summarized in RPKM values which is used to depict how strongly certain parts of the genome are expressed. **ChIP-seq** identifies regions of DNA bound by specific proteins and Slide 6 shows example signal tracks which indicate occupancy.

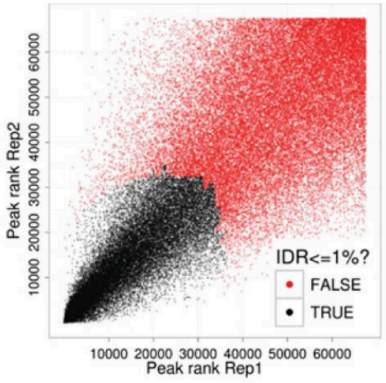


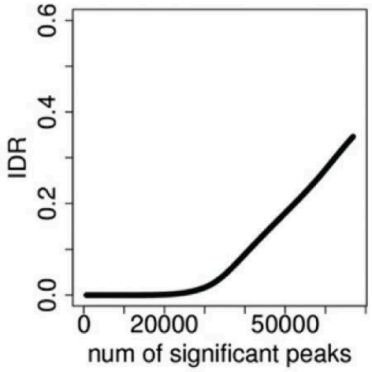


- **Peak Calling and Reproducibility**

To interpret noisy sequencing data, peak calling is performed. Tools such as MACS and PeakSeq are used to compare the signal of interest to a control and highlight enriched regions by applying statistical thresholds. The reproducibility across replicates is assessed using metrics like IDR since these data can be variable. High IDR values indicate low replicability and they are used to filter out spurious signals.

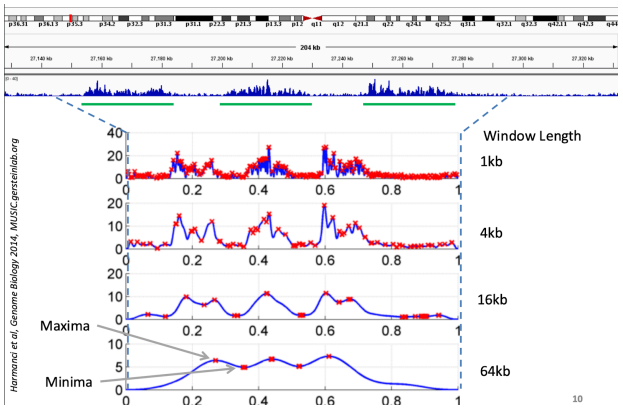


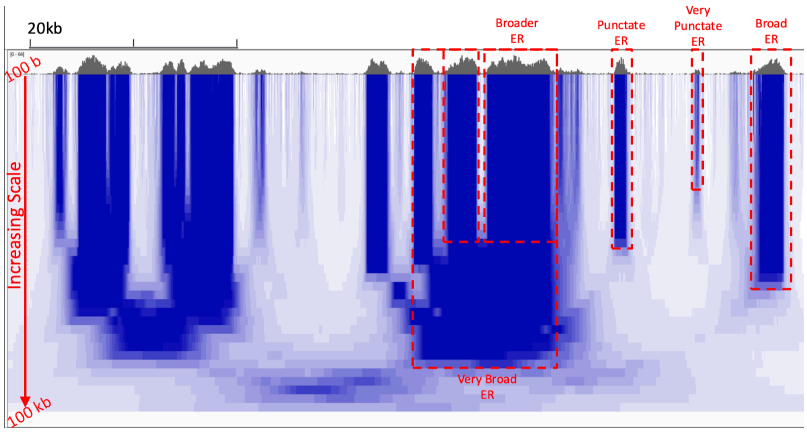




- **Multi-scale Analysis of Epigenomic Signals**

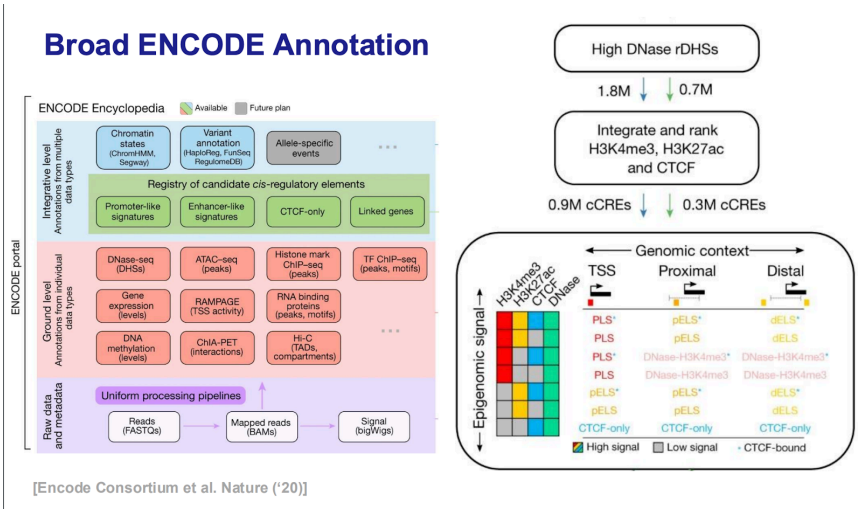
  Genomic signals from ChIP-seq may reflect both narrow binding peaks and broader regions of regulation. Multi-scale approaches analyze data in windows with different sizes. This could capture both fine details and larger domains.





- **Integration of Multiple Data Types (ENCODE)**

  Since numerous assays can be combined, major consortia like ENCODE compile and segment the genome into regulatory annotations. Computational tools like ChromHMM assign each segment to functional labels. This integrated view simplifies interpretation of large, complex datasets.

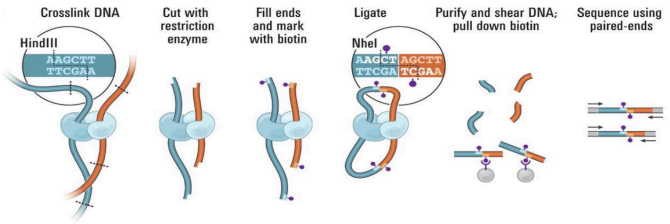


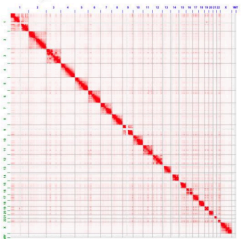
[Encode Consortium et al. Nature ('20)]

- 3D Genome Architecture and Hi-C

  Hi-C experiments produce contact maps indicating physical proximity between genomic segments. It Identifies TADs where regulatory elements and genes cluster in 3D space.

Hi-C contact map

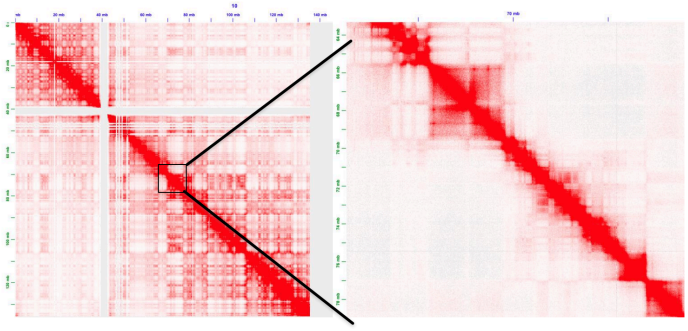




A network can be expressed as an adjacency matrix **A** where $\mathbf{A}_{ij} \neq 0$ if *i* and *j* are in contact & 0 otherwise

Science 2009, 5950: 289-293

Topologically associating domains (TADs)



TADs have apparent hierarchical organization

5. Discussion/Comments

- Even though the basic peak calling is conceptually straightforward, which is thresholding signal v.s. control, the real world applications could be complicated since it requires advanced statistical corrections and replicates
- Hi-C adds a new dimension to annotation, which allows researchers to connect distal enhancers to promoters through physical proximity

6. List all suggested reading:

Alexander, R. P. et al. (2010). Annotating non-coding regions of the genome. Nature Reviews Genetics, 11(8), 559–571. Focus on the concept of annotating regulatory features.

The ENCODE Project Consortium et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710. Emphasizes large-scale integration of epigenomic data.

Mackenzie, R. (2024). RNA-Seq: Basics, applications and Protocol. Genomics Research From Technology Networks.

Readings are helpful for understanding the breadth of genomic features. The targeted sections clearly explain large-scale integration. The RNA-seq primer is also a solid overview for transcriptomics

7. References ISL/ESL

- ISL: Ch. 2-4(regression and classification) can help to understand how supervised methods are used for genome annotation.
- ESL: Ch. 8–9(model-based clustering, HMMs, and advanced classification) can help to understand multi-state segmentation

8.. Suggest references for many of the key concepts

- Ernst, J., Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 12, 2478–2492 (2017). https://doi.org/10.1038/nprot.2017.124Any other reference material you would like to add

  It introduces to an approach to genome-wide segmentation