Lecture Title and Date

DATA - Genomics II, Jan 29, 2025

Objectives of the Lecture

1. Understanding how cells annotate the genome in chromatin
2. Understanding how RNA-seq works and how it differs from DNA-seq
3. Strategies to deal with rRNA and enriching for mRNA using polyA tails
4. Understand short and long read RNA-sequencing and ways to analyse RNA levels from RNA-seq

Key Concepts and Definitions

- **SNP (Single Nucleotide Polymorphism):** A variation in a single nucleotide at a specific position in the genome that can affect traits or disease susceptibility
- **CNV (Copy Number Variation):** A structural variation in the genome where sections of DNA are duplicated or deleted, leading to differences in the number of copies of a particular gene or genomic region among individuals
- **RNA Polymerase II:** An enzyme that transcribes DNA into mRNA
- **Enhancer:** A regulatory DNA sequence that increases the transcription of associated genes by providing binding sites for activator proteins, often functioning over *long genomic distances*
- **Promoter:** A DNA sequence located right upstream of a gene, where RNA Polymerase II and other transcription machinery assemble to initiate transcription
- **Transcription Factor:** A protein that binds to specific DNA sequences to regulate gene expression by either activating or repressing transcription
- **Heterochromatin:** A tightly packed form of chromatin that is transcriptionally inactive, often associated with gene silencing and repetitive DNA sequences
- **Modified histone:** the specific histone protein in the nucleosome that is being modified
- **Residue:** the amino acid in the histone protein where the modification occurs
- **Monomethylation:** addition of one methyl group
- **Dimethylation:** addition of two methyl groups
- **Acetylation:** addition of an acetyl group
- **RNA-Seq:** A high-throughput sequencing technique used to analyze gene expression, transcript isoforms, and RNA processing events across the transcriptome
- **mRNA:** Messenger RNA is a type of RNA that carries genetic information from DNA to ribosomes for protein synthesis, has a poly-adenylated (polyA tail)
- **rRNA:** Ribosomal RNA does not encode proteins but may dominate total RNA; it is typically removed from RNA-Seq libraries to enhance transcriptome coverage.
- **Junction Reads –** RNA-Seq reads that span exon-exon boundaries, enabling the detection of alternative splicing and novel transcript isoforms.
- **Normalization –** A method to adjust RNA-Seq read counts based on gene length and sequencing depth, allowing for accurate quantification of gene expression.

- **Unique molecular identifiers (UMIs):** short, random barcode sequences added to RNA or DNA molecules before amplification in sequencing. UMI helps distinguish original molecules from PCR duplicates, allowing accurate quantification of gene expression by eliminating amplification bias in sequencing.
- **Epitope**: the exact part of something that an antibody binds to
- **Crosslink:** get transient interactions to become more permanent through stronger bonds

Main Content/Topics

Annotation of the genome in chromatin

All cells have the same DNA, but their gene expression differs depending on the time and spatial location. Understanding sequencing data requires analyzing not only genomic expression but also the spatiotemporally resolved chemical, biochemical, and epigenetic annotations. DNA is wrapped around nucleosomes composed of histone proteins. This packaging, along with regulatory factors, influences the accessibility of different genomic regions, determining which DNA sequences are available for transcription and interactions.

An example of factors influencing DNA accessibility is histone modifications. These histone modifications participate in distinct genomic functions, such as recruitment of RNA Polymerase II to active transcription sites, marking enhancers, facilitating transcription factor binding, or signalling inactive chromatin states like heterochromatin, which suppress gene expression. Here are some related nomenclature:

- Modified histone (e.g., H3, H4): the specific histone protein in the nucleosome that is being modified
- Residue (e.g., K for lysine): the amino acid in the histone protein where the modification occurs
- Type of modification (e.g., me1 for monomethylation, ac for acetylation): the chemical modification made to the residue

Sequencing techniques are used to annotate the genome by identifying regulatory elements (e.g., ATAC-Seq), analyzing chromatin composition (e.g., ChIP-Seq, CUT&Tag), tracking RNA Polymerase activity (e.g., ChIP-Seq), and studying 3D genome folding (e.g., Hi-C). These methods enable targeted (ex. only care about 1 TF) or global exploration (ex. want to know all regulatory elements near 1 gene) of gene regulation, chromatin state, and structural interactions.

Those sequencing techniques largely depend on the deformations that regulatory proteins cause in the chromatin 3D structure, as those deformations control access for DNA-cutting enzymes to their substrate. In conjunction with various wet lab methods to separate resulting DNA fragments, the fragments can be sequenced and mapped to the known genome to determine exactly what those regulatory elements are.

ChIP is especially important for enriching sequencing samples for regulatory elements.
Chromatin **I**mmuno**P**recipitation involves several steps:

- Crosslink the regulatory proteins to the DNA. Since those proteins naturally bind only transiently, a fixative such as formaldehyde is applied to freeze them in place.
- Shear chromatin. Big pieces make it easier to match to the genome. Small pieces facilitate identification of the exact regulatory sequence in the DNA.
- Enrich with an antibody for the regulatory protein. Ideally this only enriches for the target protein, but antibodies aren't guaranteed to be specific.

After the sample is prepared, both the baseline and the enriched samples are sequenced. Comparing them to each other allows peaks to be called (i.e., sites with sufficiently increased reads are identified as target sites).
ChIP-Seq limitations:

- Cross linking isn't perfect.
- Antibodies aren't perfect.
- Even good antibodies can't always reach the epitope.
- Not all DNA-bound proteins actually promote transcription.

ChIP-Seq extensions:

- ChIP-exo: as mentioned earlier, smaller pieces mean higher-resolution regulatory sequences. Exonucleases make those pieces extra small.
- Spike-in normalization: add reference standard nucleosomes to get better quantitation.
- CUT&RUN/CUT&Tag: use antibodies to target nucleases directly to regulatory sites associated with a specific regulatory protein or histone PTM. Use magnetic beads instead of IP to remove everything except the DNA of interest before amplification.
- Apply technology to RNA factors instead of protein factors.

Finally, we can annotate the genome with respect to spatial correlations. When regulatory factors interact with two different loci at the same time, they demonstrate spatial correlation between those loci. Those simultaneous interactions can be captured by:

- First, chemical fixation and DNA fragmentation
- Second, ligating the ends of the interacting loci to each other in a loop

There are several different methods for accomplishing this and analyzing the output. They are differentiated by the interactions being considered, whether one experiment targets single interactions, the interactions between one locus and the rest of the genome, the interactions between many different sites with each other, or all locus-locus interactions. See Chromatin Interaction Data reference for more information.

Some rules of thumb for sequencing :)

- Sequence comprehensively if you have some elbow grease handy
- Target features if you know something about them
- Enrich for something if you have a lot of sample
- Cut DNA with targeting enzymes if they have the right target
- Use read-end location data for single-nucleotide info
- Use clever chemistry for latent info
- Tag with UMIs to get more from reads

- Mix and match methods for the best of both worlds (see Hi-ChIP reference below)
- There's no silver-bullet platform for everything. Know the pros and cons.

RNA Sequencing methods and applications

RNA sequencing (RNA-Seq) is a powerful technology that lets you do transcriptome-wide analysis of gene expression. It can provide insights into cellular states, regulatory mechanisms, and alternative splicing events, etc. Unlike DNA sequencing, which focuses on the static genome, RNA-Seq is also pretty cool because it captures the dynamic range of RNA molecules, including messenger RNA (mRNA), non-coding RNA, and other processed transcripts.
The workflow involves:

1. Isolating RNA and removing unwanted contaminants,
2. Fragmenting and converting it into complementary DNA (cDNA),
3. Adding sequencing adapters,
4. Sequencing the cDNA to reconstruct the transcriptome.

This allows for the quantification of gene expression, the identification of novel transcripts, and the analysis of transcript isoforms.

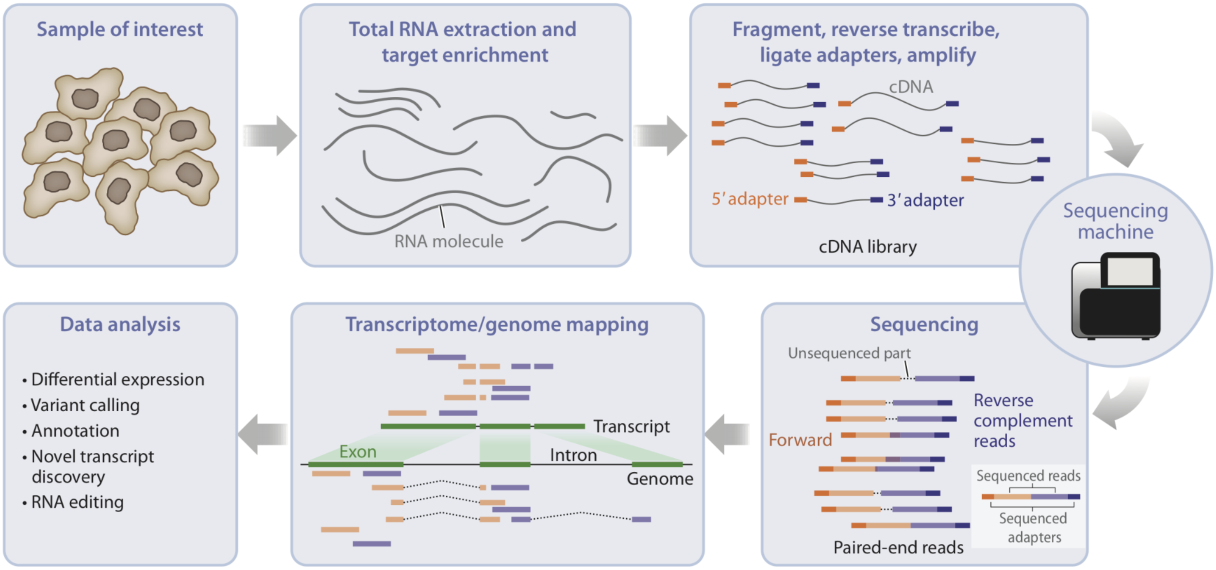

Figure: RNA sequencing workflow (source: Koen Van den Berge & Lieven Clement, StatOmics)

RNA-Seq vs DNA-seq: RNA-Seq differs from standard DNA sequencing in several key ways. RNA exists in a wide dynamic range of concentrations and requires careful handling due to its susceptibility to degradation by RNases and spontaneous chemical hydrolysis. Unlike double-stranded DNA, RNA is single-stranded and often requires a strand-specific sequencing protocol. It also undergoes extensive processing, such as splicing, capping, and polyadenylation, which affects its stability and function. It can also contain chemical modifications that either interfere with reverse transcription or remain undetectable in sequencing. In Matthew's words, for the skeptics, working with RNA is completely doable, you just need to take care! :)

Dealing with rRNA**:** Ribosomal RNA (rRNA) constitutes the majority of total RNA in cells and can dominate sequencing reads if not selectively removed. Since rRNA does not provide useful gene expression information, its presence increases sequencing costs while reducing coverage of protein-coding and regulatory transcripts. To achieve meaningful transcriptomic data, rRNA must be depleted or circumvented during RNA-Seq library preparation. For sequencing human mRNA without rRNA, poly(A) tail selection is an effective strategy. Most eukaryotic mRNAs have polyadenylated (poly(A)) tails, which can be selectively captured using oligo(dT)-attached magnetic beads. This approach enriches mature mRNAs while excluding rRNA and other non-polyadenylated RNAs.

RNA-Seq can mostly capture reads from exons, and includes junction reads that span exon-exon boundaries. These enable the identification of splicing events and alternative transcript isoforms. Mapping these reads to a reference genome allows us to quantify gene expression, infer splicing patterns, and detect novel transcripts. RNA levels can be analyzed using three main approaches:

- alignment to existing gene annotations,
- reference-guided transcript assembly, or
- de novo assembly when a reference genome is unavailable.

Proper normalization is necessary for accurate quantification. Internal methods like RPKM/FPKM (Reads/Fragments Per Kilobase per Million mapped reads) adjust for transcript length and sequencing depth, while external normalization methods use spike-in controls to account for variability.

Emerging technologies such as direct long read RNA sequencing using Oxford Nanopore Technology (ONT) enable sequencing of full-length RNA molecules without requiring cDNA synthesis or PCR amplification, reducing biases from traditional library preparation. ONT also allows direct detection of RNA modifications and poly(A) tail length. This approach enhances RNA-Seq's ability to capture splicing dynamics and transcript heterogeneity while minimizing artifacts introduced in conventional workflows.

Principles of sequencing methodology

1. Global vs. Targeted Approaches: Global methods offer comprehensive coverage but require more sequencing, while targeted methods provide better feature resolution but need prior knowledge.
2. Biochemical and Enzymatic Enrichment: Biochemical enrichment expands possibilities but demands more starting material, whereas enzymatic methods enhance sensitivity but are limited by enzyme efficiency and specificity.
3. Single-Nucleotide and Latent Information: Mutations, read-end positions, and chemical modifications can reveal detailed nucleotide-level insights.

4. Unique Molecular Identifiers (UMIs): UMIs can provide more information about sequencing reads.
5. Modular and Combinable Methods: Many sequencing techniques can be integrated for enhanced analysis.
6. Platform-Specific Considerations: Each sequencing platform has distinct advantages and limitations.

Targeting Sub-populations of RNAs

1. Depleting Unwanted RNAs
    - rRNA removal (e.g., Ribo-minus)
    - Enzymatic degradation (e.g., RNase H, Cas9)
    - Global degradation (e.g., uncapped RNAs degraded by 5'→3' exonuclease)
2. Enriching/Amplifying Specific Transcripts
    - Targeted RT primers for specific sequences
    - Hybridization-based capture for biochemical enrichment
    - 5'/3' modification-based selection (e.g., miRNA with 5'-phosphate and 3'-hydroxyl)
3. Selecting Newly Made RNAs
    - Fractionating chromatin-associated RNAs
    - Filtering intron-containing RNAs
    - Immunoprecipitating RNA Pol II-engaged RNAs
    - Metabolic labeling with short pulses
4. Selecting Modified RNAs
    - Immunoprecipitation using modification-specific antibodies
    - Chemical methods inducing RT stops or mutations
5. Selecting RNAs from Specific Cells
    - Cell microdissection or FACS sorting
    - TU-tagging (targeted metabolic RNA labeling in specific cells)
    - Single-cell RNA-seq (scRNA-seq)

Applications of RNA sequencing

1. Examining RNA structure with chemical probing
   Chemical probing methods, such as DMS and SHAPE, assess RNA structure by modifying accessible or flexible nucleotides, causing reverse transcription stops or mutations detectable via sequencing.
2. Examining cell heterogeneity with scRNA-seq
   scRNA-seq reveals cellular heterogeneity, overcoming the limitations of bulk RNA sequencing by uncovering rare cell populations and lineage trajectories.
3. Spatial sequencing methods
   Detection -> identification -> measurement -> spatial localization of molecular signals to map gene expression in tissue contexts.
4. CRISPR screens
5. Massively parallel reporter assays (MPRA)

Suggest references for many of the key concepts

- Infographic for Histones
- CUT&RUN vs CUT&Tag
- Chromatin Interaction Data
- Combo method: Hi-ChIP
- How does RNA sequencing work?
- If you want a deeper dive into RNA sequencing, this paper is a good place to start: Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. Cold Spring Harb Protoc. 2015 Apr 13;2015(11):951-69. doi: 10.1101/pdb.top084970. PMID: 25870306; PMCID: PMC4863231.
- RNA structures: Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. Nat Rev Genet 19, 615-634 (2018).
- scRAN-seq: Aldridge S, Teichmann SA. Single cell transcriptomics comes of age. Nat Commun. 2020 Aug 27;11(1):4307. doi: 10.1038/s41467-020-18158-5. PMID: 32855414; PMCID: PMC7453005.
- Spatial transcriptome: Dario Bressan et al. ,The dawn of spatial omics.Science381,eabq4964(2023).DOI:10.1126/science.abq4964
- Genome-wide quantitative sequencing: Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Science. 2013 Mar 1;339(6123):1074-7. PMID: 23328393.