

Genomics Part II

Applications of Sequencing Technology

Biomedical Data Science: Mining and Modeling
CB&B 752 · MB&B 452
Matt Simon
2/2/22

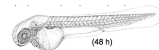
Overview

- Genomics I (Monday's lecture): Focus on sequencing technology and genomes.
- Genomics II: (Today's lecture): Focus on applications of sequencing technology.
 1. Annotation of the genome in chromatin
 2. RNA-seq methods and applications
 3. Topics suggested on Monday.

Workflow

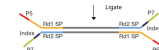
1. Isolation of sample.

e.g., Isolate DNA and shear.



2. Library preparation

e.g., Clean up and ligate Y-adaptors.



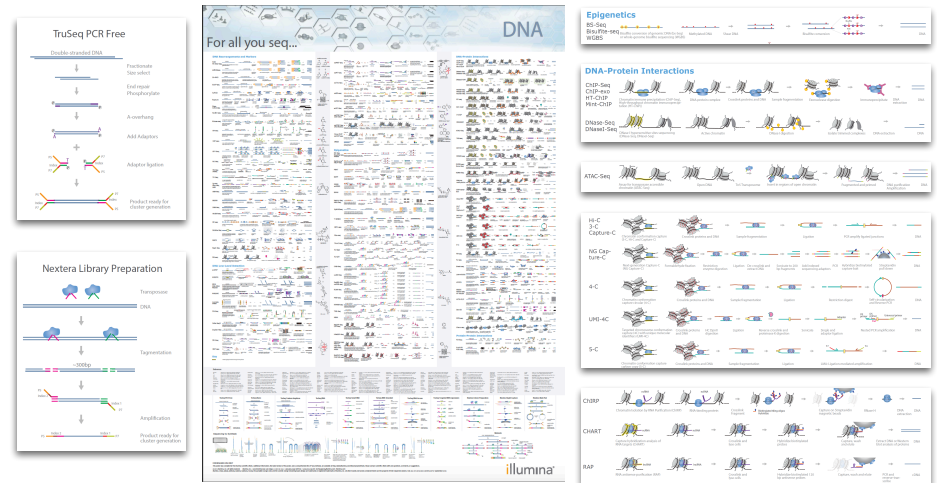
3. Sequencing

e.g., Illumina HiSeq



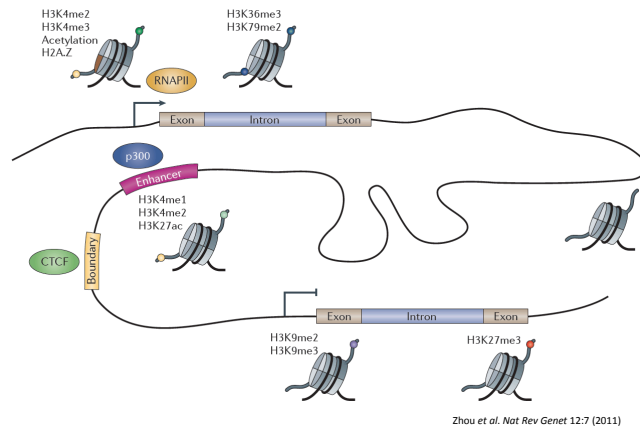
4. Analysis

e.g., Map to genome and interpret.

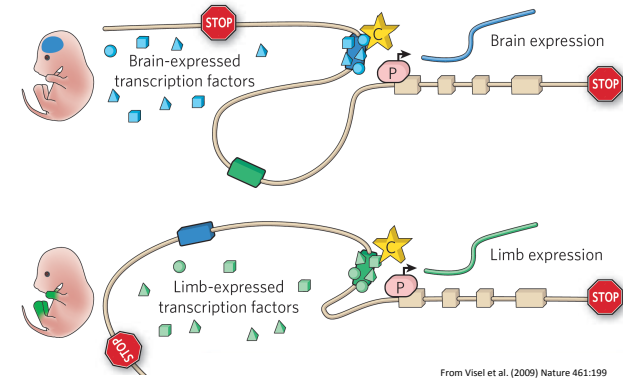


<https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf>

Chromatin modifications correlate with different genomic functions.



Regulation is temporally and specially controlled

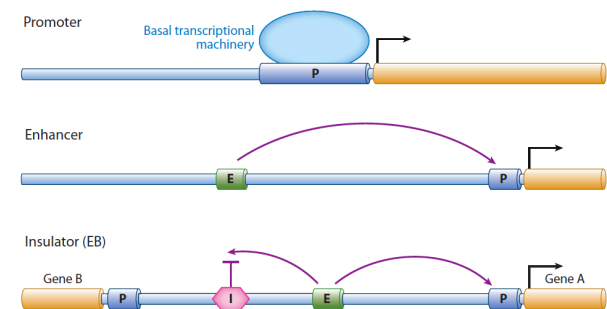


Using sequencing to annotate the genome

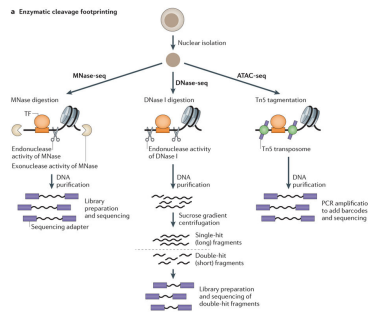
- Where are the cis-acting regulatory elements in DNA?
 - DNase I hyper-sensitivity mapping (**DNase-Seq**).
 - FAIRE** to map regulatory elements.
 - ATAC-Seq** to map regulatory elements.
- How does the chromatin composition vary across the genome?
 - ChIP-seq** of transcription factors (or in high res, ChIP-exo)
 - CUT&RUN** and **CUT&Tag** for small scale/single cell analysis.
- Where is RNA polymerase transcribing?
 - ChIP-Seq** of polymerase.
 - GRO-Seq**, **PRO-Seq**, **TT-seq** and **NET-Seq** to measure RNA polymerase activity.
- What sites are methylated in the genome?
 - Bisulfite-Seq** to measure mC levels.
 - Methyl-Seq** to measure mC levels.
- How is the genome folded in the nucleus?
 - Hi-C** to measure ligation/contact frequencies.
 - 3C/4C/5C** to measure looping at specific loci.

Targeted approaches v Global approaches

How do we identify regulatory elements in the genome?



Using differences in biochemical properties of regulatory elements to identify them by Seq

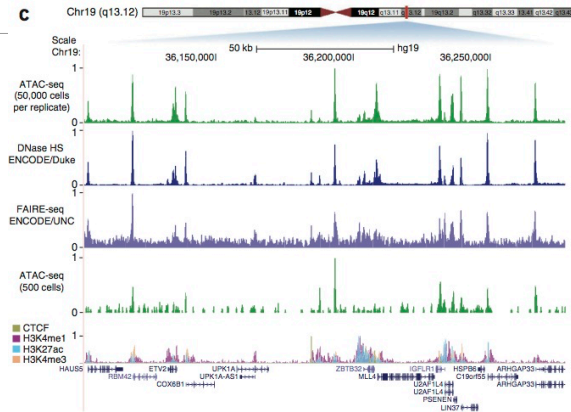


Zentner GE, Henikoff S. High-resolution digital profiling of the epigenome. *Nat Rev Genet.* 2014;15: 814-827. doi:10.1038/nrg3798

Transcription factor binding frequently deforms the B-form DNA, making it hypersensitive to DNase I and transposases.

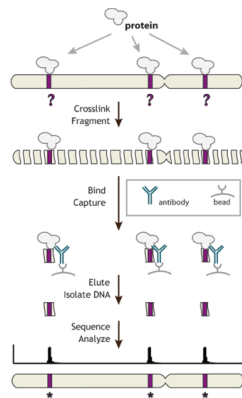
Changes in **accessibility of chromatin** can provide information about regulation

- ATAC-seq (shown)
- MNase-Seq (shown)
- DNase-Seq (shown)
- FAIRE-Seq (not shown)



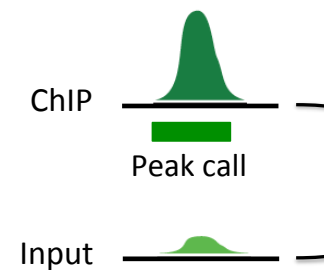
Buenrostro JD, Giresi PG, Zaba LC, Chang HY, and Greenleaf WJ. (2013) "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature Methods*

Localization of *specific proteins* in the genome with chromatin immunoprecipitation (ChIP-Seq)



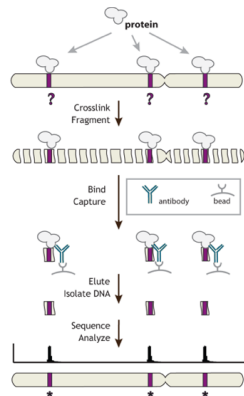
1. **Crosslink** the cells with formaldehyde to "fix" factors in place.
 Exception: Native ChIP with histone antibodies.
2. **Shear chromatin** to smaller pieces.
 Shear size determines resolution. Note: ChIP-exo uses an exonuclease at a later step to increase resolution.
3. **Enrich** target using an antibody.
 Enrichment is only as good as the antibody.

Determining sites of enrichment from ChIP-Seq



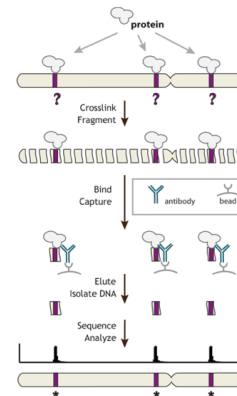
1. **Align** reads to the genome.
2. **Compare to input** to look for enrichment.
 Input coverage is not even.
3. **Call peaks** to determine statistically significant sites of enrichment.

Limitations of ChIP-Seq



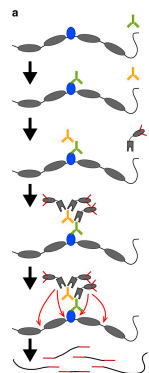
1. **Cross linking** efficiency is not necessarily uniform.
2. Enrichment is dependent on the **quality of antibody**.
e.g., Site and degree of histone modifications.
3. Enrichment is dependent on the **accessibility of the epitope**.
Comparing different sites to each other in the genome can be problematic.
4. Output is **descriptive**.
Hard to infer function without more experimentation.

Extensions of ChIP

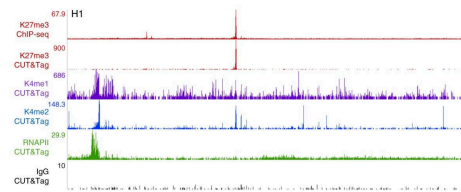


1. Using a nuclease to achieve **higher resolution** (ChIP-exo).
2. Analysis of **small samples or single cells** (CUT&RUN or CUT&Tag).
3. Extension to **RNA factors**.

CUT&Tag



Concept: Use factor-specific antibodies to target a transposase to direct the addition of DNA tags.

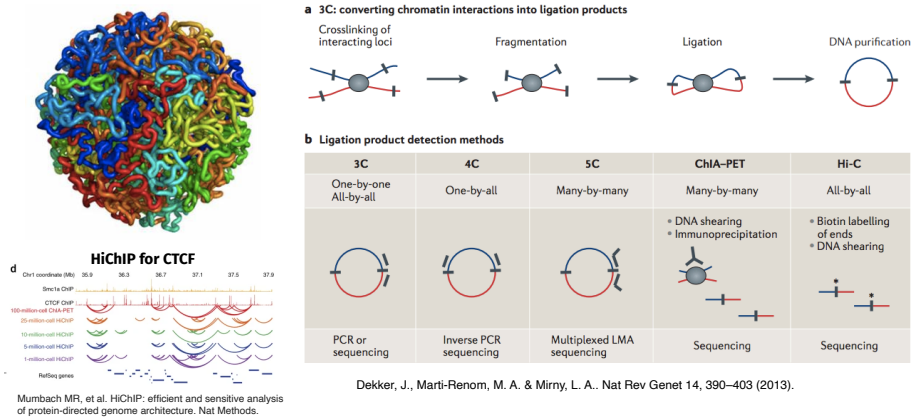


Kaya-Okur... & Henikoff (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun*

All of the following are advantages of CUT & Tag EXCEPT:

- (A) Using a transposase simplifies library preparation.
- (B) CUT & Tag allows analysis of endogenous proteins without high affinity antibodies.
- (C) Can be performed with very few cells.
- (D) Avoids artifacts from chromatin sheering.

Mapping genome folding (and rearrangements)

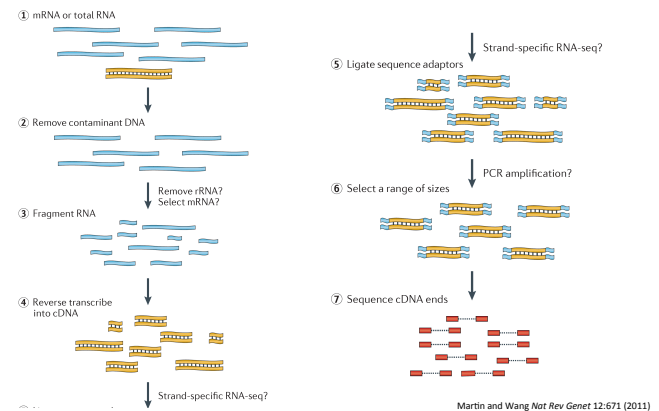


Nine principles of sequencing methodology

1. **Global approaches** can be (mostly) comprehensive, but require more sequencing.
2. **Targeted approaches** can provide better coverage of features of interest but require prior information.
3. **Biochemical enrichment** provides many opportunities, but generally requires more starting material.
4. **Enzymes** can often provide more sensitive approaches to target specific types of nucleic acids, but are limited by the efficiency and specificity of the enzymes that are available.
5. Single nucleotide information can be revealed through **mutations** or the **location of the end** of a read.
6. **Chemistry** can often be used to reveal latent information in a sequencing experiment.
7. Unique molecular identifiers (**UMIs**) can provide **additional information** about each read.
8. Many sequencing methods are **modular** and can be combined with one another.
9. Each **sequencing platform** has unique advantages and challenges.

Part 2: RNA-Seq and applications of RNA-Seq

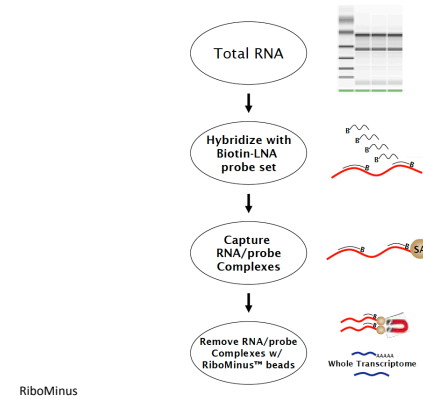
Example of an RNA-Seq workflow



How is RNA-Seq different from standard DNA-seq?

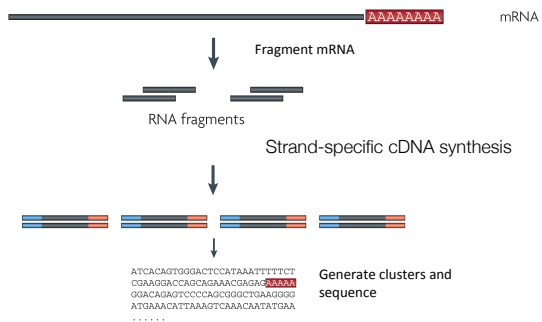
- Wide dynamic range of RNA concentrations.
- RNA is strand specific (unlike dsDNA)
- RNA degrades easily (RNase and spontaneous chemical hydrolysis)
- RNA is processed (e.g., capped, spliced, polyA)
- RNA can have modifications that can block RT or be invisible (e.g., tRNAs).
- There are a wide range of sizes of RNAs and specialized protocols are necessary for studying shorter RNAs (e.g., miRNA, short capped RNAs)
- RNA has secondary structure (possible blocks to reverse transcriptase).

Ribosomal RNA will dominate the sequenced reads unless removed

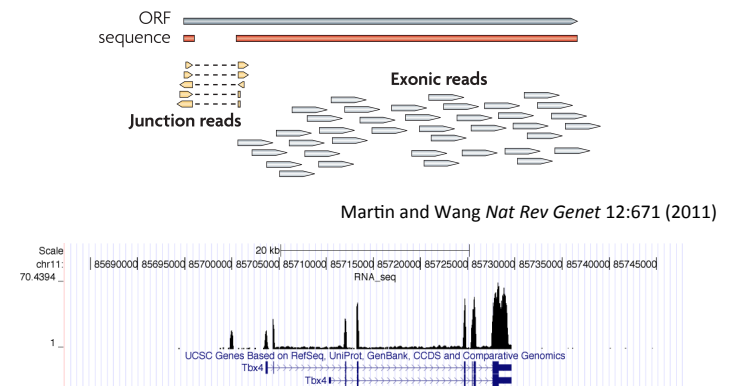


Illumina RNA-seq workflow

- Capture poly-A RNA with poly-T oligo attached beads (100 ng total) (2x)
- RNA quality must be high – degradation produces 3' bias
 - Non-poly-A RNAs are not recovered



RNA-Seq reads map mostly to exons



How does one analyze RNA levels from RNA-Seq?

Use existing gene annotation:

Align to genome plus annotated splices
 Depends on high-quality gene annotation
 Which annotation to use: RefSeq, GENCODE, UCSC?
 Isoform quantification?
 Identifying novel transcripts?

Reference-guided alignments:

Align to genome sequence
 Infer splice events from reads
 Allows transcriptome analyses of genomes with poor gene annotation

De novo transcript assembly:

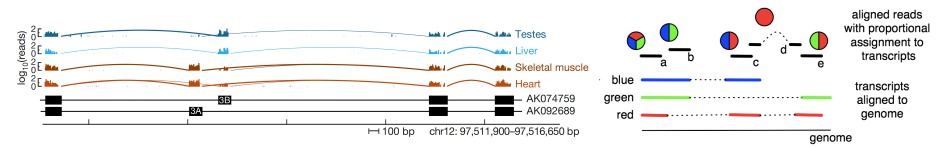
Assemble transcripts directly from reads
 Allows transcriptome analyses of species without reference genomes

RNA-seq reads contain information about the abundance of different transcript isoforms

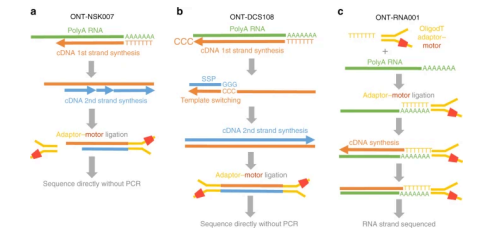
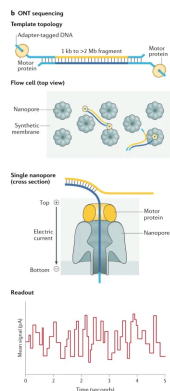
Normalization :

Internal: Reads or Fragments per kilobase of feature length per million mapped reads (RPKM or FPKM)

External: Reads relative to a standard "spike"



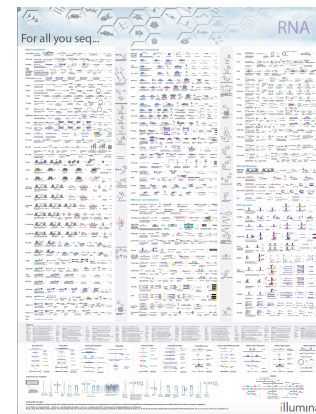
Direct RNA sequencing using ONT



Soneson C, et al. Nat Commun. 2019 Jul 31;10(1):2359. doi: 10.1038/s41467-019-11272-z. PMID: 31366910; PMCID: PMC6668388.

Long reads identify each RNA transcript isoform.
 Potential to identify RNA modifications directly.
 Can measure the length of RNA polyA tail.
 In principle avoids challenges/biases of library preparation.

Examples of applications of RNA-seq



- Characterizing transcriptome complexity
 - Alternative splicing
 - RNA modifications
 - RNA structures
- Differential expression analysis
 - Gene- and isoform-level expression comparisons
- Novel RNA species
 - lncRNAs and eRNAs
 - Pervasive transcription
- Translation
 - Ribosome profiling
- Allele-specific expression
- Measuring RNA half-lives and decay
- Examining protein-RNA interactions
- Effect of genetic variation on gene expression
 - Imprinting
 - RNA editing
 - Novel events

Nine principles of sequencing methodology

1. **Global approaches** can be (mostly) comprehensive, but require more sequencing.
2. **Targeted approaches** can provide better coverage of features of interest but require prior information.
3. **Biochemical enrichment** provides many opportunities, but generally requires more starting material.
4. **Enzymes** can often provide more sensitive approaches to target specific types of nucleic acids, but are limited by the efficiency and specificity of the enzymes that are available.
5. Single nucleotide information can be revealed through **mutations** or the **location of the end** of a read.
6. **Chemistry** can often be used to reveal latent information in a sequencing experiment.
7. Unique molecular identifiers (**UMIs**) can provide **additional information** about each read.
8. Many sequencing methods are **modular** and can be combined with one another.
9. Each **sequencing platform** has unique advantages and challenges.

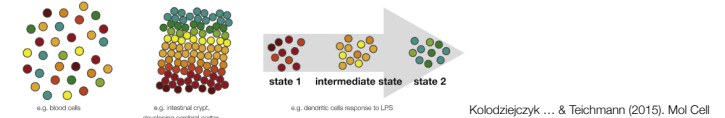
Nine principles of sequencing methodology

1. **Global approaches** can be (mostly) comprehensive, but require more sequencing.
2. **Targeted approaches** can provide better coverage of features of interest but require prior information.
3. **Biochemical enrichment** provides many opportunities, but generally requires more starting material.
4. **Enzymes** can often provide more sensitive approaches to target specific types of nucleic acids, but are limited by the efficiency and specificity of the enzymes that are available.
5. Single nucleotide information can be revealed through **mutations** or the **location of the end** of a read.
6. **Chemistry** can often be used to reveal latent information in a sequencing experiment.
7. Unique molecular identifiers (**UMIs**) can provide **additional information** about each read.
8. Many sequencing methods are **modular** and can be combined with one another.
9. Each **sequencing platform** has unique advantages and challenges.

Examples of how to target sub-populations of RNAs

1. Deplete unwanted RNAs
 - A. Ribo-minus etc. to remove rRNA.
 - B. Enzymatic removal using targeted enzymes (e.g., RNase H, Cas9)
 - C. Globally degrade unwanted RNAs (e.g., uncapped RNAs with a 5'-to-3' exonuclease).
2. Enrich/amplify specific transcripts.
 - D. Use targeted RT primers.
 - E. Biochemically enrich RNAs/cDNAs using capture hybridization.
 - F. Use knowledge of 5' and 3' modifications (e.g., miRNA with 5'-phosphate and 3'-hydroxyl)
3. Select newly made RNAs
 - G. Fractionate chromatin-associated RNAs.
 - H. Only consider intron-containing RNAs.
 - I. Immunoprecipitate RNA PolII engaged RNAs.
 - J. Metabolic labeling with short pulses.
4. Select modified RNAs
 - K. Immunoprecipitate with a modification-specific antibody.
 - L. Use chemistries that induce RT-stops or mutations.
5. Select RNAs from specific cells.
 - M. Microdissect cells or FACS sort cells of interest.
 - N. TU-tagging (targeted metabolic labeling of RNAs in certain cells)
 - O. Single-cell RNA-seq (scRNA-seq)

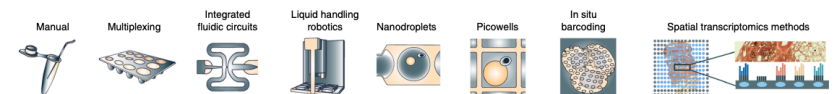
Examining cell heterogeneity with scRNA-seq



Bulk RNA-seq averages over the RNA content of many cells masking differences.

These differences can be revealed by sequencing the RNA from individual cells using single cell RNA-seq (scRNA-seq)

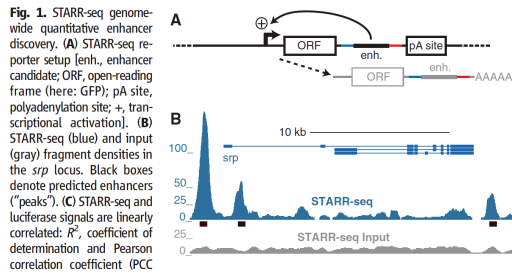
Analysis of RNA transcripts in individual cells can reveal rare cell populations and lineage trajectories.



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.

Almost any assay can be adapted to sequencing!

- CRISPR screens
- Massively parallel reporter assays (MPRA)

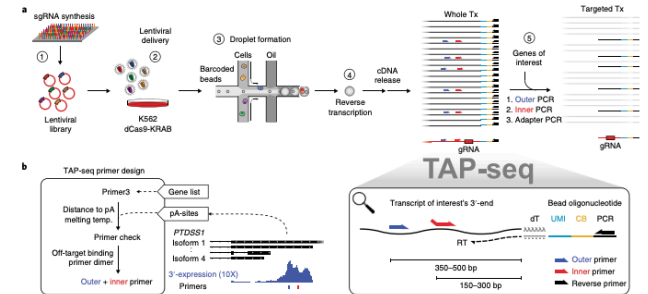


Arnold CD, Gerlach D, Steitzer C, Boryk LM, Rath M, Stark A. Science. 2013 Mar 1;339(6123):1074-7. PMID: 23328393.

Summary

- Genomics I: Deep sequencing gives us access to information on a genomic level.
- Genomics II: These approaches provide a diverse set of tools to study life at a genomic scale.
- * Sophisticated use of data from genomics requires an integrated understanding of the biological experiment, sample preparation and down stream computational analyses of the data.

Example. Targeted Perturb-Seq



Schraivogel D, Gachwind AR, Milbank JH, Leconte DR, Jakob P, Mathur L, Korbel JO, Mertens CA, Velten L, Steinmetz LM. Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat Methods. 2020 Jun;17(6):629-635. doi: 10.1038/s41592-020-0837-5. Epub 2020 Jun 1. PMID: 32483332; PMCID: PMC7810614.