Multi-omics Principles and Applications

Yuchen SHEN, ZJU-UoE Institute, Zhejiang University

Originally drafted for GP2 final examination

ZJE 浙江大学爱丁堡大学联合学院
ZJU-UoE INSTITUTE

Omic Supermap

Yuchen SHEN

Table of contents

Next generation sequencing

scRNA-seq, ChIP-seq, ATAC-seq... are the adaptations of NGS. Performing genome assembly, variant calling, RNA quantification, finding open chromatin are the following actual applications of NGS.

Next generation sequencing

Sequencing by synthesis (SBS)

Approaches

Note: sequencing by synthesis is the underlying principle of next-generation sequencing. It tells us that to get *sequence information*, you first have to provide the base for synthsis, while *single-end sequencing*, *paired-end sequencing*, and *mate pair sequencing* describe in what way do you collect base information. What NGS essentially captures is the DNA information, instead of the possibility to capture the RNA base identity.
Principle: the longer each read is, the easier it is to map it to the reference genome.

- **Single-end sequencing**
- **Paired-end sequencing:** short-insert
- **Mate pair sequencing:** long-insert

Steps

- Randomly fragment genomic DNA
- Ligate adapters to both ends of the double-stranded DNA fragments
- Denature to single strand
- Randomly bind single-stranded DNA fragments to the inside surface of the flow cell channels
- Bridge amplification using *unlabeled* neuclotides
- Denature the double stranded DNA
- Add 4 types of *labelled* neucleotides, terminators, primers and DNA polymerase
- First chemistry cycle
  - One-base amplification then terminated
  - Image by laser excitation to get the first base
  - Remove the terminator from 3' terminus
  - Remove the fluorophore from 3' terminus
- Repeat the chemistry cycles

Data preprocessing

The raw, intact data obtained from the sequencing machine are in FASTQ format.

FASTA format

FASTA format is for the reference genome which can be acquired from *UCSC* and *ENSEMBLE*.

- First line: reference genome name starting with >
- Second line: sequence it self

```
>21 dna:chromosome chromosome:GRCh38:21:1:46709983:1 REF
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCTAACCCTAACCCTA
```

Note: Sometimes you need a specific version; most other times it may be the best to use the most recent version.

FASTQ format

The fastq format is the same for both DNA and RNA NGS results.

- Line 1: identifier (ID)
- Line 2: read data (AGCT)
- Line 3: anything (usually "+")
- Line 4: quality score (the probability of ERROR)

Initial quality control

- Tool: FastQC
- Information: visualization of distribution (box plot) of quality scores across all reads
- Green-orange-red graph
  - Indication
    - Green: very good
    - Orange: good
    - Red: bad
  - Distinguish between good & bad
    - Good
      - Decay is normal
      - Initial bases are within high quality range
      - No box plot average in red
    - Bad
      - Decay is normal

            - Initial bases are not in high quality range
            - Some parts of box plot average in red
- N-content graph
  - N = not A nor G nor C nor T
  - Red peak can be visually examined

Alignment

- Definition: the process of reads get mapped to the reference genome
- Algorithms
  - BLAST (basic local alignment search tool)
    - *Not ideal*
    - Too slow
    - BLAST's algorithm assumes that errors in the reads are randomly distributed. In NGS data, errors are often clustered
    - NGS does not need the complete alignment, just the location it maps to
  - Burrows-Wheeler transform
    - Steps
    - Advantages
      - Fast
      - Begins alignments at the end of the k-mer which corresponds to the high-quality 5' end of sequencing reads.
    - Applications
      - bowtie2: for spliced alignment (genome)
      - STAR
      - bwa: for fast and accurate short read alignments (genome)
- Files
  - Both SAM and BAM formats store the same alignment information, but BAM is a compressed, binary version of SAM.

Alignment quality control

- Both with software reports (shallow way) or with down-stream analysis (deep-diving)
  - Shallow
    - Software reports % of reads mapped/unmapped
  - Deep-diving: Ask experiment-specific questions
    - More Y chromosome read in male and more X chromosome read in female?
    - Able to find known marker genes in RNA-seq?
    - Similar results in replicates?

Genuine variant identification algorithms

- Identification of variants from sequencing results = Variant calling
- Methods
  - Allele counting
  - Probabilistic methods (statistics)
    - Quantify statistical uncertainty based on observed allele frequency of multiple samples
  - Heuristic approach (thresholds)
    - Thresholds for read depth, base quality, variant allele frequency…
    - Most common method

Summary

Barcodes & UMI (unique molecular identifiers)

Barcode

- What: a short (8-12 nt) DNA sequence
- Other names: DNA barcoding, DNA indexing, indexing
- Feature: unique to each sample
- Purpose: identify sample source to enable multiple sample sequencing in one run, a process called multiplexing
- Significance:
  - Enhanced throughput with lowered cost

  - Metagenomics studies: multiple species or strains can be analyzed simultaneously in environmental samples

UMI

- What: also a short DNA sequence
- Alias: molecular barcodes
- Feature: unique to each read
- Purpose: identify and eliminate PCR duplications (not "prevent" PCR duplications) to enhance quantification accuracy

Applications

Principle: applications of NGS are interpretations of mapped data from different perspectives.

- Genome assembly
  - *de novo* genome assembly *(details see next chapter)*
    - Do not need a reference genome
  - Reference-guided assembly
    - Need a reference genome
- Genuine variant identification (variant calling, *details see next chapter*)
  - After getting studied genome, variant calling is followed, but as a separate process from genome assembly
  - Methods
    - Allele counting
    - Probabilistic methods (statistics): quantify statistical uncertainty based on observed allele frequency of multiple samples
    - Heuristic approach (thresholds): thresholds for read depth, base quality. variant allele frequency. The most common method.
- Transcript Assembly / Quantification
  - Reads are mapped to different transcripts or isoforms
  - The basis of DEG analysis
- Peak calling in ChIP-seq
  - Regions with a local enrichment of sequencing reads are called as peaks
- Genome re-sequencing
- Clinical applications
- Sequencers as counting devices

Wet-lab considerations

For successful sequencing results, it is necessary to provide a nucleic acid sample for sequencing that has: (i) accurately measured concentration, (ii) sufficient purity, and (iii) excellent integrity.

DNA

An accurate concentration of DNA is necessary for optimal cluster density in NGS sequencers.

RNA

The integrity of RNA is often problematic, as RNA is easily degraded. RNA integrity can be expressed as *RIN value* (RNA Integrity Number), from 1 to 10, with 1 being the most degraded and 10 the most intact. For most experiments, an RIN value greater than 8 is desirable. Unwanted RNA degradation will have a major effect in calling DEGs.

Single molecule sequencing

Also called third generation sequencing.

Single molecule sequencing

Advantages in *de novo* assembly

- Longer read lengths
  - Single molecule sequencing technologies, such as PacBio and Oxford Nanopore, produce much longer reads (10–100 kb) compared to traditional short read sequencing (150–300 bp).
  - This significantly simplifies the assembly process by resolving complex regions such as repeats, structural variants, and GC-rich regions.
- Improved contiguity
  - Longer reads bridge repetitive regions, which means
    - More contiguous
    - More complete genome assemblies
    - Fewer scaffolds to deal with
- Less amplification bias
  - Single molecule sequencing does not rely on PCR amplification, avoiding amplification biases
- More accurate SV detection
  - Long reads enable the direct identification of large insertions, deletions, and other structural variants that are often missed or misrepresented by short read assemblies.

Application in variant calling

Category	Long-read	Short-read
Accuracy	Superior in calling variants in repetitive or complex regions of the genome due to its ability to span large segments of DNA uninterrupted	Struggles with repetitive or complex regions due to short fragment length, often resulting in fragmented assemblies
Large indel detection	Much better at identifying indels accurately since they can capture entire insertion or deletion events, even when they span several kilobases	Limited in capturing large indels accurately; small indels are detectable but with less clarity in complex regions
Throughput	Lower	Higher
SNP detection	Traditionally less accurate for SNP calling due to higher raw error rates	More accurate for SNP detection due to lower raw error rates; ideal for single nucleotide precision
Best strategy	Excellent for structural variants and de novo assembly; critical for resolving complex genomic regions	Optimal for SNP detection and high-throughput applications
Hybrid Approach	Best strategy is to combine both technologies: Long-read for structural insights and short-read for high-accuracy SNP detection.	Provides high-depth coverage for fine-resolution variant calling, complementing the broader structural insights from long-read data.

Challenges

- Higher sequencing error rates: although high fidelity reads and error-corrected reads technologies are improving this
- Cost and throughput: more expensive with a lower throughput per run
- Computational requirements: more complex computational analysis for error correction, alignment, and assembly
- Storage and data handling: the large volume of data generated by long read sequencing demands substantial storage and efficient processing pipelines

Applications

- *de novo* assembly
- Variant calling
- Epigenomics: directly detect base modifications (e.g., methylation) without requiring bisulfite conversion, providing more original epigenetic information
- Transcriptomics: captures full-length transcripts for accurate gene structure annotation and identification of splice variants

Whole genome sequencing

Genome sequencing = DNA-seq. The ultimate goal is to know the complete genetic information

Whole genome sequencing

What is genome

- Definition: the *complete* set of genetic information of an organism.
- In human: 23 pairs of chromosome in nucleus + small circular DNA in mitochondria

Genome sequencing workflow

- Assembly process
- Evaluation of assembly quality
- Genome annotation

Assembly process

- Collect sample like cells, tissues…
- DNA sequencing, options include:
  - Sanger sequencing
  - NGS
  - Single molecule sequencing
- Pairwise read overlaps (find the overlaps between two reads and attach them together)
- Draft string graph
- Construct contig
  - Contig is scattered but complete sentences
  - Contig is linear sequences assembled by overlapped reads
- Construct scaffold
  - Scaffold are a paragraph which consists of complete sentences despite some gaps
  - The relative positions of contigs are known though some of the spaces between them is not known
  - Contig connected by large-insert reads produced by paired-end/mate-pair sequencing
- Fill in the gaps
  - Direct PCR, primer walking, or Hi-C

Evaluation of assembly quality

- High contiguity (measured by *N50*)
  - Sort all the contigs by length from largest to smallest, and then cumulatively add their lengths until the total reaches 50% of the total assembled genome length. The length of the contig at this point is N50.
  - AKA the median contig length.
  - The higher N50, the better the contig set is.
- High completeness
  - = coverage
- Accuracy

Genome annotation

- Definition: a process of attaching biological information to sequences (either contigs or chromosomes)
- Steps
  - Structural annotation: identify elements through computational gene prediction methods (gene identity, start/stop site, intron, exon)
  - Functional annotation: attach biological information to these elements (protein-coding/non-coding, signaling pathway, more advanced...)
- How to find a gene from just its sequence: (1) identical to a known gene *in the same species* (2) highly similar to a known gene *in another species (BLAST)* (3) highly similar to some *other functional data* (e.g. massspec proteomics, or possibly RNA-seq)

Variant calling

The reference genome of the species is known. We wish to identify variants throughout the genome of that species.

Types

- Single nucleotide aberrations
  - Single-nucleotide polymorphisms (SNPs)
    - Differs from SNVs for just frequencies
    - Mutation is well-characterized and found at an appreciable frequency
    - Documented
  - Single-nucleotide variations (SNVs)
    - Mutation found in only one or a few individuals
    - Rare
- Short indels
- Larger structural variants (SVs), *such as,*

  - Copy number variation
  - Short indels
  - Inversion, translocation
  - Chromosome rearrangement

Workflow

- 1. Mapping
  - Objective: align reads to a reference genome
  - Tools: bwa, bowtie, Novoalign
  - Conceptual correction:
    - Genome assembly $\neq$ alignment
    - Mapping = alignment
    - Alignment is a step of variant calling
  - Process:
    - Reads getting mapped
    - Initial alignment results are refined
      - Local realignment around indel regions
      - Removal of PCR duplicates
    - Output stored in a sequencing platform-independent format called SAM/BAM
- 2. Discovery of raw variants
  - Objective: identify which sites with an alternative alleles have statistical evidence
  - Tools: GATK, VarScan2, SAMtools
  - Process:
    - Uses SAM/BAM files to count alleles
    - Apply statistical methods to identify variants
      - Direct allele frequency counting
      - Probabilistic methods: estimate the probability of a variant existing
      - Heuristic approaches: rule-based threshold identification
  - Output: VCF (Variant Call Format) containing raw SNVs and indels
- 3. Variants analysis
  - Objective: filter and annotate raw variants to identify truly polymorphic sites
  - Tools: Variant Effect Predictor (VEP)
  - Process:
    - Filtering: based on quality metrics (e.g., read depth, quality score).
    - Annotation:
      - Synonymous variants
      - Non-synonymous variants
      - Stop-gain variants
  - Output: a VCF file containing analysis-ready variants

VCF file

Structure

- Header
  - VCF version, variant caller (tool) version, and metadata annotations
  - Mandatory header line: `##fileformat=VCFv4.0`
  - Optional header line: other information
- Body
  - Each data line contains information about a single variant
  - `#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2`

Example

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",t
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ...
```

Applications

- Identify novel pathogens and their evolution
  - SARS-CoV-2 genome and their evolution: draw a evolution tree
- Microbiome analysis
- The 1,000 genomes project
  - Identify all common genetic variants with frequencies of at least 1 across global populations
- Pan-Cancer Analysis of Whole Genomes (PCAWG)
  - Identify common cancer mutation patterns across whole genome
  - Identify primary cancers and their matching normal tissues
- GWAS

FAQs

The conceptual relationship between NGS and WGS?

- NGS are single molecule sequencing are two milestones in WGS technology.
- But Sanger sequencing can be used for WGS too!
- That is, NGS and single-molecule sequencing are the technology, while WGS is one of the main applications of NGS and single-molecule sequencing, because WGS basically equals *de novo* assembly.
- WGS has its application, too.

Genome spatial information by sequencing

How transcription is regulated by genome spatial conformation? How does chromatin conformation capture work?

Genome spatial information by sequencing

Only 2% of the genome encodes protein, whereas more than 10% are now explored regulatory regions.

Enhancers

Definition

Enhancers are *cis*-regulatory elements. It can act, even through long distances, to regulate the activity of promoters by making a DNA loop to attach itself to promoters and thus potentiate transcription.

ENCODE

- 12.6% of genome is devoted to enhancer function
- A piece of sequence cen be a promoter *and* an enhancer at the same time.

Long-distance regulation

- *TAD (topologically associating domains)* is the structural unit of chromatin. The enhancer and promoter located within the same *TAD* tend to interact more often. The correlation of certain histone modifications is higher in one TAD than between TADs.
- The border line of *TAD* is maintained by *CTCF (insulator transcription factor)* and *Cohesin*. Specifically:
  - *Cohesin* suppresses compartments but is required for TADs and loops
  - *CTCF* define *TAD* boundaries and loop anchors
  - *WAPL* define the length of loops
- *TADs* are conserved across species, across all cell types.
- By *chromatin looping*, the spatial distance of the enhancer and its linked promoter will be reduced. That is, ***linear (sequence) distance* $\neq$ *spatial (actual) distance***.
- Genes and their regulatory elements tend to be segregated into TADs that are insulated from adjacent TADs.

Enhancer trap

- Enhancer trap is a tool to localize the enhancer and identify its functions.
- The basic methodology is to randomly incorporate a small segment of the sequence (inactive minimal promoter + linked reporter gene such as *lacZ, GFP*) into the mouse / fruit fly genome.
- If there is an active enhancer near the location of insertion, the reporter gene will be expressed.

From *3C* to *Hi-C*

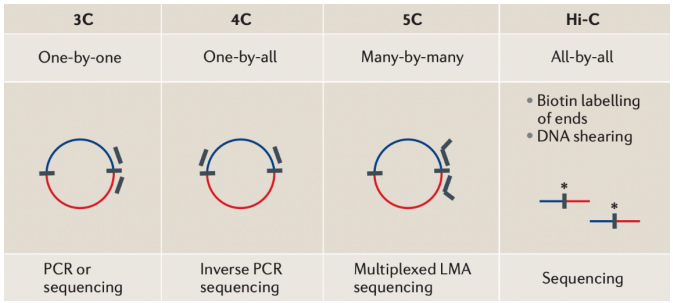
Both 3C and Hi-C are chromatin conformation capture technologies.

Goal

Chromatin conformation capture (CCC, 3C) is a technology that aims to determine whether two DNA segments are close enough *spatially*.
3C experiments can be used to study coregulated genomic regions.

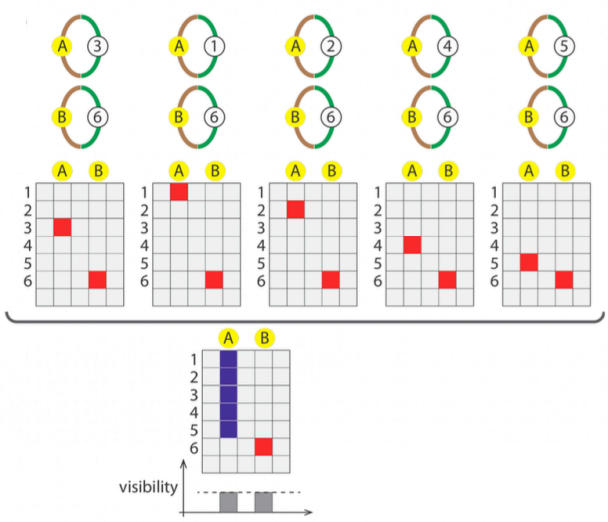
Methodology

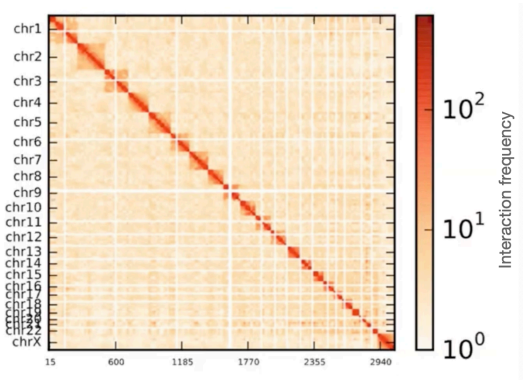
1. Crosslink to preserve native 3D structure.
2. Restriction enzyme digestion. If two DNA segments are close enough, they will be adjacent after digestion.
3. Ligation. If two DNA segments are close enough, they will be ligated.
4. De-crosslinking and purify DNA with PCR.



Hi-C adds biotinylation which, after sequencing, provides a matrix that records the contact frequency of one chromatin position with another.

Contact heatmap





- x and y axis: different position on chromatin.
- Color/number: interaction frequency.
- Interpretation:
  - Diagonal line: sequentially close, not surprised
  - Out of diagonal line: one enhancer "touched" with promoter. We should be surprised.
- The effect of resolution on results
  - 50kb resolution: fine compartments
  - 10kb resolution: TADs
  - 5kb resolution: loops

From *FISH*, *smFISH* to *MERFISH*

FISH

Fluorescent *In Situ* Hybridization (FISH) used fluorescence-labeled DNA or RNA probe to tag the target DNA or RNA in the cell.

smFISH

Single-molecule Fluorescent *In Situ* Hybridization (smFISH) uses multiple short probes to target the same mRNA. Every probe has a florescence label with different colors. So after hybridization, the targeted RNA will exhibit strong fluoresce.

- Exon probe: tag mature mRNA
- Intron probe: tag translating mRNA
- Merge with DAPI: tag mRNA and expressing region

MERFISH

Multiplexed Error-Robust FISH (MERFISH) = FISH (to visualize DNA or RNA) + super-resolution microscopy (to visualize single molecules) + multiplexing

- Use one fluorescence-coded probe to represent one mRNA (one color combination = one mRNA)

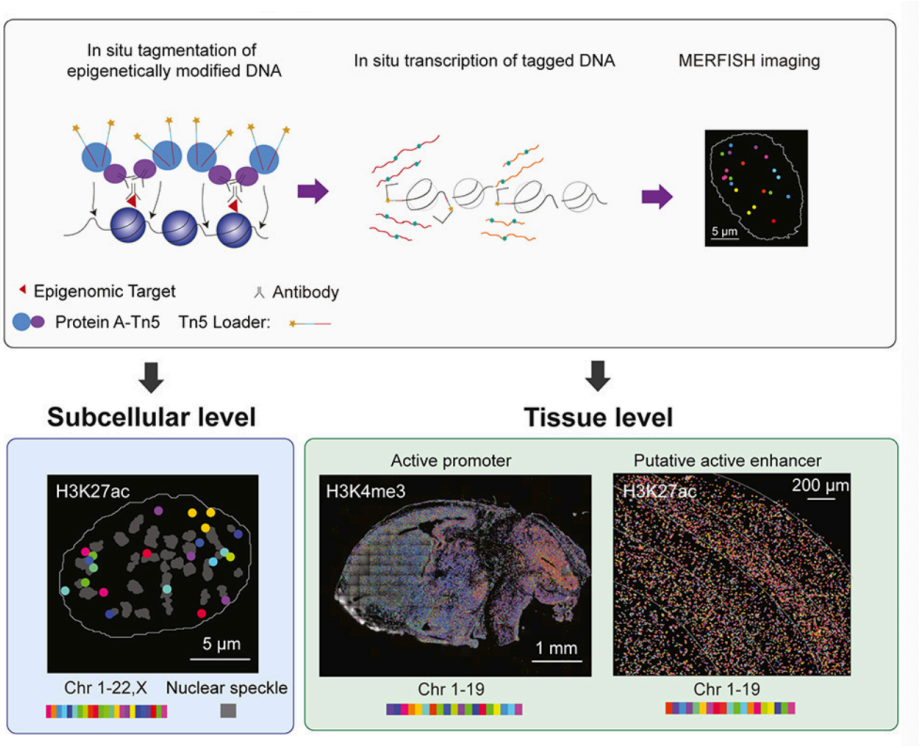
- Allows for testing thousands of mRNA at the same time
- Each fluoresce is activated to exhibit fluorescent light at one sequential time points
- Use sequential florescence color combination to distinguish a specific RNA

Applications

- Cell typing
- Use RNA velocity with pseudotime to determine cell state transition
- Anything that can be done with scRNA-seq

EpiMERFISH

Expand the functionality of MERFISH to DNA to tag chromatin modification localization.



DBiT-seq

Deterministic Barcoding in Tissue for spatial omics sequencing (DBiT-seq) is used for high-spatial-resolution multi-omics profiling. Basically, it is spatial scRNA-seq + proteomics.

- A tissue section is placed on a microscope slide.
- The first microfluidic chip flows barcode A1–A50 across the tissue (horizontal direction), labeling both RNA and protein. Reverse transcription occurs *in situ*, converting mRNA to cDNA with barcode A.
- The second chip flows Barcode B1–B50 in the perpendicular (vertical) direction.
- At each intersection point, Barcodes A & B are ligated to form a unique spatial coordinate (X, Y).
- The resulting cDNAs contain both transcript and protein information and spatial coordinates.
- These are then sequenced to reconstruct spatial expression maps.
- Reveals concordance (or discordance) between transcription and translation spatially.

Transcriptomics

The study of genes, specifically the protein-coding genes. Not all RNAs encode proteins. However, transcriptomics only includes mRNAs. The gold standard is wet lab validation experiments. Ideally qRT-PCR or microarray.

Transcriptomics

Comparison of DNA-seq & RNA-seq

If NGS is used for genome sequencing, the purpose is mainly to assemble the genome or call variants. However, if NGS is used for the transcriptome, the purpose is to quantify gene expressions.
mRNAs are spliced, so they contain only exon information, while the genome is not spliced, so they contain all the genetic information. So RNA-seq can only capture gene *expression* information, instead of complete genome sequence which is something DNA-seq do.
In DNA-seq, the coverage level (read count) represents sequencing depth (how many times a sequence is sequenced); however in RNA-seq, the coverage level *not only* represents sequencing depth, but also, most importantly, gene expression level.

Workflow

- Collect sample
  - Replicate samples
  - Comparison samples
- Isolate RNA from samples
- Fragment RNAs
- Reverse RNAs into cDNAs
- NGS on cDNAs
- **Count** the gene expression: map sequencing reads to the genome or transcriptome *(Step 1 of expression analysis)*
  - Coverage (level) measures read number mapped to each gene
  - Tools:
    - Gapped aligners
      - Extended from previous mappers for genome assembly
      - Allow gaps in alignment to improve mapping
      - Tools: bowtie2, STAR, bwa, Cellranger
      - More precise
    - Pseudo-alignment
      - Estimates which k-mers of transcripts each read belong to without full alignment (not necessarily need base-to-base alignment)
      - Tools: kallisto, salmon
      - Faster, more efficient, less computationally intensive
      - Need pre-existing transcript models to predict associations
- **Normalize** read count to library depth *(Step 2 of expression analysis)*
  - Context: raw counts are not enough for fair comparison
  - Purpose: adjust various variation factors across samples
  - **Approach 1**
    - Different sample → different sequencing depth
      - Counts → FPM (fragments per million)
      - Process: total read counts of all genes per sample represent the sequencing depth of that sample (after unit conversion divided by million)
    - Different gene length → different mapping coverage
      - FPM → FPKM (fragments per kilobase per million)
  - **Approach 2**
    - Different gene length → different mapping coverage
      - Counts → FPK (fragments per kilobase)
    - Different sample → different sequencing depth
      - FPK → TPM (transcripts per million)
  - **Approach 1** and **approach 2** are both correct and the same in terms of effect, but a different order gives different values. Only TPM has the same total library depth across samples, whereas FPKM does not.
- **Differential expression analysis** (Step 3 of expression analysis)
  - Purpose: to identify whether the differences between groups are statistically different.
  - Previous gaps:
    - What test to use?
      - T-test requires normality which is violated in FPKM/TPM measures even after log transformation
      - Traditional tests requires more samples, but RNA-seq the sample number is often less than 4
      - Poisson distribution failed for the assumption of equal mean and variance often not hold
      - Solution: DESeq2 count data using a negative binomial distribution to account for the observed

overdispersion (variance greater than the mean)
* What to do for the increased false positives?
  - Bonferroni: lowers the chance of false positives but increases false negatives
  - FDR (False Discovery Rate)/Benjamini-Hochberg: more balanced, commonly used in RNA-seq
  - Q-value/Storey Method: estimates the proportion of false positives

Normalization *in depth*

Goal

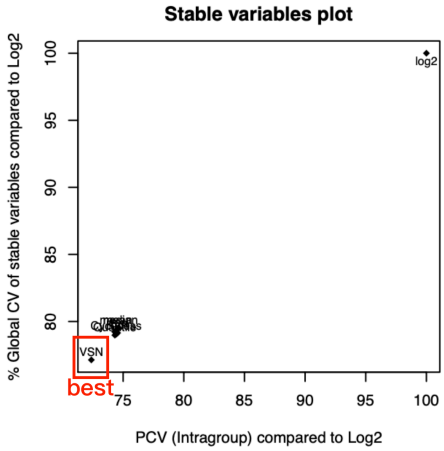
- Control for unwanted biological, but mostly technical variation such as *batch effects* and *sequencing depth*.
  - Batch effects are technical sources of variation that have been added to the samples during handling, for example, RNA purification was performed in two different days due to large sample number.
- Minimize the variation between the *replicate* samples.
- Ensure a reliable comparison of *between* samples.

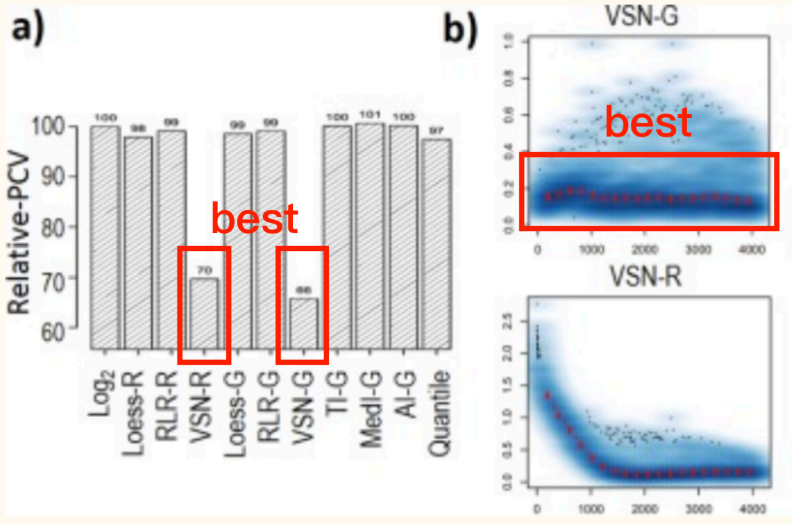
Features

- Different types of 'omics' experiments including microarrays, bulk RNAseq, scRNAseq and proteomics may prefer a specific type of normalization.
- There is no single normalization procedure that works for all cases because the technical variation in each experiment is unique.
- Although in general normalization should have a positive effect on the data, it is still possible that inappropriate normalization makes your analysis worse due to "overfitting" of your data.
- The best thing to do when comparing the effects is to try out several methods and have some sort of quantitative approach for evaluating the result.
  - Statistical means
  - Enrichment results in response to the normalisation

NormalyzerDE

R package `NormalyzerDE` for evaluating the effect of various normalization methods.





Networks and enrichment analysis

Network terminologies

- Node: an object within the network, such as a protein or gene
- Edge: the connection between two nodes
- Degree centrality: the number of connections a node has. It indicates how well connected a node is within the network. A higher degree centrality means the node has more connections, suggesting it may be more influential or central in the network
- Hub: high-degree nodes are often considered "hubs" in the network

Network database

- One system to embed your hits into networks (also has function enrichment functionality) is the *STRING* database.
- But the interactions are not physical PPIs, but evidence found based on text mining, gene coexpression experiment, gene fusion… basically any evidence that can be found to support the relatedness of genes / proteins.

Application

- DEGs
- Understand diseases
- Detect new genes/transcripts
- Understand how alternative splicing or RNA editing might affect gene expression

Replicates in RNA-seq

Statistically vs. biologically significant expression change

- Statistical
  - Not due to random chance
  - p-value
- Biological
  - Cause an impact
  - (1) FC & (2) linked to known pathways or functions

Best packages for identifying DEGs in bulk RNA-seq

Using different packages instead of increasing the number of replicates can be a viable strategy to improve accuracy without extra costs.

- DESeq2
- edgeR
- limma (voom)

Microarray vs. RNA-seq

Principles of microarray

A wet lab technique to detect the expression of thousands of genes simultaneously.
The principle is hybridization between complementary DNA (cDNA) strands.
A microarray consists of a grid of tiny spots on a glass slide, each spot containing DNA probes specific to known gene sequences.

When a sample containing fluorescent-labeled cDNA is washed over the array, complementary sequences hybridize to the probes on the plate.
The amount of fluorescence at each spot is measured, which is correlated with the targeted gene expression level.

Advantages of microarray

- Cost-effectiveness: less expensive than sequencing, especially for large sample sizes
- Quick turnaround: the workflow for microarrays is faster than NGS, making it efficient for time-sensitive project
- Data analysis simplicity: data generated is less complex and easier to analyze compared to NGS data

Microarray over sequencing

- Only want to measure *some* of the *known* genes' expression level, but not all of them
- When cost-effectiveness and efficiency is prioritized
  - Clinical setting
  - Cancer study
  - Pharmacogenomics

Sequencing over microarray

Information provided by RNA-seq that are not possible with microarrays:

- Novel transcript discovery: identify novel genes, splice variants, and non-coding RNAs that are not present on microarray chips
- Single-nucleotide resolution: detect SNPs and mutations
- Dynamic detection range: allows the detection of both low and highly expressed genes accurately

Single cell transcriptomics

Single cell transcriptomics is a dimension of Single cell genomics. Single cell genomics = scRNA-seq + scDNA-seq + scBS-seq (BS is bisulphite sequencing, which detects methylation) + scATAC-seq. The biggest picture is that scGenomics allows us to study diverse cell type populations, cells at different stages or under different regulatory states

Single cell transcriptomics

Applications:

- The advance of scGenomics directly ignites the establishment of the **HUMAN CELL ATLAS** project, which describes all cell types in the human body.
- In mouse, the publication of ***Mapping the Mouse Cell Atlas by Microwell-Seq*** covered all the major mouse organs and constructed a basic scheme for a mouse cell atlas (MCA).

Bulk RNA-seq vs. scRNA-seq

Advantages of bulk RNA-seq

Detects the average transcriptional state

- Higher quality
- Cheaper
- Less batch effect
- Less technical variation
- Less bias
- Standardized analysis
- Best for *pure cell populations*

Advantages of scRNA-seq

- Capture rare cell types or even new cell types
- Independent of prior knowledge
- Able to capture differentiation lineage
- Captures multimodality
- Best for *complex tissues*

Sequencing platforms

Smart-seq2

for each cell, sort it into a well on a plate then lyse it
Compared to 10X Genomics, this method is low-throughput, labor-intensive, and high abundant-transcript bias (dominated by highly expressed genes), not strand-specific (cannot discriminate whether an mRNA is forward or reverse), but high quality.

- Poly A capture
- Reserve transcription
- Template switching to complete double-stranded DNA using a single primer
- Pre-amplification (only a few cycles)
- DNA fragmentation and adapter ligation
- Gap repair
- Enrichment PCR
- Add final index sequences

Drop-seq

- A cell is capsuled in a droplet which has barcoded bead with oligo-DT oligos for RNA capture
- Cell lysis, RNA hybridized to each bead
- Pull beads out of droplet
- Reverse transcription
- Now all contents are in a bead with all uniformly-barcoded DNAs
- Amplification (PCR)
- DNA-seq

10x Genomics

Compared to Smart-seq, this method is high-throughput, labor-free, strand-specific, capture 65% of cells, but of low quality.
Very similar to Drop-seq

- Mix cells and barcodes and then add beads to make emulsion (different from Drop-seq here)
- Have reverse transcription *inside* (different from Drop-seq here) droplets as well as tagging
- Sequencing library in a single tube and normal sequencing

Barcodes and UMIs

Barcodes

Identify each cell (a cell as a sample).

UMIs

Identify each transcript.

Sources of errors

Not elaborated but exist: sample preparation, batch effect, genetic background, technical variation…

Doublets

2 cells encapsulated in the same droplet. It needs *cell doublets filtering.*

Barcode errors

Barcode errors can be induced because of synthesis and sequencing. It can be corrected by a special class of *barcode error correction* methods.

Workflow (*gross*)

Sequencing: barcode and UMI

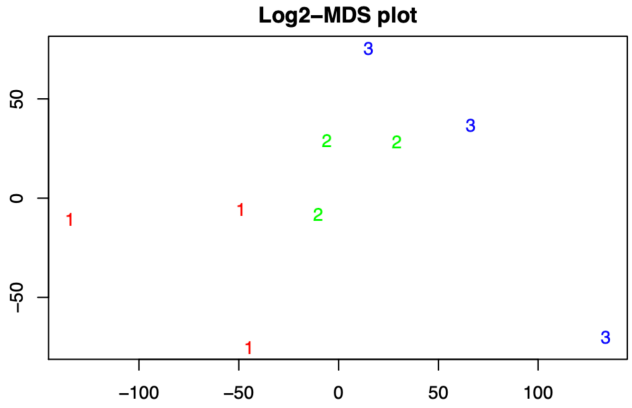
Every read consists of barcode, UMI and cDNA(read itself). cDNA is what is mapped to the reference. Every cell is an individual sample, so barcode is unique to cells.

Dimensionality reduction

Gene expression matrix → cell-cell distance matrix → cluster dentrogram → principal component analysis (PCA), multidimensional scaling (MDS), t-Distributed Stochastic Neighbor Embedding (t-SNE), UMAP.

The general idea of dimensionality reduction is to reduce the number of measured variables to a set of principal variables that capture most of the variation in the dataset.

- **PCA & MDS**: sample clustering and normalization



- **t-SNE & UMAP**: represent original points from the high-dimensional space in a lower-dimensional space. They can be seen as advanced versions of PCA and MDS.
  - UMAP appears to be computationally faster than t-SNE and better preserves the global structure of the data.
  - t-SNE is designed to preserve the local structure of the data.
    - In t-SNE plots, distances between clusters are not always meaningful in terms of global relationships.
    - Clusters (which is a local feature) suggest subsets of data that share similar characteristics.
    - Overlapping clusters can suggest regions of mixed features or transitional states.
    - Gradients or continuous paths indicate a progression.
  - Perplexity is a parameter that controls the balance between local and global aspects.
    - Low perplexity: local, for smaller datasets
    - High perplexity: global, for larger datasets

The challenges in UMAP and t-SNE applications are that you can only compare dimension reduction maps prepared with *same setting* and even using the same settings does not always guarantee that an identical map is produced.

Workflow (*detailed*)

Basic quality control

- To filter out genes which expression is igonrable: `min.cells = 3`
- To filter out cells which contribution is igonrable: `min.feature = 200` with `subset = nFeature_RNA`
- To filter out mitochondrial genes: `percent.mt < 6`

Normalize data

Make cell-to-cell gene expression statistically comparable

Feature selection

Calculate a subset of features that exhibit high cell-to-cell variation which can be used to perform PCA.

Scaling the data

- Shifts the expression of each gene, so that the *mean* expression across cells is 0
- Scales the expression of each gene, so that the *variance* across cells is 1

This step gives equal weight in downstream analyses, so that highly expressed genes do not dominate (for example when plotting a heatmap)

Dimensionality reduction - PCA

By default, only the previously determined variable features are used as input. PC is chosen based on `ElbowPlot(pbmc)` results.

Cluster the cells

Louvain algorithm (default) or SLM to iteratively group cells together.

Nonlinear dimensionality reduction - (UMAP/tSNE)

As input to the UMAP and tSNE, Seurat suggests using the same PCs as input to the clustering analysis (note that while this simplifies the analysis, it also means that the PCA and non-linear reduction are not independent.)

Examine marker gene expression

DGE helps to find marker genes.

Single cell spatial transcriptomics

Name methodologies used to study mRNA distribution in the cell.

Single cell spatial transcriptomics

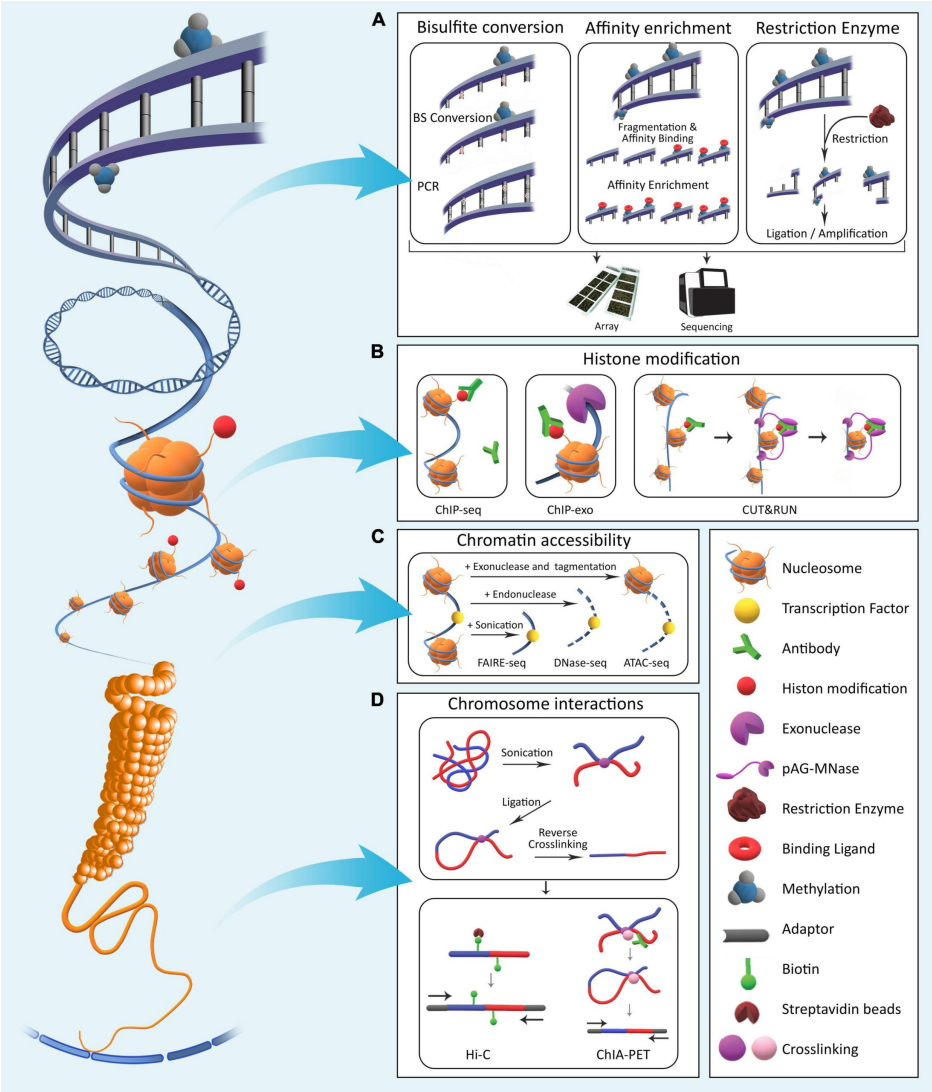
- Single cell spatial transcriptomics is a dimension of spatial multi-omics.
- Why spatial multi-omics: single-cell technologies sacrifice crucial information by taking cells out of their biological context.
- Underlying principles: mRNA is captured from intact tissue using barcoded oligos that can be matched to a particular position in the sample

Epigenomics

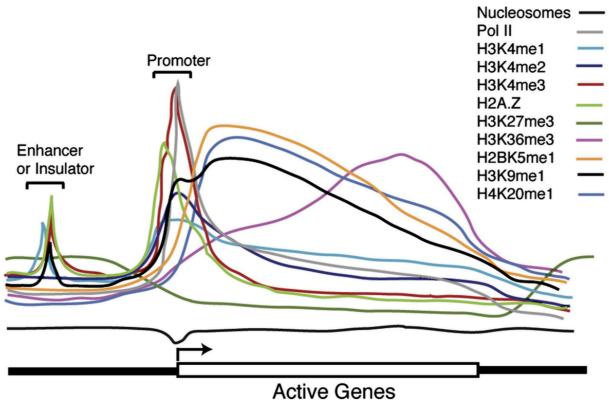
Epigenetics involves the use of a wide array of technologies. Quite a number of them overlap in applications

Epigenomics

Overview



Histone modifications can be divided into those on the coding regions and those on the regulatory regions (promoter, enhancer).



DNase-seq

- Purpose: map DNase I hypersensitive sites (DHSs) in the genome. These sites are chromatin regions that are more open and therefore more accessible, usually indicating active regulatory elements such as promoters, enhancers, and TF binding sites.
- Steps:
  - Detergent lyses cells, chromatin exposed
  - Chromatin treated with DNase I, an enzyme that cuts DNA primarily at accessible regions (DHSs), that are not tightly wrapped around nucleosomes
  - Ligate the both end of the cutout gap
  - Amplification
  - Peak = DHS enrichment = more exposed chromatin regions

ChIP-seq*

ChIP-seq is a integral technological part of the ENCODE (The Encyclopedia of DNA elements) project, which has provided standard guidelines for ChIP-seq practices.

Introduction

- Full name: Chchromatin imunoprecipitation + high-throughput parallel sequencing
- Purpose: identify protein-DNA interactions and identify the location of the genome bound by proteins.
  - Protein can be one of those…
    - TF → TF binding sites
    - RNA polimerase
    - Histone modifications
    - ... other DNA-binding proteins
- Protocol
  - Crosslinking: glue the all proteins (including the ones that are not of interest) tightly to the DNA with formaldehyde.
  - Isolating the DNA (with proteins) from chromatin. and fragmentation by sonication (open chromatin will be more fragmented).
  - Immunoprecipitation using protein-specific antibody to extract the DNA-protein complex of interest. However, some cross-reactivity may occur, which is normal.
  - Reverse cross-linking: detach protein from DNA by heating.
  - DNA Purification: eliminate the protein residuals
  - NGS: the fewer cycles, the better to reduce amplification bias
    - Alignment to genome
  - Following computational workflow
  - Repeat all above steps for control groups.

Computational workflow

- Initial QC, trimming and alignment
- Convert BAM files into BedGraph files for visualization
  - The normalized read counts are given for each position by piling up all the reads
- Peak calling
  - Purpose: to find out all the peaks in *Oct4* ChIP-seq data (the regions that *Oct4* likes to bind)
  - Use `macs2`, which builds a model to compare the tag coverage of experimental group and control group to find out the significant peaks
- Motif search
  - Purpose: to identify what DNA sequences does *Oct4* tends to bind
  - Use `HOMER`, which calculate the weights of nucleotides in each position in the peaks by comparing to random genomic sequences
- Peak annotation
  - Purpose: to find out what are the nearest genes to the peaks, and where are the peaks located to the genes
  - Use `HOMER` to annotate the peaks based on the gene annotation
- Binding profiles of *histone modifications* around *Oct4 binding regions*
  - Purpose: to find out the distribution pattern around genes of different histone modifications.
  - Use `HOMER` to generate density plot, heatmap…

Technical Concerns

- Results are highly influenced by the choice of the antibody (quality & specificity)
- Cross-linking is time dependent. Excessive cross-linking reduces antigen accessibility (harder for antibody to bind) and sonication efficiency
- Open chromatin regions are easier to shear than closed chromatin resulting in higher background noise, which will induce bias

- Sonicating for too long will disrupt nucleosome-DNA interactions: DNA fragment size should not be smaller than 200 bp to preserve necessary chromatin structure
- Sequencing efficiency of DNA regions is dependent on base composition (AAAA is different than GGGG)
- 1-2 replicates is a standard practice

Experimental design (the control group)

- **Background noise**: cross-linked, sonicated but not immunoprecipitated DNA sample (no antibody is added)
  - Provides baseline for background noise and sequencing bias. Reflects the general DNA fragmentation profile without specific protein binding enrichment.
- **Detect non-specific interactions in experimental group**: IgG mock-ChIP that used nonspecific antibodies to generate signals that should be random
  - Ideally, the IgG mock-ChIP should produce random DNA fragments with minimal specific enrichment.

Applications

1. **Histone modifications**
   - On coding regions, particularly on promoters and enhancers
   - Histones are the proteins wrap around DNA, and their chemical modifications (like methylation [mono, di, tri] and acetylation) influence gene expression. The effect varies with the type of histone (H3K4, H3K9, H3K27…).
2. **CTCF**
   - on non-coding regions
   - CTCF (CCCTC-binding factor): a major protein creates TADs (Topologically Associating Domains, a loop), which are regions of the genome that interact more frequently with themselves than with other regions.
   - CTCF binding sites act as boundaries between active and inactive chromatin.
   - Genes and their regulatory elements tend to be segregated into TADs that are insulated from adjacent TADs. However, DNA methylation can interfere with the binding of CTCF to its binding sites.

Bisulfite sequencing

Biological background

- 5-Methylcytosine (5-mC): the main form of DNA methylation in vertebrates, primarily at CpG sites
- Functions
  - Gene silencing (like X-chromosome inactivation)
  - Genomic imprinting (parent-specific gene expression)
  - Protection against transposable elements
  - Making DNA less accessible and thus reduce gene expression

Purpose

Map DNA methylation

Protocol

- Bisulfite conversion converts unmethylated cytosine (C) to uracil (U), U then replicated into T while methylated cytosine (mC, not just 5-mC) remain unchanged (still C)
- After sequencing, comparison with the reference genome reveals which cytosine (C) are methylated

MeD-seq with LpnPI

Not recommended compared to bisulfite sequencing. Technically challenging and costly, but with higher sensitivity (sequencing depth less than one-tenth required for whole genome bisulfite sequencing).

Biological principle

Same as in the previous one.

Introduction

- *LpnPI* is an enzyme that targets methylated DNA
- Applications
  - Differentially methylated regions in patients
  - Organ-specific methylation patterns
  - Promoter methylation differences in the inactive X chromosome

ATAC-seq

Introduction

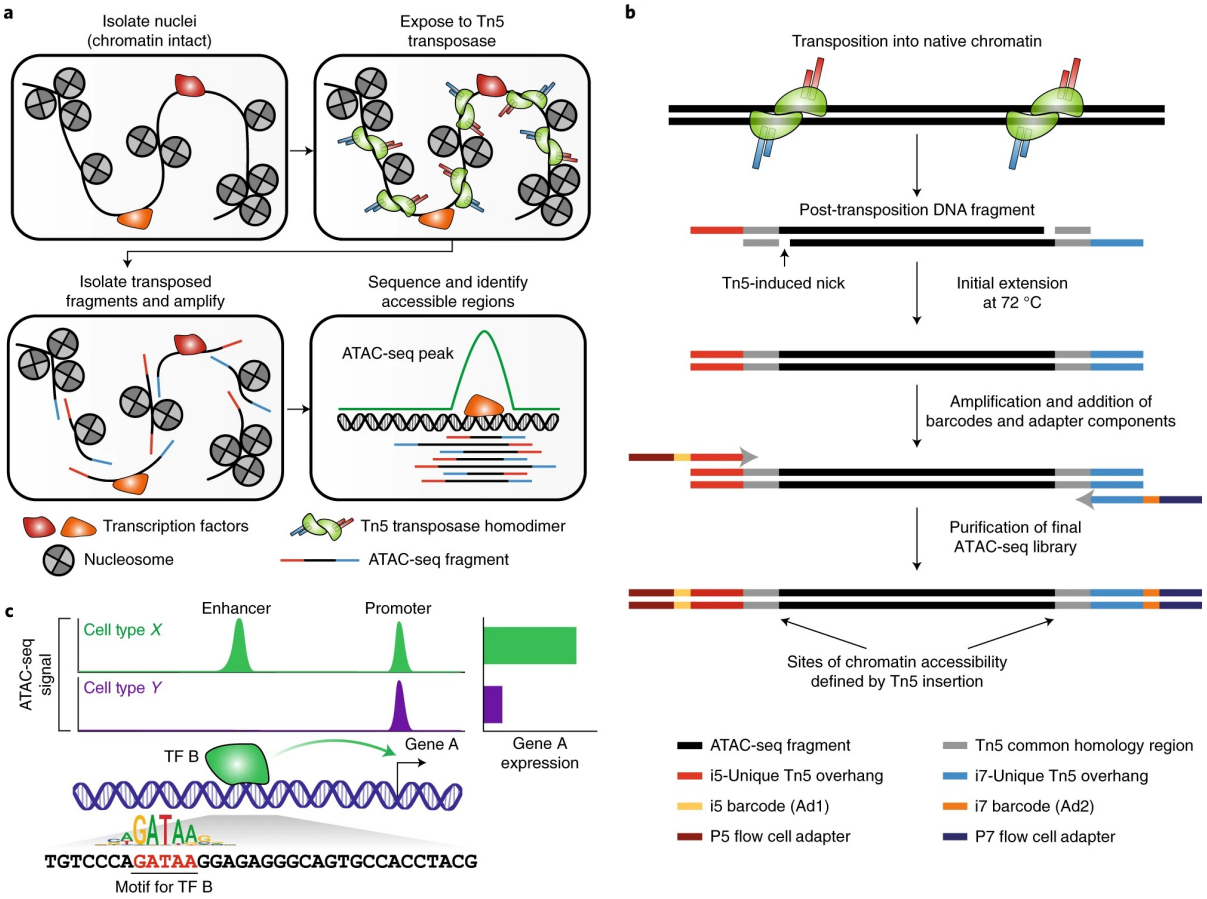
- Full name: Assay for Transposase-Accessible Chromatin
- Purpose: measure nucleosome free regions (open regions) on gene regulation
- Feature: antibody or tags free (reduced bias)
- Principle:
  - Use *hyperactive Tn5 transposase* to integrate into active regulatory regions and fragment DNA
  - Different fragment size and their enrichment relates with different elements like [mono/di/tri/tetra]

nucleosomes, CTCF, TSS, enhancer, promoter…

- Applications:
  - Detect chromatin regions
  - Detect regulatory regions like promoters and enhancers
  - Detect nucleosome packing
  - Detect nucleosome positioning
  - Detect TF footprints

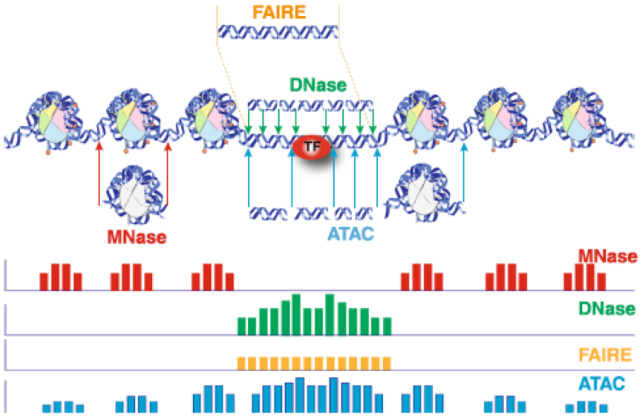
Protocol

- Insert *Tn5 transposase with adapters*
- PCR amplification
- Data processing
  - QC, read trimming
    - Remove mitochondria reads
  - Alignment
  - Peak calling
    - Normalize across samples
    - To identify TF binding regions
      - Count-based: count with statistics
      - Shape-based: model the shape
  - QC again, visualization
  - Downstream analysis



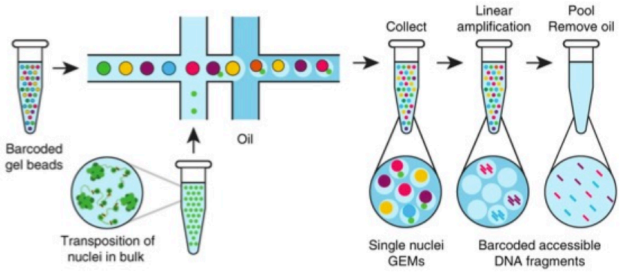
ATAC-seq vs. DNase-seq

Both ATAC-seq and DNase-seq can be used to examine chromatin accessibility. The following is their comparison.



scATAC-seq

Always used in conjunction with scRNA-seq. Any epigenetic study is meaning less if not in the context of gene expression.



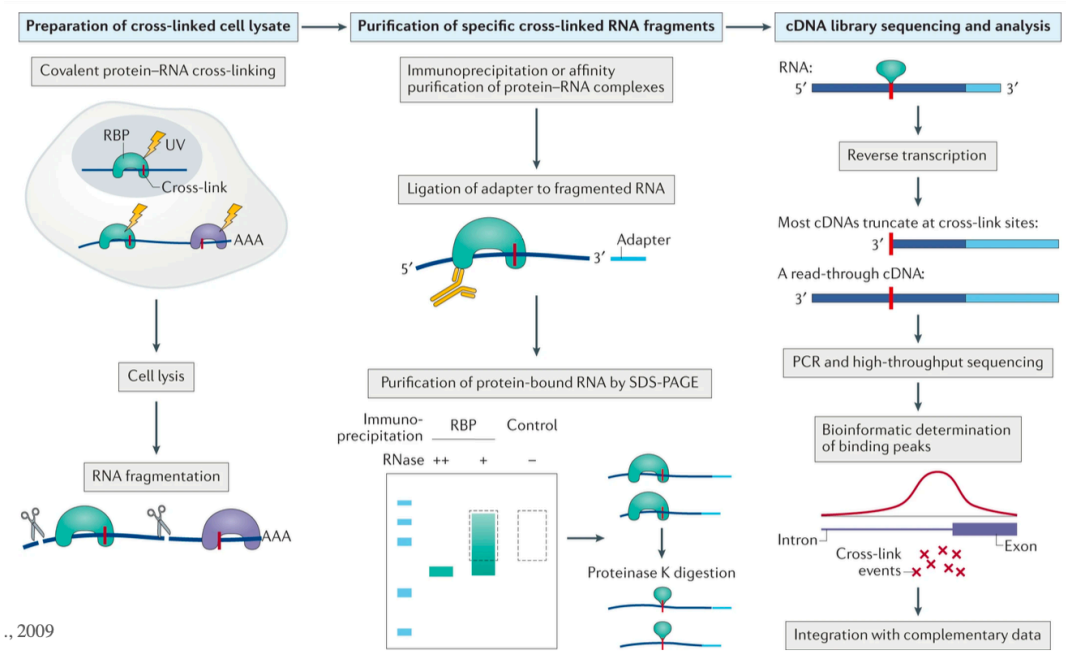
CLIP-seq

Introduction

- Full name: Cross-linking and Immunoprecipitation Sequencing
- Purpose: identify RNA molecules that interact with specific RNA-binding proteins (RBPs) to understand how RBP regulates RNA splicing, translation and stability

Protocol

The whole process is just like an RNA version of ChIP-seq.



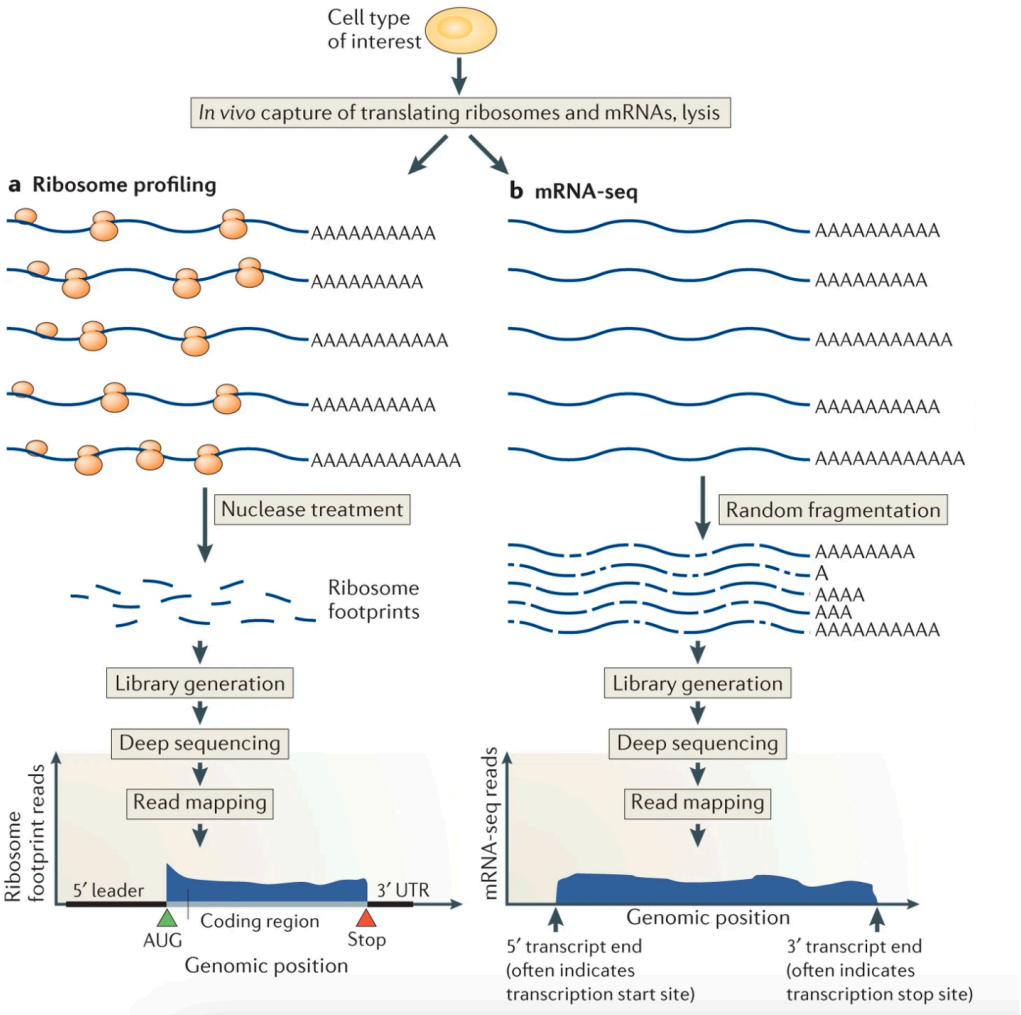
Ribosome profiling (Ribo-seq)

Introduction

- mRNA includes translated regions (from exons) and untranslated regions (from introns)
- It provides a snapshot of protein synthesis
  - Which genes are being actively translated
  - How much protein is being produced
  - Where ribosomes are pausing or moving quickly

Comparison with RNA-seq

- Ribosome profiling:
  - "Translatome"
  - Focus: from mRNA to protein
  - These regions are tended to be coding region of great biological importance
- mRNA-seq:
  - "Transcriptome"
  - Focus: from DNA to RNA
  - Includes all transcribed but some untranslated regions
- Ribosome profiling is an active subset of the transcriptome.



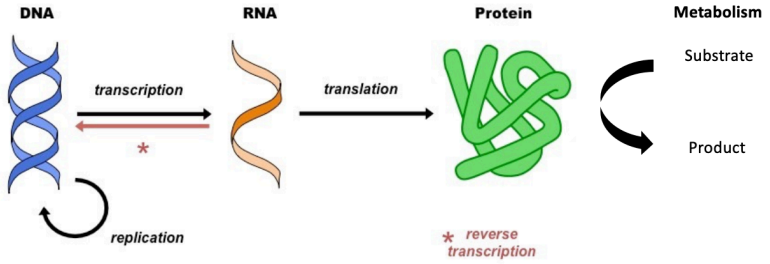
Proteomics

The most exiting thing about proteomics is that proteomics not only captures ***intracellular*** *protein information, but within all biological sample range. The main application of proteomics is to identify proteins and even quantify the protein level.*

Proteomics

Introduction

- Definition: proteomics is the large-scale study of proteomes (many individual proteins)
- Why proteomics & metabolomics: genome sequence and RNA expression do not always reflect biological perturbations. Proteomics provides complementary information to DNA/RNA based genomics data. This will be discussed in detail below.



- Background: challenging but promising
  - Proteomics data is inherently more variable than nucleic acid-based data.
  - Although genomics HAS produced clinically useful results, proteomics HAS not yet produced clinically approved biomarkers. All protein-based biomarkers in clinical use originate from the screening and use of antibodies, not from proteomics analysis.
  - But protein and metabolite biomarkers are still the best tools to analyze human physiology *objectively.*
  - DNA is "digital", proteins are not. Proteins cannot be amplified by PCR, purified in a unified way, or quantified by counting. Protein measurements are based on *signal intensity.*
  - Finding a decent proteomics service is challenging.

mRNA and protein correlation

Reasons for 0.2~0.9 correlation

The below layers of regulation vary significantly across different cell types, conditions, and species:

- Post-transcriptional regulation (splice/edit/degrade) and post-transcriptional modifications (phosphorylated/glycosylated/ubiquitinated)
- Translation efficiency in ribosome
- mRNA instability
- protein stability
- Degradation rates
- RNA change rapidly to environmental stimuli, whereas proteins have longer half-lives

Which is more important for protein levels: transcriptional or translational regulation?

The dominance of control varies by context:

- **Quick adjustment:** often, translational control has a stronger impact on protein abundance, especially for proteins that are tightly regulated or need rapid response times (e.g., stress response, cell cycle regulators)
- **Stable supply:** for highly abundant, stable proteins, transcriptional control is more predictive of protein levels

Proteomics is more direct then why we still perform RNA-seq?

- They provide complementary information, not mutually replaceable
- RNA-seq is more cost-effective and easier to perform
- RNA-seq has higher throughput
- RNA detection technologies are more sensitive, especially for low-abundance transcripts
- Transcriptomics has a more established analytical pipeline, with robust computational methods and abundant databases (transcriptome atlas)

Beyond expression levels, what information can be obtained using proteomics that is not accessible with RNA-seq?

- Post-translational modifications (PTMs)
- Protein-protein interactions
- Protein's subcellular localization
- Isoform diversity
- Protein turnover (synthesis and degradation) rates
- Functional states (active versus inactive states)

Single-cell proteomics

Challenges

The typical workflow with cell lysis, protein purification, and digestion is not practical since, for example, proteins tend to stick to plastic and are therefore lost.

The progress of scProteomics is relatively slow compared to scRNA-seq.

Benefits

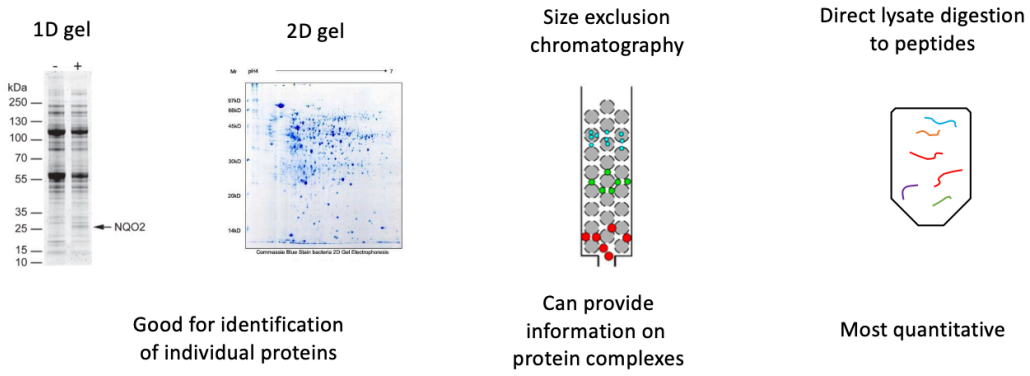
Proteins cannot be amplified to increase sensitivity, but this may actually be beneficial, which is proven by the observation that in proteome measurement, cells have higher correlation than in the droplet-based and the SMART-Seq2 method.

Applications

- When and where proteins are expressed
- Rates of protein production, degradation, and steady-state abundance
- Post-translational modifications
- Localization of proteins in different subcellular compartments (or extracellularly for secreted proteins)
- Protein-protein interactions

Separation

LC is commonly used in biological samples or other complex samples before MS. This is what the term "LC-MS" means. Separation is needed because proteome coverage may be better after the separation steps. Separation of complex mixtures allows for better coverage, but increases variance and decreases reproducibility by adding additional steps into sample processing.



Mass spectrometry (MS)

Principle

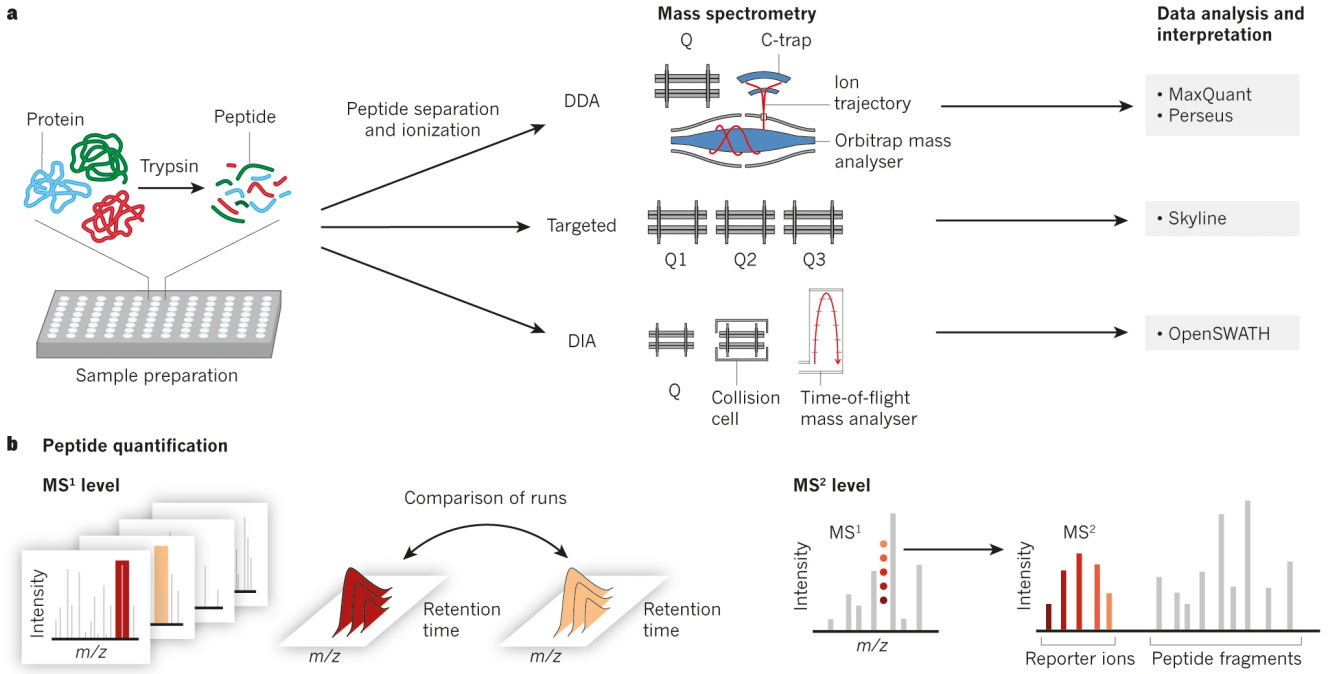
MS is used to identify and/or quantity (relatively or absolutely) molecules based on their mass-to-charge (m/z) ratio. Ions are generated by loss or gain of a charge and directed by electric field into a mass analyzer where they are separated according to m/z and detected.

The identification of proteins by MS is based on probabilities of peptides with a certain m/z rather than the exact AA sequence.

Tandem mass spectrometry (MS/MS) offers additional information about specific ions. Selected ions of interest are filtered based on their m/z during the first round of MS and fragmented by a number of different dissociation methods.

Workflow

Separation (not necessary) then detect. LC then MS.



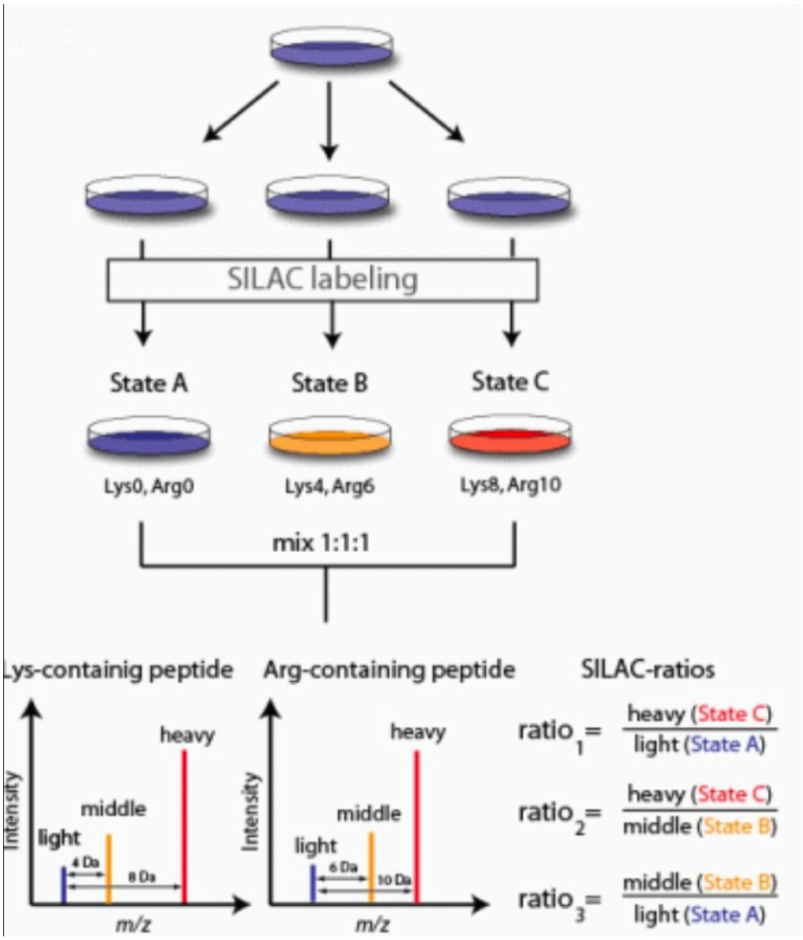
- Extracted proteins are digested by a sequence-specific enzyme such as trypsin. These smaller peptides are measured using a MS instrument.
- The observed peptide masses are compared to the calculated masses of peptides derived from *in silico* digested proteins.
- Individual peptide of a certain mass could be derived from one or more proteins, but if *many* peptides match the expected peptides derived from *in silico* digest, this is taken as a positive identification of a protein.
  - In complex mixtures such as cell or tissue lysates, perhaps only one peptide originating from a low-expressed protein are typically identified.
- Statistical rules have been defined what constitutes a reliable identification.

Considerations

- The mass measurement also has an error, usually very small. Nevertheless, it does not mean we need to eliminate it.
- Specifying too tight (small) mass tolerance when searching peptide matches may lead to missing of true matches.
- It IS possible to "sequence" individual peptides using MS, but this is rarely done. Instead, we are happy with "good enough" identification.
- MS is more likely detect abundantly expressed proteins. Typically a proteome-wide experiment detects 4000-7000 proteins, less than 15000 mRNAs got from typical RNA-seq.
  - The assumption when doing GO and network construction that your background gene set is the whole whole genome is incorrect.
  - Your hit list of proteins is enriched for high-abundance proteins such as ribosomal proteins.
  - So, the background gene set should contain the proteins that are ACTUALLY detected in the experiment.
  - For the same reason, a more relevant control for a proteomics experiment could be to randomly select a few sets of proteins from all those identified in your MS experiment.

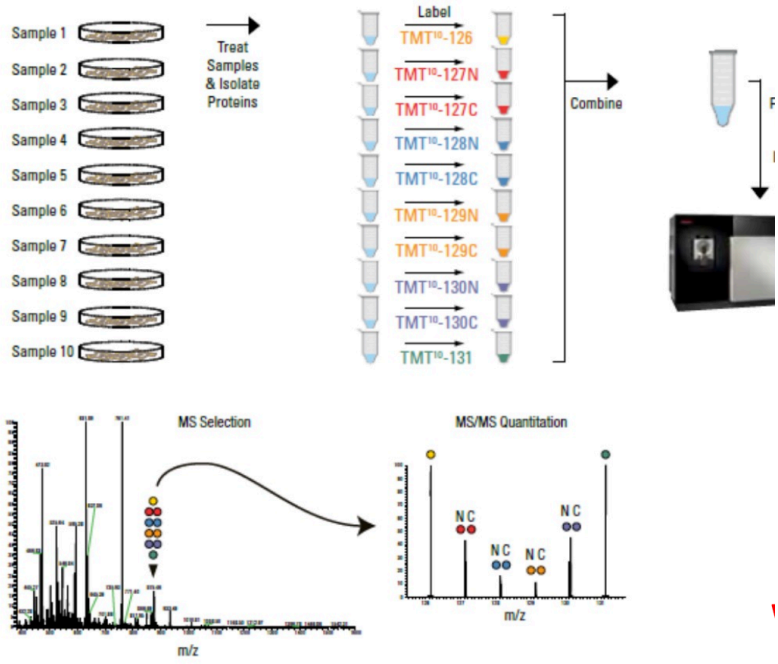
Multiplexing MS

Metabolic labeling: stable isotope labeling with amino acids in cell culture, SILAC



- **Principles:**
  - Light and heavy proteins are identified by the mass shift that distinguishes identical peptides from different samples.
  - 2~3
- **Limitation:**
  - Expensive
  - Only for cell, not for tissue
  - labeling efficiency low
- **Benefit:**
  - All the labels directly incorporated into the protein, so it is accurate.
  - No further chemical labeling, all done in cells.
  - No batch effect.

Chemical labeling: tandem mass tags, TMT



- **Principles:**
- TMTs are small molecules which contain different numbers and combinations of 13C and 15N isotopes. These are incorporated into the peptides by a chemical reaction.
- The peptides of each sample are identified as a result of their predictable mass differences.
- Resembles iTRAQ.
- 10~15
- **Limitation:**
  - Expensive
- **Benefit:**
  - High throughput
  - Applied to tissues
  - No batch effect

Data analysis

Starting point

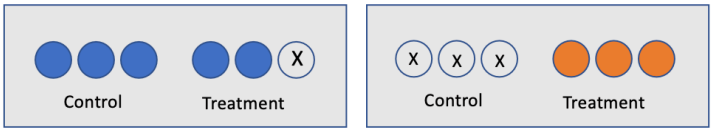
Analysis can start from
- The quantified protein levels (ion intensities)
- The *.raw* mass spectrometry files and process them yourself with MaxQuant

Inconsistent detection

If dealing with multiple MS runs, the problem that some proteins were detected only in one run but not the others is often faced. Multiplexing by SILAC or TMT helps solve this problem.
How to deal with them?

What to do with missing values? Filter them out! Why?

Because they are likely to be low abundant proteins, with low confidence quantification
But1, it might be good to keep proteins that are quantified in at least two out of three replicates.
But2, what about if the protein is detected in all treatment samples and none of the control samples?



Minor issue? *Biologically meaningful?*

Variance

Variance can be low for these several factors:

- Isolation (tissue dissection and protein extraction)
- Trypsin digestion and clean-up
- Run-to-run instrumental variance
  - Solved with multiplexing
- Long term (>2 week) drifts in instrument stability
- Labelling of proteins

Solution: normalization. Choose a conservative normalization option! This will greatly improve "hit" connectivity in network construction.

Challenges

- **In terms of experiment**
  - Low abundant protein: deep proteome coverage is not of particular quantitative meanings due to challenges in sampling of low abundance peptides. This is especially challenging in serum/plasma.
    - Solution: use *Plasma Immunodepletion Kit* (based on immunoaffinity ligands and antibodies) to remove 20 most abundant proteins from plasma to enrich less abundant proteins.
  - Detect sequence variants: almost all proteomic analyses ignore variants because MS identifies peptide by mass.
    - Solution (not effectively achieved yet): algorithms and cloud computing.
  - Huge dynamic range of cellular proteins: structural protein are more easily detected than regulatory proteins such as TF.
  - Membrane protein detection: the problem lies in solubilizing, separating, and digesting membrane proteins. They are hydrophobic meaning they often lack charged lysine (K) and arginine (R) residues. Many membrane proteins are also often of low abundance.
- **In terms of statistical analysis**
  - In a proteomics experiments, the *power* to detect real changes can be low for technical reasons:
    - Ratio compression in TMT experiments (co-fragmentation reduces the differences)
    - Few replicates due to high cost and instrument availability
  - So 5% FDR may fail to detect any true positives even they exist!

Phosphoproteomics

Proteomics is able to provide genome-scale information on PTMs including phosphorylation. This is not easily achievable with any other method. AP-MS (PPI analysis) and PTM proteomics are separate studies in the area of proteomics, but a parallel analysis of expression levels still helps to normalize the modification signals and PPI levels. Phosphoproteomics is a PART of PTMs. Most common PTMs include phosphorylation, acetylation, methylation, glycosylation, and ubiquitination. Modifications add additional peptide mass. Because PTMs are rare and non-stoichiometric, their analysis typically requires an enrichment step. For phosphopeptides, this is typically done with immobilized metal ion chromatography or TiO2 based; for most other PTMs, antibodies are used to bind. PTMs are rare, but can occur in billions of sites

Phosphoproteomics

PTMs

Note: true genome-wide coverage of PTMs is not feasible.

Dynamics

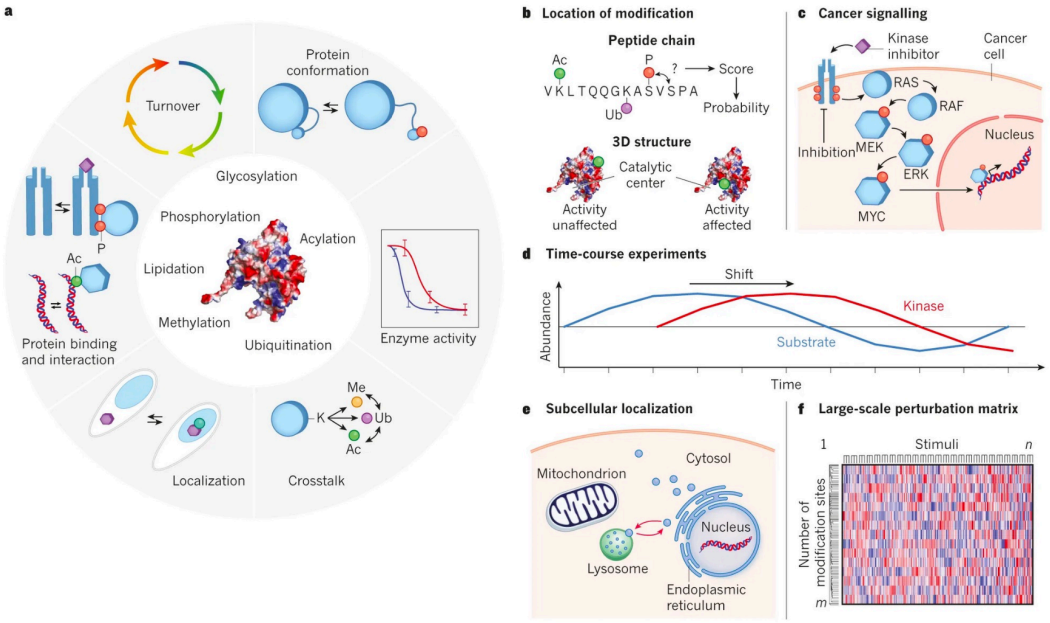
PTM levels typically change much faster than protein levels. For example, some kinases and phosphatases can act in seconds or minutes. This may raise a problem for sample preparation, as one needs to preserve the modifications as well as possible. This problem is commonly minimized using inhibitors for specific PTM (e.g., phosphatase inhibitors).

In summary, PTM data may be much noisier than protein level data.

Applications of PTM analysis

Studying PTMs by proteomics helps us to learn:

1. Protein turnover rate
2. Protein conformation change
3. Enzyme activity
4. Protein crosstalk
5. Protein localization
6. Protein binding and interactions
7. ...



Overview

Identification workflow

Generated data: position and probability

- *MaxQuant* output files: `proteinGroups.txt` and `Phospho (STY)Sites.txt`
- The phosphosites (paired with each protein) give you information about the positions of detected phosphorylation sites
- These are probabilistic identifications.
  - A peptide may have several potential phosphorylation sites. You will only get a probability that the site will be modified.
- Note: the phosphosite occupancy (stoichiometry) is rarely 1. In other words, the site is not always phosphorylated in all protein molecules)
  - 0.2% to 0.1% or 100% to 50%. They represent two *fundamentally different* regulatory strategies.
- Since protein levels also vary, it is best to normalize the level of phosphopeptides with the total protein expression level.

Interpretation

After knowing where the phosphomodifications are, we also ask the question whether they are biologically relevant and what their biological functions are. Use ML methods!

Known-sites searching

Another way to interpret the data is to search them online. One source of known phosphosites is PhosphoSitePlus

Identify kinases for phosphorylations

One of the goals in phosphoproteomics is to identify which kinase(s) are responsible for the phosphorylation. Only 5-20% of the phosphoproteome has been linked to identified kinases.

Human protein reference database provides some of the known phosphorylation motifs (http://hprd.org/PhosphoMotif_finder).
There are motif enrichment analysis programs to search for these (e.g, http://motif-x.med.harvard.edu/). You can also search, which kinases are most likely to phosphorylate your protein sequence (http://networkin.info/index.shtml).

Examine quality

We are also curious about whether the reported phosphosite identifications are of high quality.
The low true positive percentage suggests that the accumulation of phosphosite identifications from multiple independent searches might be strongly enriched for false positives.

This is once again a demonstration that you shouldn't consider everything as true what you find in the literature and databases.

Challenges

Although studying phosphorylation sites is beneficial, it may also add noise to the data as sample preparation increases variance.

Protein-protein interaction

Concept

You pull down your protein of interest (*bait*, can be endogenous or a artificial construct with affinity tag) from the cell or tissue lysate and identify interacting proteins (*prey*) by MS. The later one is called AP-MS (affinity purification-mass spectrometry). If you do this for multiple proteins, you can start building a PPI map.

Tandem affinity purification

- Purpose: to solve the problem that many protein co-purified without actually binding to the protein of interest (bait).
- Approach: add a second affinity tag and go through two rounds of purification.
- Challenge: some contaminating proteins still remain.

BioID

- Purpose: to solve the problem that some weak (but still factual) interactions may not survive protein purification conditions.
- What is it: a *in vivo* labeling method
- Principles:
  - Protein of interest is tagged with a mutant form of biotin ligase (BirA*), which labels other proteins within ~10 nm radius
  - These proteins can be enriched using biotin-streptavidin affinity purification
- Challenge: this method labels anything in the proximity of the bait protein, not just those actually physically interacting

Identify *genuine* PPIs

- **SAINT:** Significance Analysis of INTeractome
  - Principle: use a probability-based model use label-free quantification to derive the probability for an interaction to be true
- **CRAPome:** Contaminant Repository for Affinity Purification
  - Principle: uses negative controls from various groups to collect information about proteins which most frequently contaminate MS data
  - One of the most frequent contaminant is keratin. Because they are everywhere and highly expressed. In addition, they are easily introduced during sample handling.
- **Basic practices:**
  - With careful quantification, it is possible to estimate interaction stoichiometry (the molar ratio of prey protein and the bait protein) AND the relative cellular abundances of the proteins.
  - Therefore, stable protein complexes can be identified from weak or nonspecific interactions.
  - Ideally, bait proteins should be expressed at endogenous control or at similar expression levels.

- Otherwise, baits may ARTIFICIALLY interact with proteins that are not true biological partners, resulting in false positives.

Online tools and databases

A wide reference of tools for multi-omics.

Online tools & databases

Resources

Reasons to use it	**Web link**
Is this gene expressed in this organ/tissue?	proteinatlas.org
Proteomics	uniprot.org
The Rednote of biology	biology.stackexchange.com
The Wechat subscription account posts of biology	labworm.com
RNA-seq pipelines (workflow, not code)	rna-seqblog.com
Datasets integration intersection	omicsdi.org
The Cancer Genome Atlas (TCGA)	cancer.gov
Papers related to cancers	cell.com
Cancer genomics	cbioportal.org
A web-based platform for data-intensive, bioinformatics-dependent research	usegalaxy.org

Heuristic algorithm

An heuristic algorithm yields reasonable (approximate/trade-off) results, even if it is not probably optimal or lacks even a performance guarantee. It balances effect with solution speed.

Reproducibility

Versions, releases, and day of downloading must be specified to ensure reproducibility.

Benchmarking

What factors to consider when performing a benchmark

- Computational efficiency (speed, memory usage)
- Ease of use and deployment
- Reproducibility of results
- Flexibility and adaptability to different datasets
- Scalability for large datasets

Issues with gold standards

Systematic benchmarking based on a gold standard should in principle help to evaluate computational tools. The gold standard can be prepared by computer simulation or curated by human experts.
Potential issues with this gold standard include:

- Simulated data may not capture real-world biases and errors
- Real data may lack complete ground truth, especially for complex genomic regions
- Human-curated standards can be subjective and error-prone, with possible cherry-picking bias
- Difficulty in reflecting all biological variability and edge cases

Importance of analysis software vs. experimentation in scRNA-seq

- Both are critical, but normalization and library preparation have the largest impact on results.
- Normalization is crucial because it corrects for sequencing depth and technical variability, significantly affecting the DEG analysis.

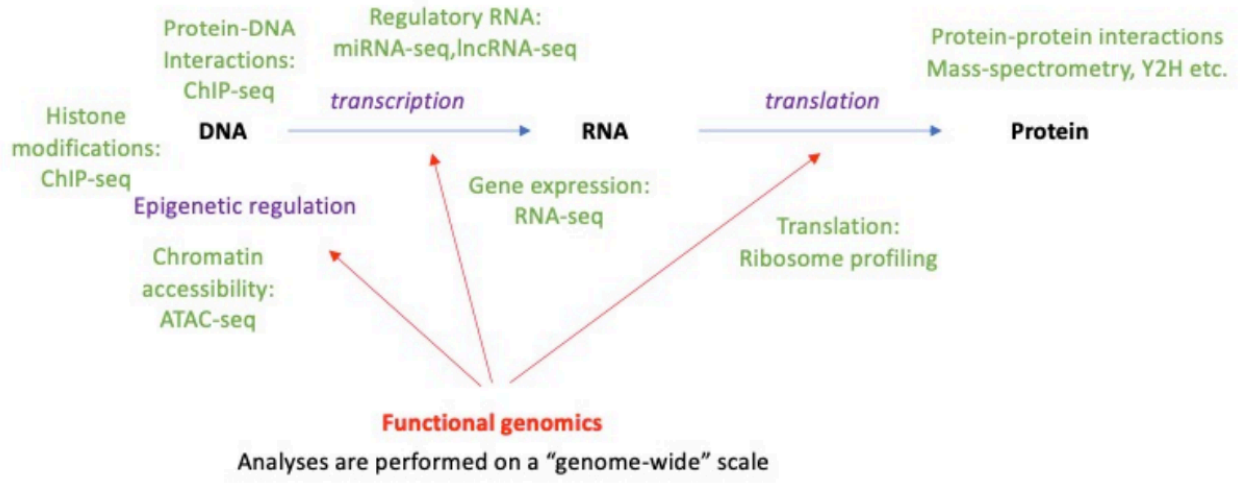
When to stop using an old tool?

- When a superior AND validated successor exists (e.g., HISAT2 replacing TopHat)
- When community support and updates cease (CellChat v2)
- When the older tool obviously introduces inefficiencies or inaccuracies in analysis

Functional genomics

Functional genomics is the very interface between dry-lab technologies and wet-lab considerations. Functional genomics is the study of how the genome, transcripts/genes, proteins, and metabolites work together to produce a particular phenotype, that is, functional genomics is the application of genome information. It studies the gene expression, gene functions, and regulation of thousands of genes at once. In other words, functional genomics focuses on the dynamic expression of gene products in a specific context. Functional genomics is highly technology dependent.

Functional genomics



Fundamental biological questions answered with the advancement of functional genomics

- Q: How many genes are essential for the viability of human cells?
  - A: 2,000 (under standard conditions.) Since there are ~20,000 protein-coding genes. There are only 10% truly essentials for life.
- Q: Are essential genes in human cells also essential in other organisms?
- A: Essential genes are more evolutionarily conserved. Many essential genes in human cells are also essential in yeast.
- Q: What are the functions of essential genes in humans?
  - A: Enriched in translation, transcription, and DNA replication, but not signalling.
- Q: What are the differences in the set of essential genes across different types of human cells and genotypes?
  - A: Many. But a core set of common essential genes (~700 genes) exist.

Applications

- Functional identification for all genes in all organisms in all conditions
- Understanding of mechanisms of antibiotic resistance
- Identification of drug targets, pahrmacogenomics and more
- Move onto more specific ones…
  - Genome and exome sequencing have yielded extensive catalogs of genetic variation. How does the presence of mutations upstream of promoters affect the ability of the sequence to be transcribed through *in vitro* transcription?
    - Solution: make library
      - Covert to RNA, RT-PCR and sequence (Is it expressed successfully?)
      - DNA, PCR and sequence (Successfully constructed? Already in DNA?)
      - See what variants are significantly different between these two approaches.
  - What are the phenotypic effects of different cancer molecular targets?
    - Experimental designs can vary (cell lines, mouse models) to study different phenotypes.

Model organisms

Why use model organisms?

- Better model than cell lines
- Fewer ethical problems
- Experimentally handlable, e.g. quick generation time (Drosophila), transparent (zebrafish), simple physiology (yeast)

Note

Phenotypes are often more diverse than 'on-off'. They have "severity", a "degree".

Forward & reverse genetics

- Forward genetics
  - Traditional genetics
  - Phenotype in hand, linked genotype unclear.

  - Start from phenotype and looking for what genotype is responsible for that phenotype.
- Reverse genetics
  - = Functional analysis
  - Genotype in hand, linked phenotype unclear.
  - Start from genotype and looking for what phenotype correspond to this genotype.

Functional genomics aims to prove the casual relationship between genotype and phenotype.

Reverse genetics: systematic phenotypic screens using pooled populations

Feature: quick, cheap and less labor-intensive

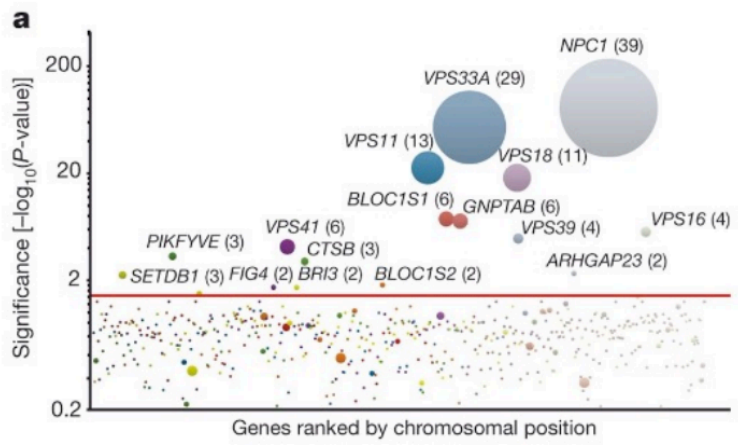
- Systematic deletion of each gene in yeast.
- Knocked-out genes are replaced with a KanMX cassette with unique barcodes (~20nt long) for identification, which allows easy tracking of each depletions.
- These mutant yeast strains are pooled and grown together under a specific condition.
- Microarray technology identifies which mutants thrive or fail under specific conditions.

Reverse genetics: RNAi & cell-based screening

In 2006, RNAi emerges as a tool for gene silencing. In addition, sequencing technology, *Illumina Genome Analyzer*, was introduced, capable of 1 Gb per run.

RNAi technology enables cell-based screening to study genes' regulatory effects. Coupled with arrayed screening, each gene's function can be tested individually. But this is not necessary. You can just test the function of one or two genes. Arrayed screening typically uses microwell plates and involved 3 replicates for each gene. It allows different phenotypes to be tested. Microwell plates allows high-throughput.

An example of result is like this, able to achieve the data presentation of thousands of genes in one single figure.



Reverse genetics: advaned toolbox

Cutting-edge tools - Y2H: used to examine PPI or Protein-DNA interactions - CRISPR, RNAi - RNAseq, proteomics and metabolomics

Chemical genomics

- Purpose: how small chemical molecules interact with cells. Not pure functional genomics in the strictest sense because it uses principles of genomic screens but with drug libraries instead of gene manipulation.
- Feature: phenotypic drug screens do not reveal drug targets, but can be combined with sgRNA/RNAi tools
- Role:
  - Functional genomics screens may reveal *mechanism* of action
  - Chemical genomics identify function for an uncharacterised genes
- Approaches
  - Forward chemical genetics: starts with phenotype to identify molecular targets (receptor, etc.)
  - Reverse chemical genetics: starts with a molecular target (receptor, etc.) chemical modifications to study its phenotypic impact.

Data analysis options

- Volcano plots
- Enrichment
- Network analysis
- The use of *Venn diagram*
- Binding motifs
- Correlation between samples

- Heatmaps
- Clustering…

Shortcomings

- Functional redundancy of genes prevents identification of observable phenotypes.
  - Explanation: many genes have functional redundancy, meaning that multiple genes can perform similar or overlapping biological roles. Because of this, when you knock out or silence one gene, another gene might compensate for its function, leading to no observable phenotype.
- Genetic interactions module phenotype
  - Explanation: gene interactions can significantly modulate phenotypes. Genes rarely act alone; they interact with other genes, proteins, and pathways. – How representative is the data from a single cell line or animal?
  - Explanation: when experiments are done in a single cell line (like HeLa or MCF7) or a single animal model (like C57BL/6 mice), the results may not be representative of other biological systems.
- Phenotype may be heavily dependent on environmental conditions, e.g. cell culture medium, temperature
- Information overload and multiple-testing problems due to large-scale experimentation
- Independent validation required
- Your results are only as good as your measurements!
  - Explanation: if your measurements are inaccurate, contaminated, or technically flawed, all subsequent analyses and conclusions will be unreliable.

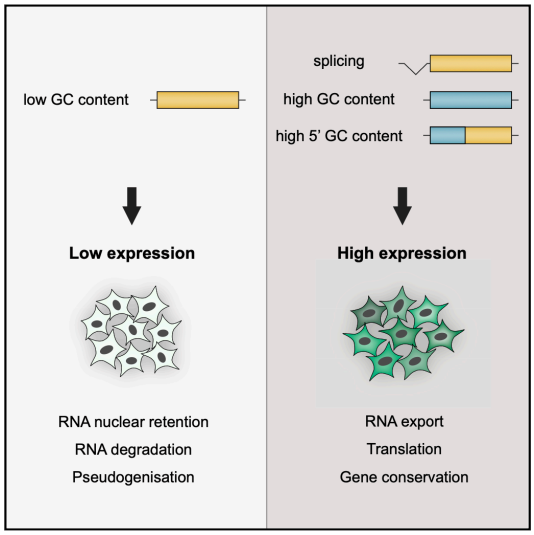
Multi-omics data integration

Describe a published example of a multiomics project that integrates multiple datasets. Understand why researchers want to integrate the genomic and proteomic data. Identify the difficulties and compromises required in this type of research.

Multi-omics data integration

An example of multi-omics data integration: *Codon Usage and Splicing Jointly Influence mRNA Localization*

Finding summary



- GC4 (the 3_{rd} position of codon) content usually has a positive effect on gene expression in unspliced genes, including:
  - transcription initiation (GRO-cap, transcription start sites),
  - cytoplasmic stability (exosome mutant data),
  - total RNA (RNA-seq),
  - cytoplasmic enrichment (RNA-seq from cell fractionation),
  - translation rate (ribosome profiling vs RNA-seq),
  - protein (SILAC)
- Not so much for protein stability (SILAC/translation rate)

Multi-omics summary

To support this conclusion, the information we aim to get are as follows:

Genomic data	Transcriptomic data	Proteomic data
1. GC4 content	2. Transcription initiation	5. Translation
	3. RNA stability	6. Protein expression
	4. RNA expression/localisation	7. Protein stability

GC4 content

Purpose

The goal is to annotate the genome to understand the distribution of GC content at specific locations called fourfold degenerate sites.

Fourfold degenerate sites are codon positions where any nucleotide change does not alter the amino acid (e.g., GG<u>A</u>, GG<u>C</u>, GG<u>G</u>, GG<u>T</u> all code for Glycine).

Steps

1. Extract transcript sequence.
2. Identify four fold degenerate sites.
3. Score the GC-content at these sites.

GENCODE

Annotations are from GENCODE project (from gene to transcripts to protein correspondence information). It is used because it fitted with other ongoing projects in the lab. It is used throughout because all the sites need to be directly comparable, especially with precision.

Transcription initiation

Purpose

To identify where transcription starts in the genome and quantify the nascent RNA at those sites.

Steps

1. Use *Global Run-On sequencing* (GRO-cap, a little like ChIP-seq), which tags the 5' ends of nascent RNA molecules. This captures where transcription is initiated.
2. The reads are then mapped using `.bedGraph` file from GEO.
3. Promoters are defined as regions around GENCODE Transcription Start Sites (TSS), specifically -300 to +100 bp. This window size is chosen because of standard definitions in the literature for promoter regions.
4. BedTools is used to perform genomic intersections to count the reads for each transcript.

RNA localisation and expression

Purpose

To understand how RNA is distributed between the nucleus and cytoplasm and at what level it is expressed.

Steps

1. RNA-seq data is collected from ENCODE, also from additional in-house experiments. The reason for combining both ENCODE existed RNA-seq data and in-house experiments is that although ENCODE provides high-quality annotations but was missing some *HeLa cell* comparisons, which is why in-house data is used.
2. These reads are aligned using Kallisto, which is efficient for pseudo-alignment.
3. Data is quantified in TPM to standardize across samples.

RNA localisation

Purpose

To measure how much of the RNA is in the cytoplasm versus the nucleus.

Formula

$$\text{Relative Cytoplasmic Concentration (RCC)} = \frac{\text{Cytoplasmic TPM}}{\text{Nuclear TPM} + \text{Cytoplasmic TPM}}$$

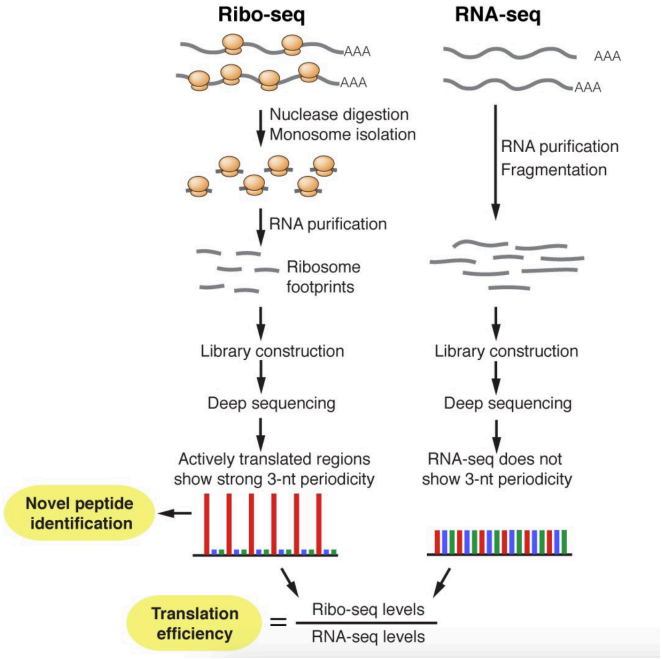
- This value ranges from 0 to 1, representing the fraction of total RNA that is in the cytoplasm.
- It avoids simple ratios that could skew distributions, ensuring a more meaningful interpretation.
- This was the preferred method by experimental collaborators for its clarity and interpretability.

Translation process

Purpose

To understand the translation efficiency of mRNA transcripts by analyzing the association of mRNA (RNA-seq) with ribosomes (Ribo-seq).

Steps



1. Captures ribosome footprints, which indicate regions of mRNA currently being translated. This is done by isolating ribosomes and sequencing the encapsulated mRNA fragments. Actively translated regions show a 3-nt periodicity, reflecting the triplet nature of codon translation.
2. Provides a broader view of all mRNA abundance by sequencing all RNA fragments, not just those protected by ribosomes. It does not show the 3-nt periodicity, as it includes non-translated regions as well.

But Ribo-seq data from GEO is simply counts per gene. However, both Ribo-seq and RNA-seq only looks at transcripts. So it remains a question whereas to adopt gene-level or transcript-level analysis because if we attempted to assign each gene count to all GENCODE transcripts, some measurements will be inflated.

Protein expression

Steps

For each gene…

1. There are three independent counts, mapped to gene names or UniProt IDs.
2. Mean expression values are calculated across triplicates.
3. However, some GENCODE transcripts map to multiple UniProt genes, causing redundancy or conflicts in annotation.
4. Just remove conflicting mappings to simplify analysis.

Protein stability

Purpose

To measure the stability of proteins relative to their mRNA levels.

Formula

$$\text{Protein Stability} = \frac{\text{Protein Expression (mass spec)}}{\text{Whole cell RNA-seq}}$$

- A higher ratio suggests that the protein remains stable longer relative to its mRNA levels. - A lower ratio could indicate rapid degradation or inefficient translation.

Critical analysis

This might not be justified because this formula assumes a. RNA-seq *accurately* represents mRNA abundance and b. Mass spectrometry is a *direct* indicator of steady-state protein levels.

Normalization

Purpose

To normalize expression data for fair comparison across samples.

Steps

1. Pseudocount added to avoid zeros and allow for the following log2 transformation.
2. All values are transformed using log2 for easier interpretation of fold changes.
3. Transcripts with an expression value of 0 are removed to prevent infinite values in log transformation.

Multi-omics integration issues

1. Multi-omics data often comes from different sources: patient samples, different diseases, or even distinct species. It makes direct comparison challenging because variations may be due to sample differences rather than biological effects.
2. Different omics layers (e.g., genomics vs. proteomics) measure distinct aspects of biology. These layers are not always directly translatable, causing issues when trying to integrate them for analysis.
3. Researchers often make practical and subjective decisions during processing, such as "which transcripts to include?", "which threshold to use?" These decisions can introduce bias and affect the comparability of datasets.
4. When combining different datasets, annotations may not always match perfectly. A transcript may be detected in RNA-seq but not show up in proteomics, leading to gaps. So important biological signals could be missed, or data interpretation may be skewed.
5. Different omics platforms produce data in various formats (e.g., `.bed`, `.bam`, `.fastq`, `.csv`). So researchers can spend an excessive amount of time simply reformatting and harmonizing datasets instead of analyzing them.
6. Even in this single example, the analysis referenced multiple lectures and external knowledge sources. This complexity demands more expertise to correctly interpret results.
7. Data alone is not enough; experimental validation and understanding are crucial. Effective collaboration with experimental biologists helps validate findings and align analysis with biological reality.
8. Every multi-omics project is different, making it impossible to have a "one-size-fits-all" workflow. Researchers must adapt their methods for each new dataset and question, increasing complexity.

AI in omic studies

AI in omic studies

Geneformer

Overview

- What is Geneformer: a context-aware, attention-based deep learning model, grounded in Transformer architecture, pretrained on a large-scale scRNA-seq data to enable context-specific predictions in network biology.
- Approach: it employs a perturbation function that simulates gene expression changes to predict alterations in cellular states.
- Functions: mapping gene networks to learn the connections between genes

Main steps

- Enhanced data-preprocessing
  - Supports single-cell *multi-omics* data
  - Improves inter-class separability of cell types
- Upgraded model structure
  - Groups genes into functional modules, enhancing interpretability and pathway-level interactions.
  - A new module designed to support multi-omics data fusion
- Advanced pretraining strategies
  - Causal masked learning: simulates gene regulatory networks to predict the effects of perturbations more accurately.
  - Adversarial training: utilizes adversarial techniques to boost the model's robustness against noisy biological data.
- Fine-tuning
  - Zero-shot: enables natural language instructions to be mapped to biological predictions directly.
- New application scenarios
  - Provide spatial omic information
- Deployment & optimization
  - Lightweight version
  - Cloud platform version
- Unified multi-omics modeling
  - Process various omics data types in a unified analytical pipeline.
- …

Highlights of V2

- Dynamic rank encoding (data balancing): address the oversight of rare cell types in scRNA-seq data
- Hierarchical sparse attention (efficiency): achieve computational efficiency while preserving biological relevance
- Causal masking (biological logic): mimic biological causality in gene regulation
- Adversarial training (data synthesis): enhance realism of generated single cell data
- Perturbation impact score: quantifies global effects of gene perturbation on cell states
- Spatial interaction modelling: identifies gene interactions networks in spatially adjacent cells

Transformer

- Embedding: It turns words into numbers so it can understand them, just like you turn apples into slices to eat them.
- Attention: It focuses on the most important parts, like when you pay attention to your teacher's instructions.
- MLPs (Multi-Layer Perceptrons): It learns deeper and more complicated things, like solving puzzles faster each time.
- Unembedding: It takes all its learning and makes it back into words or pictures that you can understand.

Algorithms

- Self-supervised large-scale pretraining
- Multi-task fine-tuning
  - Zero-shot learning
  - Limited task-specific data fine-tuning

scGPT vs. Geneformer

Aspect	ScGPT	Geneformer
Data processing and tokenization	Retains relative expression magnitudes through binning. Multimodal design suits cross-omics analysis.	Emphasizes gene ranking. Excels in transcriptome-specific feature extraction.
Model architecture and pre-training objectives	Allows direct simulation of gene perturbations (e.g., knockout effects).	Focuses on feature-based tasks like cell state classification.

Aspect	ScGPT	Geneformer
Applications and transfer learning	Excels in exploratory tasks such as novel drug target prediction.	Offers deeper insights into disease mechanisms.
Biological interpretability	Visualizes attention weights to reveal gene regulatory networks (e.g., transcription factor-target interactions).	Infers functional gene relationships via embedding similarities (e.g., clustering dosage-sensitive genes).
Strengths	Multimodal integration, generative prediction, zero-shot adaptability.	High-dimensional feature extraction, disease target validation.
Ideal use cases	Cross-omics analysis, perturbation simulation, novel hypothesis generation.	Cell type classification, gene network inference, mechanistic studies.
Design philosophy	Adopts a language-model-inspired approach for generality.	Prioritizes biological relevance.
Choice dependencies	Best for multi-omics and generation-based tasks.	Best for single-omics and feature-based analysis.