ML 4 Genomics: Notes and Self-check Questions

ETH Zürich - Autumn 2024

Week 1 - Introduction to Genomics

Glossary: Basic Biology and Genomics

- **Cell**. A cell is the basic structural, functional, and biological unit of all living organisms. Cells contain the organism's DNA housed in the nucleus (in eukaryotic cells) or freely in the cytoplasm (in prokaryotic cells).
- **Protein**. Proteins are large, complex molecules made up of amino acids. They perform a vast array of functions within organisms, including catalyzing metabolic reactions, providing structural support, and responding to stimuli.
- **DNA (Deoxyribonucleic Acid)**. DNA is a molecule that contains the genetic code of living organisms. It is composed of two strands forming a double helix and consists of nucleotides made of a sugar, a phosphate group, and a nitrogenous base: adenine (A), thymine (T), cytosine (C), guanine (G). It usually exists as a double-stranded molecule. It has a direction: we distinguish between the 5' and 3' ends, which refer to the orientation of the carbon atoms in the sugar molecule of the DNA backbone. DNA is always read from the 5' to the 3' end. So that reading the reverse complement is equivalent to reading the reverse of the DNA sequence.
- **Base Pair (bp)**. A base pair is a pair of complementary nucleotide bases in DNA. Adenine (A) pairs with thymine (T), and cytosine (C) pairs with guanine (G). Base pairs are the building blocks of the DNA double helix structure.
- **Genome**. A genome is the complete set of DNA, including all of its genes, in an organism. It encompasses all the genetic material present in the organism's cells. The human genome, for example, consists of approximately 3.2×10^9 base pairs.
- **Gene**. A gene is a segment of DNA that contains the instructions to make a specific protein or set of proteins. Genes are the basic units of heredity and are passed from parents to offspring.
- **Chromosome**. A chromosome is a long, thread-like structure made of protein and a single molecule of DNA. It contains many genes and regulatory elements. Humans typically have 23 pairs of chromosomes.
- **Genomics**. Genomics is the study of the structure, function, evolution, mapping, and editing of genomes. It involves the comprehensive analysis of the entire genome, including its interactions and effects.
- **Genetics**. Genetics is the study of genes, heredity, and the variation of organisms. It examines how traits are passed from one generation to the next and how genes contribute to an organism's development and function.

Summary of Key Differences

- **DNA vs. Gene**: DNA is the molecule, while a gene is a specific section of DNA that codes for a protein.
- **Genome vs. DNA**: The genome includes all of an organism's DNA, both coding and non-coding regions.
- **Genomics vs. Genetics**: Genomics is the study of the entire genome, while genetics focuses more on individual genes and inheritance.

- **Chromosome vs. Gene**: Chromosomes are structures made of DNA that carry multiple genes.
- **Protein vs. Gene**: Genes provide the instructions for making proteins; proteins are the end products that perform cellular functions.

Additional Basic Questions

1. **Is the genome the same for all individuals?** No, the genome varies between individuals due to small differences in DNA sequences, known as genetic variation. Most of these variations are minor and do not affect the organism's functioning, but some variations are responsible for differences in traits like appearance, behavior, or susceptibility to diseases. Humans share about 99.9% of their genome, with the remaining 0.1% accounting for individual differences.
2. **How many nucleotides per gene?** The length of genes varies widely. Some genes can be as short as a few hundred nucleotides, while others can span millions of nucleotides. On average, a human gene contains about 27,000 base pairs of nucleotides, though many of these include non-coding regions (introns) that do not directly code for proteins.
3. **How many genes in the human body?** The number of genes in the human genome is estimated to be around 20,000-25,000.

Central Dogma of molecular biology

There are 3 major classes of cell biopolymers[1]: **DNA** (A,T,C,G), **RNA** (A,U,C,G), and **proteins** (20 amino acids).

The central dogma of molecular biology explains the flow of genetic information within a biological system, establishing the framework for how genes encoded in DNA are expressed as proteins, which perform critical cellular functions. This flow follows two main steps: transcription and translation.

- **Transcription**: Transcription is the process where the genetic information encoded in a segment of DNA is copied into messenger RNA (mRNA). This occurs in the nucleus of eukaryotic cells and the cytoplasm of prokaryotic cells. The enzyme RNA polymerase binds to a specific region of the DNA called the promoter, unwinding the DNA double helix. RNA polymerase then reads the DNA template strand, synthesizing a complementary RNA sequence. The resulting mRNA carries the genetic code out of the nucleus to be used in protein synthesis.
- **Translation**: Translation is the process where the mRNA is decoded by the ribosome to produce a specific protein. This occurs in the cytoplasm, where the mRNA sequence is read in codons (three-nucleotide sequences), each specifying a particular amino acid. Transfer RNA (tRNA) molecules bring the appropriate amino acids to the ribosome, which links them together in the order specified by the mRNA, forming a polypeptide chain that folds into a functional protein.

$$\text{DNA} \xrightarrow{\text{Transcription}} \text{RNA} \xrightarrow{\text{Translation}} \text{Protein}$$

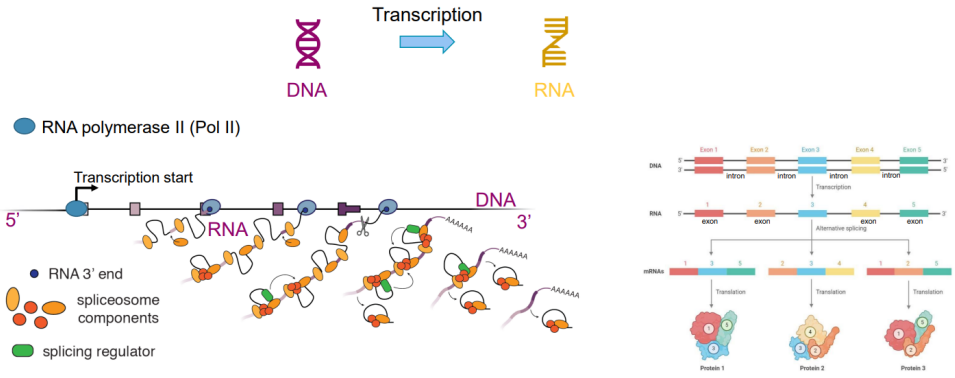

Figure 1: Transcription Phase

[1] large, naturally occurring molecules such as proteins, nucleic acids, and polysaccharides that form the structural and functional components of living cells.

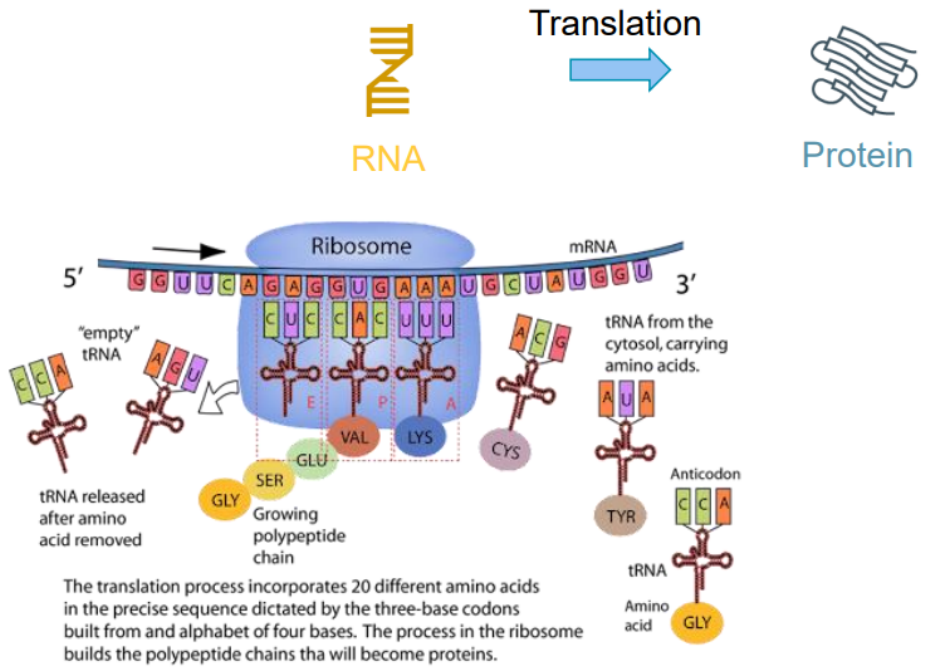

Figure 2: Translation Phase

Key Concepts

- **DNA** (deoxyribonucleic acid) is the hereditary material in cells, and it contains the instructions for building proteins.
- **RNA** (ribonucleic acid) serves as the intermediary molecule that carries the genetic code from DNA to the ribosome.
- **Proteins** are the functional products of genes, responsible for performing a wide variety of roles in the cell, including catalysis (as enzymes), structure, and signaling.
- **Template Strand** is the DNA strand that is read by RNA polymerase during transcription. The RNA produced is complementary to this strand.
- **Non-Template Strand** (also called the coding strand) is the DNA strand that is not transcribed, but its sequence is the same as the mRNA (except thymine (T) in DNA is replaced by uracil (U) in RNA).
- **Messenger RNA (mRNA)** is the RNA molecule transcribed from DNA that serves as a template for protein synthesis during translation. Its sequence is complementary to the DNA template strand and is later translated into a protein by ribosomes.
- **RNA Polymerase** is the enzyme responsible for synthesizing RNA from a DNA template during transcription. It binds to the promoter region of a gene and catalyzes the formation of an RNA strand by matching RNA nucleotides to the complementary DNA bases.
- **TSS (Transcription Start Site)** is the specific location on the DNA where transcription begins, marking the point at which RNA polymerase starts synthesizing RNA.
- **Untranslated Regions (UTR)** are regions of mRNA that are not translated into protein. The **5' UTR** is located upstream of the coding sequence, and the **3' UTR** is located downstream. These regions play roles in mRNA stability, localization, and regulation of translation.
- **Introns** are the non-coding regions of a gene that are removed during RNA splicing. They do not code for proteins but may play roles in gene regulation.
- **Exons** are the coding regions of a gene that remain in the mRNA after splicing and are translated into proteins. They include both the protein-coding sequence and untranslated regions (UTRs).

Careful: not all exons are translated into proteins, in fact, in humans only about 30% of the exons are coded into proteins, and the rest aid regulating the gene expression.

- **RNA Splicing** is the process of removing introns (non-coding regions) from the pre-mRNA transcript and joining the exons (coding regions) together to form a mature mRNA molecule ready for translation.
- **Alternative Splicing** is a process by which different combinations of exons are joined together, allowing a single gene to produce multiple protein variants. This increases the diversity of proteins that can be encoded by the genome. In other words, from the same gene, different mature mRNAs can be produced, leading to different proteins.
- **Promoter** is a DNA sequence located upstream of the transcription start site (TSS) where RNA polymerase binds to initiate transcription. It helps regulate gene expression by controlling when and how much a gene is transcribed. It determines the direction of transcription.
- **Enhancer** is a DNA sequence that can increase the transcription of a gene. Enhancers function by binding transcription factors and are often located far from the gene they regulate, influencing gene expression over long distances. It boosts the rate of transcription.
- **Codon** is a sequence of three nucleotides in mRNA that corresponds to a specific amino acid or a stop signal during protein synthesis. For example, the codon `AUG` codes for the amino acid methionine, which is often the start signal for translation.
- **Anticodon** is a sequence of three nucleotides in a tRNA molecule that is complementary to a codon in mRNA. It allows the tRNA to bring the correct amino acid to the ribosome during protein synthesis.
- **Mature mRNA** is the fully processed mRNA transcript that has undergone splicing, capping, and polyadenylation, and is ready to be exported from the nucleus to the cytoplasm for translation.
  Capping: The addition of a modified guanine nucleotide to the 5' end of the pre-mRNA to protect it from degradation and assist in translation.
  Polyadenylation: The addition of a long tail of adenine nucleotides (poly-A tail) to the 3' end of the pre-mRNA, which enhances mRNA stability and aids in its export from the nucleus.
- **Transcription Factors** are proteins that bind to specific DNA sequences near genes (such as promoters or enhancers) and help regulate the transcription of those genes, either activating or repressing gene expression.

More Insights into the Transcription Process

For transcription to be initiated, the polymerase must first gain access to the promoter region at the beginning of the gene. This can be impaired by the condensation of chromatin, the complex of DNA and proteins that makes up chromosomes. The histones are proteins that help package DNA into a compact structure, forming the core around which DNA is wrapped. The nucleosome is the basic unit of DNA packaging, consisting of a segment of DNA wound around a core of histone proteins.

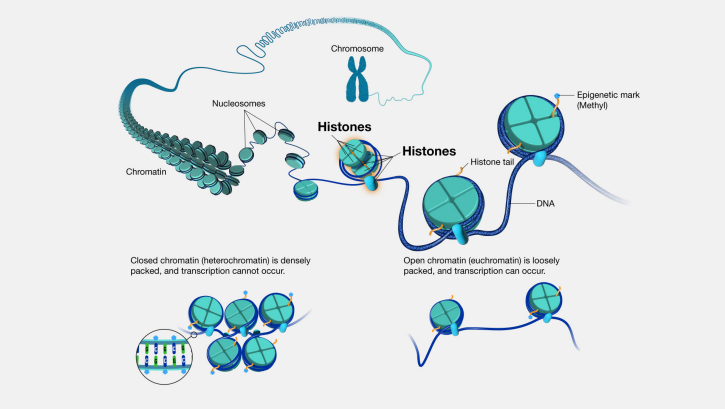

Figure 3: Histone Modification and Gene Regulation

Role of Transcription Factors, Histone Modification and Nucleosome Displacement

- **Transcription factors** are proteins that bind to specific DNA sequences, such as promoters and enhancers, to regulate the transcription of genes. Their role is crucial in enabling RNA polymerase to access the promoter region, especially when DNA is tightly packaged in nucleosomes. Transcription factors can recruit other proteins, such as chromatin remodelers, which help loosen or reposition nucleosomes, thereby exposing the DNA for transcription.
- **Nucleosomes** can either block or facilitate transcription. If nucleosomes are tightly packed and densely distributed, the chromatin is in a "closed" state, making the gene repressed or inactive. In contrast, if nucleosomes are loosely packed or displaced, the chromatin is in an "open" state, allowing transcription machinery to access the DNA, thus making the gene active.
- **Active genes** are typically found in regions of loosely packed chromatin, known as *euchromatin*, where nucleosomes are spaced farther apart, allowing RNA polymerase and transcription factors to access the DNA easily. In contrast, **repressed genes** are found in *heterochromatin*, where nucleosomes are tightly packed, limiting access to the DNA and preventing transcription. The degree of nucleosome compaction around a gene plays a key role in determining its expression. The denser the nucleosome arrangement, the more likely a gene is repressed, while a less dense arrangement generally correlates with gene activation.
- **Histone modification** involves chemical changes to histone proteins, such as the addition of acetyl, methyl, or phosphate groups. These modifications affect how tightly or loosely the DNA is wrapped around histones, influencing gene expression. For example, histone acetylation usually results in a relaxed chromatin structure (euchromatin) and gene activation, as it reduces the positive charge on histones, weakening their interaction with negatively charged DNA. Conversely, histone methylation can either activate or repress gene transcription, depending on the specific site of modification. Histone modifications serve as signals to the transcriptional machinery, indicating whether the chromatin is in an active or repressed state. These modifications can also recruit chromatin remodelers that further alter the nucleosome structure and dynamics.

DNA Looping and Gene Regulation

How can an enhancer influence transcription of gene whose transcription start site (TSS) is located about 27,000 bp away?

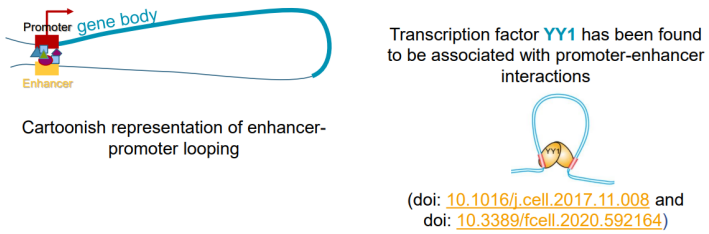

Figure 4: DNA Looping and Enhancer-Promoter Interaction

- **DNA looping** is a process where distant regions of the genome are brought into close physical proximity through the looping of the DNA strand. This allows enhancer elements, which may be located far from the gene they regulate, to interact with promoters and increase the transcriptional activity of genes. Transcription factors and other regulatory proteins play a key role in stabilizing these loops, thereby facilitating efficient communication between regulatory regions of the genome and the transcriptional machinery.

Some additional key concepts:

- **Binding Motif**: specific sequence or pattern of nucleotides in DNA that is recognized and bound by a protein, such as a transcription factor. These motifs are key to regulating gene expression, as they dictate where regulatory proteins can attach to control transcription.
- **DNA Methylation**: DNA methylation is a chemical modification where a methyl group is added to cytosine bases in the DNA, typically at CpG sites. This process can repress gene expression by

altering the structure of chromatin or preventing transcription factors from binding, contributing to gene silencing.

- **CTCF (CCCTC-binding factor)**: a protein that helps organize the 3D structure of chromatin and regulates gene expression by controlling DNA looping and insulation between different parts of the genome. CTCF plays a critical role in establishing boundaries between chromatin domains and regulating enhancer-promoter interactions.
- **TDA (Topologically Associating Domains)**: TADs are regions of the genome where DNA sequences are more likely to interact with each other than with sequences outside the domain. TADs help organize the 3D structure of chromatin and play a role in regulating gene expression by bringing enhancers and promoters into close proximity.

Missing concepts:

• ***miRNA***: (micro RNA) Small non-coding RNA molecules that regulate gene expression by binding to messenger RNA (mRNA) and preventing it from being translated into protein.

• ***lncRNA***: (long non-coding RNA) Longer RNA molecules that do not code for proteins but can regulate gene expression and influence other biological processes.

• ***DNA methylation:*** A chemical modification where a methyl group is added to DNA, often serving to silence genes and regulate gene expression.

Week 2 - Working with DNA

Transcription factors (TFs) are proteins that bind to DNA to regulate transcription. They play a crucial role in gene expression by controlling when and how much a gene is transcribed. TFs are, for example, important players in the development of cancer, as mutations in TFs can lead to dysregulated gene expression and uncontrolled cell growth.

TFs bind DNA with very high specificity, recognizing short DNA sequences called **transcription factor binding sites** (TFBSs). These sites are typically 6-20 base pairs long and are often located near the genes they regulate. The specific sequence of a TFBS is known as the **binding motif** of the transcription factor.

Terminology Summary for Genomic Variation

SNP vs Mutation

- **SNP (Single Nucleotide Polymorphism)**: A SNP is a variation in a single nucleotide that occurs at a specific position in the genome. They are the most common type of genetic variation in humans (occurring in more than 1% of the population) and can be used as genetic markers to track disease susceptibility, drug response, and other traits. Roughly 1 in 1,000 nucleotides in the human genome is a SNP. In other words, generally, two individuals will have 99.9% of their DNA in common, and the remaining 0.1% will differ in the form of SNPs.
- **Mutation**: Mutations are much rarer than SNPs (occurring in less than 1% of the population) and can have a more significant impact on gene function, although generally they have no effect on health or development. Mutations can be caused by errors in DNA replication, exposure to mutagens, or other factors.

Germline vs Somatic Mutations

- **Germline Mutation**: A germline mutation is a genetic change that is present in the DNA of an organism's germ cells (sperm or egg cells). These mutations can be passed on to offspring and are present in every cell of the offspring's body. Germline mutations are the cause of inherited genetic disorders.
- **Somatic Mutation**: A somatic mutation is a genetic change that occurs in a somatic cell (any cell other than a germ cell). These mutations are not passed on to offspring and are present only in the cells that arise from the mutated cell. Somatic mutations can contribute to cancer and other diseases.

Note: SNPs can be either germline or somatic mutations, depending on when they occur and in which cells they are present.

Coding vs non-coding variants

- **Coding Variants**: Coding variants are genetic changes that occur within the protein-coding regions of genes. These variants can lead to changes in the amino acid sequence of the resulting protein, potentially affecting its structure and function. Coding variants are more likely to have a direct impact on phenotype. They are the most studied type of genetic variation and least frequent (about 5% of all variants).
- **Non-coding Variants**: Non-coding variants are genetic changes that occur outside of the protein-coding regions of genes. These variants can affect gene expression, splicing, and other regulatory processes. Non-coding variants are often associated with complex traits and diseases, and ML methods are increasingly used to study their effects. Much more frequent than coding variants.

Validation of non-coding variants

Two experimental protocols are commonly used to validate the effects of non-coding variants:

- **Genome-wide association studies (GWAS)**: GWAS are used to identify genetic variants, typically SNPs, associated with specific traits or diseases. The basic idea of GWAS is to compare the genomes of two groups of individuals: one group with a specific trait (e.g., a disease), and a control

group without that trait. The entire genome is scanned for genetic differences between the groups. If a variant (SNP) is found to be statistically more frequent in the group with the trait, it is said to be "associated" with that trait.

How it works:

- Collect genetic data from individuals with and without the trait (cases and controls).
- Perform statistical tests to compare the allele frequencies[2] of each SNP between the two groups.
- Identify SNPs significantly associated with the trait.
- These associations suggest a correlation between a region of the genome and the phenotype, but not necessarily causation.

Key concept: GWAS link SNPs to specific phenotypes, which are observable traits, but do not directly explain the molecular mechanisms. Typically a p-value threshold of 5×10^{-8} is used to determine statistical significance in GWAS.

- **Expression Quantitative Trait Loci (eQTL) analysis**: eQTL analysis is used to identify genetic variants that influence gene expression levels. Unlike GWAS, which associates SNPs with a trait or disease, eQTL analysis focuses on how genetic variation affects the transcriptional activity of genes. This is particularly important for understanding non-coding variants, as many of them may not change proteins directly but rather influence the regulation of gene expression.

  How it works:

  - Collect both genetic data and gene expression data from a cohort of individuals.
  - Identify regions of the genome (eQTLs) where genetic variants (e.g., SNPs) correlate with changes in the expression levels of nearby or distant genes.
  - Use statistical methods to determine whether the presence of certain genetic variants is significantly associated with altered gene expression.
  - Variants identified in eQTL studies can give insight into how non-coding regions of the genome affect gene regulation.

  Key concept: eQTL analysis links SNPs to gene expression levels, allowing researchers to understand how genetic variation can regulate gene activity, often providing more mechanistic insight than GWAS.

Differences between GWAS and eQTL:

- **GWAS:** Associates SNPs with specific phenotypes (traits or diseases), but does not explain the underlying biological mechanisms.
- **eQTL:** Links SNPs with gene expression changes, offering insight into how non-coding variants may impact gene regulation.

Preference: The choice between GWAS and eQTL depends on the research goal:

- **GWAS:** Preferred when the goal is to identify genetic variants associated with a phenotype (e.g., disease susceptibility). It is widely used for discovering genetic risk factors.
- **eQTL:** Preferred when the focus is on understanding how genetic variation affects gene expression, which is especially useful for studying non-coding regions and regulatory mechanisms.

[2] An allele is a variant form of a gene, which can result in different phenotypic traits.

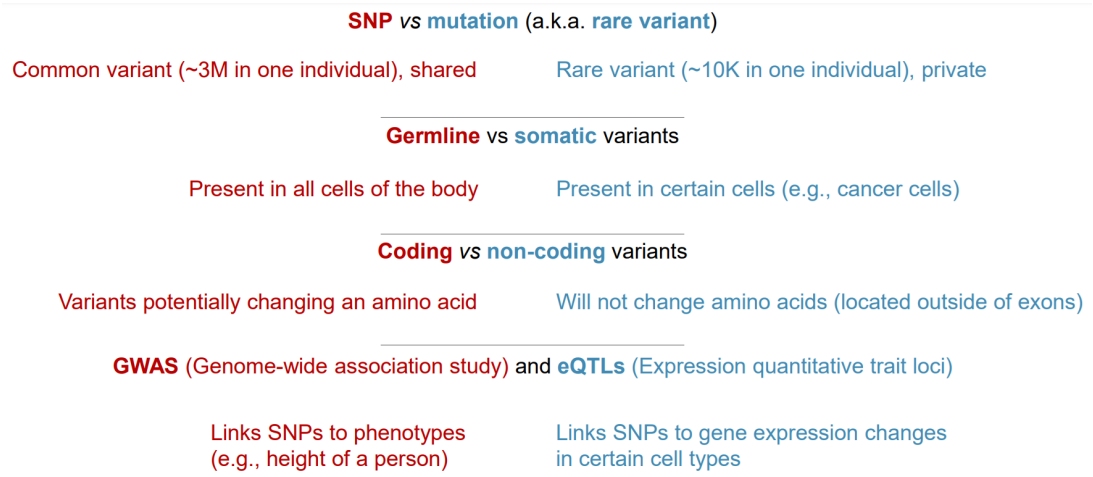

Figure 5: Summary of Genomic Variation

Prediction of Transcription Binding Sites

Prediction or experimental profiling of transcription factor binding sites (TFBS) identifies where transcription factors (proteins that regulate gene expression) bind to DNA. This task is crucial because it helps determine how genes are regulated, revealing how different conditions or mutations might influence gene activity. It is particularly useful for understanding gene regulation mechanisms, studying diseases caused by regulatory mutations, and developing targeted therapies by manipulating gene expression.

Biological Example:

In T-cell acute lymphoblastic leukemia (a type of blood cancer), there is a gene called TAL1 that acts like an "on switch" for the cancer. When TAL1 is overexpressed (producing too much of its protein), it can lead to cancer growth. Scientists noticed that in many patients with this leukemia, TAL1 was making too much protein, but they didn't know why.

Later, they discovered that certain mutations occurred in noncoding regions of DNA (the parts that don't directly make proteins). These mutations weren't in the TAL1 gene itself but in the regions that control when TAL1 is turned on or off. These noncoding mutations "activated" TAL1, making it produce too much protein, which contributed to the cancer.

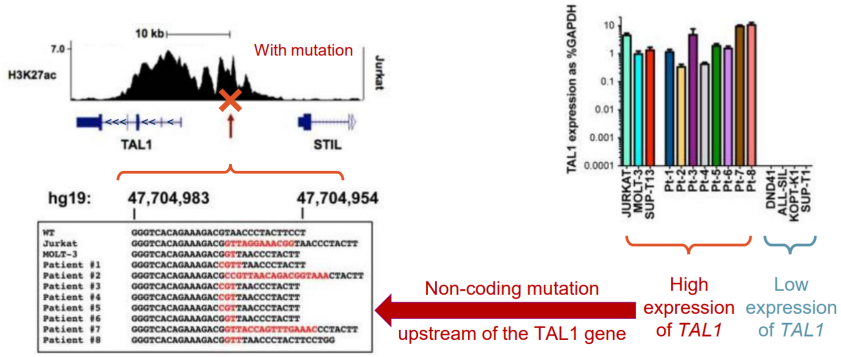

Figure 6: Reason for TAL1 Overexpression in T-Cell Acute Lymphoblastic Leukemia

Historical Statistical Models

Enumeration Method

We model the transcription factor binding site (TFBS) of the transcription factor (TF) of interest as a sequence of length k (e.g., 6-20 base pairs). We then collect a set of sequences that are known to be bound by the TF and use these as positive examples.

In inference, given a new sequence, we simply check if it matches any of the known TFBS sequences. Main limitations of this method are:

- It does not generalize well to unseen sequences, as some binding sites may be missed in the training set.
- It does not account for the variability in TFBS sequences. Binding affinity is not binary and thus a probabilistic model is needed.

IUPAC Consensus Method

The IUPAC (International Union of Pure and Applied Chemistry) code is a standard way to represent ambiguous nucleotide sequences. For example, the IUPAC code for a purine (A or G) is R, and for a pyrimidine (C or T) is Y. Using the IUPAC code, we can represent a set of sequences as a consensus sequence, where each position can have multiple possible nucleotides.

Thus, in inference, given a new sequence, we can check if it matches the consensus sequence, allowing for some variability in the binding site. This method is more flexible than the enumeration method but still has limitations in capturing the full variability of TFBS sequences, mainly stemming from the fact that it does not provide information about the most likely nucleotide at each position.

Position Weight Matrix (PWM)

A position weight matrix (PWM) is a mathematical model that represents the sequence specificity of a transcription factor (TF) binding site. It assumes that each position in the binding site contributes *indepenently* and *additively* to the overall binding affinity.

The **PFM** is a matrix of size $4 \times k$, where k is the length of the binding site and each row corresponds to a nucleotide (A, C, G, T) and each column corresponds to a position in the binding site. The value at each position in the matrix represents the frequency of the corresponding nucleotide at that position in known binding sites.

The **PPM** is a matrix of the same size as the PFM, but the values are normalized to sum to 1 in each column, representing the probability of each nucleotide at each position.

The **PWM** is derived from the PPM by taking the logarithm (usually base 2) of the ratio of the probability of observing a nucleotide at a given position to the background frequency of that nucleotide in the genome.[3] The PWM is used to score new sequences for their similarity to the known binding sites.

Thus, the PWM score is just the sum of the log-likelihoods of each nucleotide at each position in the sequence. The higher the score, the more likely the sequence is to be a binding site for the TF.

Sequence logo: graphical representation of PWM

The height of each letter in the logo represents the information content of that position in the binding site. The more information (i.e. the more conserved the position), the taller the letter.

The actual formula for calculating the information content at a position j is:

$$\text{Information content} = 2 + \sum_{i \in \{A,C,G,T\}} p_{ij} \log_2 (p_{ij})$$

where p_{ij} is the probability of observing nucleotide i at position j, and 2 accounts for the fact that the maximum entropy at each position is 2 bits (since there are 4 possible nucleotides).

Binding site prediction using PWM

Given a new sequence, we can scan the forward and reverse complement of the DNA sequence to find the highest-scoring subsequence (kind of like a sliding window method). The score of the highest-scoring subsequence is then compared to a threshold to determine if the sequence is a putative binding site.

As an output we obtain *strong* and *weak* sites, where the former are more likely to be true binding sites.

Limitations of PWMs

[3] The background frequency simply accounts for the fact that some nucleotides are more common in the genome than others.

- **Independence assumption**: The PWM still assumes that each position in the binding site contributes independently to the overall binding affinity. In reality, positions in the binding site can interact with each other, and the independence assumption may not hold.
- **Need of multiple PWMs**: TFs can have different spacers between two-block binding sites, and a single PWM may not capture this variability.
- **Fixed length**: The PWM assumes a fixed length for the binding site, which may not always be the case. Some TFs can bind to variable-length sites, and the PWM may not be able to capture this flexibility.

Convoluted Neural Networks (CNNs)

CNNs are a type of deep learning model that have been successfully applied to sequence-based tasks, including TFBS prediction. CNNs are particularly well-suited for capturing local patterns in sequences, making them effective for tasks like sequence classification and motif discovery.

This method aims to overcome all the previously mentioned limitations of classical statistical models. CNNs can learn complex patterns in the data, capture dependencies between positions in the binding site, and adapt to variable-length binding sites.

How CNNs work for TFBS prediction:

- **Input representation**: The input to the CNN is typically a one-hot encoded sequence, where each nucleotide is represented as a binary vector of length 4 (A, C, G, T). The sequence is then passed through a convolutional layer that learns filters to detect local patterns in the sequence.
- **Convolutional layer**: The convolutional layer applies a set of filters (also called kernels) to the input sequence, sliding the filters across the sequence to detect patterns. Using different filters can help capturing different binding sequences that correspond to different conformations of the same TF.
- **Pooling layer**: The pooling layer aggregates the output of the convolutional layer, reducing the dimensionality of the data and capturing the most important features in the sequence.
- **Fully connected layers**: The output of the pooling layer is passed through one or more fully connected layers, which learn to classify the sequence based on the features extracted by the convolutional layers.
- **Output layer**: The output layer produces a prediction for each position in the sequence, indicating the likelihood of that position being part of a TFBS. The model is trained using labeled sequences, where the labels indicate the presence or absence of a TFBS at each position. Also, the model can be used to solve a binary classification problem, where the goal is to predict whether a sequence contains a TFBS or not.

How to get the training data for CNNs

When training Convolutional Neural Networks (CNNs) to predict transcription factor (TF) binding, different experimental techniques are used to generate the data needed.

- **Protein Binding Microarray (PBM)**: PBM is a high-throughput technique designed to measure the binding affinity of a transcription factor to all possible DNA sequences of a fixed length (usually 8-10 base pairs). In this method, a microarray (a grid of thousands of tiny DNA spots) contains all possible DNA sequences for a given length. The transcription factor of interest is applied to the array, and the binding affinity of the TF to each sequence is measured.

  Data output: The result is a quantitative score for each DNA sequence, indicating how strongly the transcription factor binds to it. This provides a comprehensive binding profile, which can be used to train CNNs to recognize the DNA motifs preferred by the TF.
- **SELEX data**: Systematic Evolution of Ligands by Exponential Enrichment (SELEX) is a method used to identify DNA sequences that bind to a specific transcription factor. In SELEX, a random pool of short DNA sequences is incubated with the TF. Sequences that bind to the TF are separated and amplified. This process is repeated multiple times, enriching for the DNA sequences that bind most strongly to the TF.

This logic reminds a lot to an evolutionary algorithm, like a genetic algorithm, where the sequences that bind most strongly to the TF are selected and amplified, while the rest are discarded.

Data output: SELEX produces a collection of DNA sequences that show a high binding affinity to the transcription factor. These sequences are then analyzed to identify common motifs or patterns, which are used to generate training data for CNNs to predict TF binding specificity.

- **ChIP-seq data**: Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a widely used method to identify the actual DNA sequences bound by a transcription factor within the entire genome. In ChIP-seq, cells are treated to cross-link proteins (like TFs) to DNA. The transcription factor of interest is then immunoprecipitated (isolated) along with the DNA it is bound to. This DNA is sequenced to determine the exact genomic locations where the TF binds.

  Data output: ChIP-seq provides genome-wide binding profiles of the transcription factor, highlighting specific regions in the genome where the TF is bound. These regions can be used as positive examples to train CNNs, as they provide real binding locations within a living cell, allowing the model to learn binding patterns in their natural genomic context.

Self-Check Questions

1. *Can you explain in 1-2 sentences why it is important to be able to predict binding of transcription factors to a given DNA region?*
2. *Given a set of DNA sequences bound by a transcription factor of interest (e.g., from a SELEX or ChIP-seq experiment), how could one construct a model to predict binding of this transcription factor to an unseen DNA sequence?*
3. *When one trains a CNN model to predict binding affinity of just one transcription factor, why does one usually use several filters (from the biological perspective)?*
4. *When training CNN on ChIP-seq data (e.g., on DNA sequences of length 100-1000bp bound by a TF of interest), what size of filters does it make sense to use for the first layer and why?*

Week 3 - Working with DNA II

Improving the prediction of transcription factor binding

How to improve the prediction of transcription factor binding sites (TFBS)?

- By using different encodings of the DNA sequence, such as k-mer counts, or word-2-vec.
- By optimizing the CNN structure.
- By using improved architectures, such as RNNs and LSTMs.
- By using attention based mechanisms.

K-mer Encodings

K-mer counting is a method to encode DNA sequences by counting the occurrences of all possible subsequences of length k (k-mers). For example, if $k = 2$, the k-mers for the sequence `ACGT` would be `AC`, `CG`, `GT`. The counts of these k-mers can be used as features for training machine learning models.

Word2Vec (and other word embeddings)

Word2Vec is a popular technique in natural language processing (NLP) that learns distributed representations of words in a continuous vector space. These representations capture semantic relationships between words, allowing similar words to have similar vector representations.

In the context of DNA sequences, we can adapt Word2Vec to learn embeddings for the different k-mers in the DNA alphabet. These embeddings can then be used as input features for machine learning models, capturing the relationships between different k-mers based on their co-occurrence patterns.

Recurrent Neural Networks (RNNs)

Particularly useful for sequential data (such as DNA sequences), **Recurrent Neural Networks (RNNs)** are a type of neural network that can capture dependencies between elements in a sequence. RNNs have a hidden state that is updated at each time step, allowing them to remember information from previous steps. The hidden state of an RNN at each time step (along with the next input) is used to update the hidden state (and produce an output) of the next time step.

- RNNs can be **unidirectional** (information flows from past to future) or **bidirectional** (information flows in both directions).
- Main **advantage** of RNN over CNN: RNNs can capture long-range dependencies, which is important for modeling sequential data like DNA sequences. It is well suited for sequential data and is not invariant (nor equivariant) to the order of the input.
- Main **disadvantage** of RNN: It can suffer from the **Exploding/Vanishing gradient problem**, where gradients become very small during training, making it hard to learn long-range dependencies.

Long Short-Term Memory (LSTM)

LSTMs are a type of RNN that is designed to address the vanishing gradient problem. LSTMs have a more complex architecture than traditional RNNs, with additional gates that control the flow of information through the network. These gates allow LSTMs to learn long-range dependencies more effectively than standard RNNs.

Additional hidden state called: **cell state**. The cell state acts as a conveyor belt, allowing information to flow through the network without being altered. The gates in the LSTM (input, forget, and output gates) control how information is added to or removed from the cell state, allowing the network to learn when to remember or forget information.

One can use several layers of RNNs, LSTMs, GRUs (Gated Recurrent Units), or a combination of these to improve the performance of the model.

DanQ is an example of a model that uses a combination of CNNs and bidirectional LSTMs to predict TF binding sites. The CNNs capture local sequence patterns, while the LSTMs capture long-range dependencies in the sequence.

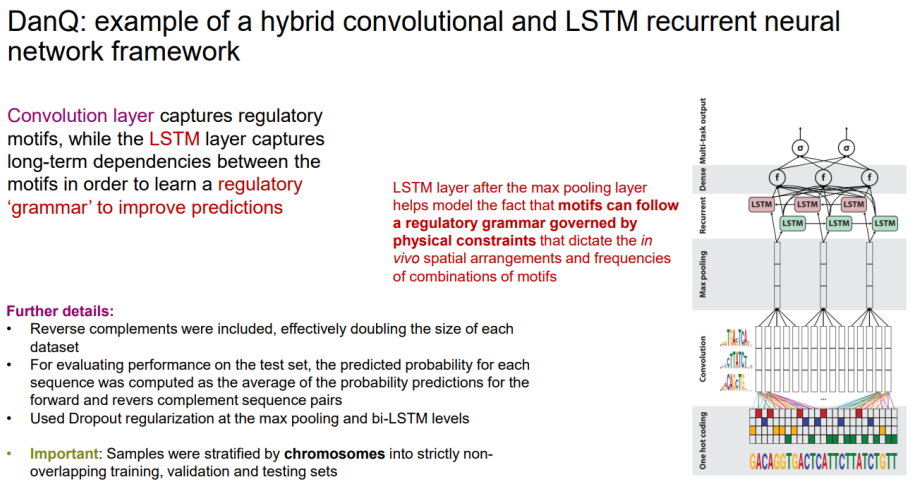

Figure 7: DanQ Model Architecture and Functionality

Using Attention Mechanisms

CNNs have two main issues: (i) they introduce the inductive bias of translation invariance, which might not always be desirable, and (ii) they are not able to capture long-range dependencies in the sequence.

RNNs and LSTMs can capture long-range dependencies, but they may struggle with long sequences and can be computationally expensive.

Attention mechanisms are a way to address these issues. Self-attention is a mechanism that allows the model to focus on different parts of the input sequence when making predictions. It assigns a weight to each input element based on its relevance to the current prediction, allowing the model to learn which parts of the sequence are most important for the task at hand.

In "*Attention is All You Need*", a model called **Transformer** was introduced, which uses a scaled dot-product as the building block of the attention mechanism.

For instance, in **TBiNet**, the authors combine CNNs, LSTMs and attention to train a model on multiple tasks: producing the binding profiles of several TFs.

Experimental Methods to Profile Open Chromatin

TF binding correlates with open chromatin regions, which are accessible to TFs and other regulatory proteins. Open chromatin regions can be profiled using experimental methods like the following:

- **DNase-seq**: DNase-seq is a method that uses the enzyme DNase I to digest open chromatin regions. The DNA fragments that are protected from digestion by TF binding are sequenced to identify the regions of open chromatin.
- **ATAC-seq**: Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a method that uses a hyperactive Tn5 transposase to insert sequencing adapters into open chromatin regions. The DNA fragments with the adapters are then sequenced to identify the accessible chromatin regions.
- **FAIRE-seq**: Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) is a method that uses formaldehyde to cross-link proteins to DNA. The DNA is then sheared, and the regions that are not bound to proteins are isolated and sequenced to identify open chromatin regions.

For instance in **CENTIPEDE**, the authors use DNase-seq data to predict TF binding sites. They use a Bayesian approach to model the relationship between DNase-seq signal and TF binding (combining PWMs and DNase-seq data).

Another example is **DeepGRN**, where the authors employ: a one-hot encoding of the DNA sequence, DNase-seq data, 35bp mapability uniqueness, and a hybrid CNN-LSTM architecture (+ ensembling with attention modules) to predict TF binding sites.

Note: **Mapability** is a measure of how uniquely a sequence can be mapped to the genome. Regions with low mapability are often excluded from analysis, as they can lead to ambiguous results.

Take Away Messages

- The model architecture should be designed based on the understanding of the underlying biological characteristics of the data: e.g., CNNs for local patterns, bi-LSTMs for long-range dependencies, Siamese networks to account for reverse complements, multi-tasking to leverage shared features, etc.
- **Additional information is encourage** to be used, when available: e.g., DNase-seq data, 35 bp mapability uniqueness, gene TSS (Transcription Start Site) locations, histone marks, etc.
- Train and **evaluate wisely**: use cross-validatioon, hold entire choromosomes out for testing, and use appropriate metrics for evaluation (e.g., AUC-ROC, AUC-PR, F1-score, etc.), compare with SotA models, perform ablation studies, etc.

Self-check Questions

1. *You wish to improve over a classic CNN architecture (e.g., DeepBind method) to predict binding sites of a transcription factor. What may you implement in your model (name at least 3 features)?*

   To improve the prediction of transcription factor binding sites, one could implement the following features:

   - Use k-mer encodings or word embeddings to represent the DNA sequences, Instead of one-hot encoding.
   - Optimize the CNN structure by adjusting the number of layers, filters, and kernel sizes. Also, using dilated kernels can help capture long-range dependencies.
   - Incorporate recurrent neural networks (RNNs) or long short-term memory (LSTM) networks to capture long-range dependencies.
   - Use attention mechanisms to focus on different parts of the input sequence.

2. *Why RNN-based models or attention-based models could be expected to offer better prediction accuracy that a classic CNN (e.g., used in the DeepBind approach)?*

   RNN-based models or attention-based models could offer better prediction accuracy than a classic CNN because they can capture long-range dependencies in the sequence, which is important for modeling sequential data like DNA sequences. RNNs and LSTMs have a hidden state that is updated at each time step, allowing them to remember information from previous steps. Attention mechanisms allow the model to focus on different parts of the input sequence when making predictions, assigning a weight to each input element based on its relevance to the current prediction.

3. *What is an alternative to using randomized weights for the initialization of the CNN filters in the first layer when predicting TFBS from DNA sequence?*

   An alternative to using randomized weights for the initialization of the CNN filters in the first layer when predicting TFBS from DNA sequence could be to use pre-computed PWMs (Position Weight Matrices) as the initial weights. These PWMs represent the sequence specificity of a transcription factor binding site and can provide a good starting point for the CNN to learn the features of the binding sites.

Week 4 - Working with Chromatin

Learning Goals:

- **What?**: Predict accessibility of the chromatin; understand the information encoded in our genome.
- **Why?**: To understand outcomes of genetic variation.
- **How?**: By using machine learning models to predict chromatin accessibility from DNA sequence: CNNs, Hidden Markov Models, Latent Dirichlet Allocation, etc.

Underlying biological intuition: genomic variants may affect TF binding. TF binding will affect the presence of histone marks, which in turn will affect chromatin accessibility, and ultimately gene expression.

Recap:

- Histone modifications are associated with active gene transcription and gene repression (silencing).
- Thus histone modifications and open chromatin in promoters correlate with gene transcription levels.

Predicting openness of crhomatin signal from DNA sequence alone is a challenging task and has an important biological application. Why? Because then we can understand how different variations in the DNA sequence can affect the chromatin state and gene expression.

→ For example, **Basset (2016)** used a deep CNN to predict DNase-seq signal from DNA sequence. They showed that the model could predict DNase-seq signal with high accuracy, indicating that the model had learned important sequence features that influence chromatin accessibility.

Basset allowed for the discovery of new motifs that caused Vertiligo, a skin disease, by analyzing the learned filters of the CNN. The newly discovered motifs were then validated experimentally, using ChIP-seq data.

Other examples include **Deep SEA (2015)**, which also used a deep CNN to predict chromatin accessibility and TF binding sites from DNA sequence. The model was able to discriminate between sites bound by TFs that just open chromatin and TFs that bring acetyltransferase to activate enhancers and promoters.

Other methods such as **ExPecto (2017)** used a deep CNN to predict the effects of non-coding genetic variants on gene expression. The model was trained on a large dataset of genetic variants and their effects on gene expression, allowing it to predict the impact of new variants on gene expression.

Basenji (2018) used a hybrid CNN-RNN model to predict chromatin accessibility and gene expression from DNA sequence. The model was able to capture long-range interactions in the DNA sequence and predict gene expression levels with high accuracy.

Statistical Models to Annotate the Genome

Chromatin States are how biologists call the latent structure of the genome, which is not directly observable. Chromatin states are inferred from histone modification data, and they represent different functional elements in the genome, such as promoters, enhancers, insulators, etc.

The goal is to decode chromatin states in an unsupervised manner, i.e., without any prior knowledge of the functional elements in the genome. This is done using **Hidden Markov Models (HMMs)** and **Latent Dirichlet Allocation (LDA)**.

Hidden Markov Models (HMMs) are a type of probabilistic graphical model that is used to model sequential data. In the context of chromatin states, an HMM can be used to model the transitions between different chromatin states along the genome.

(...)

Week 5 - Chromatin and Protein Folding

Learning Goals:

- **What?**: Prediction of 3D structure of the chromatin (importantly, promoter-enhancer interactions and CTCF loops). Also, (part II) prediction of structure of proteins.
- **Why?**: Understand outcome of non-coding mutations. Contribute to "decoding the human genome". (Part II) Understand protein structure and function.
- **How?**: CNN, LSTM, Self-attention.

DNA Looping: CTCF and TADs

- **CTCF**: CTCF is a protein that binds to DNA and plays a key role in the organization of the genome. It is involved in the formation of loops that bring distant genomic regions into close proximity, allowing for interactions between enhancers and promoters. CTCF binding determines "big" DNA loops, that are called Topologically Associating Domains (TADs).
- **TADs**: Topologically Associating Domains (TADs) are regions of the genome that are in close spatial proximity. Promoters and enhancers located within the same TAD are much more likely to interact with each other.

In different diseases, including cancer, mutations in transcription factor binding sites and CTCF binding sites may result in a different DNA looping structure, which will affect gene expression and potentially lead to disease.

But, before we can predict the 3D structure of the chromatin, we need to understand how to profile the 3D structure of the chromatin.

Experimental Methods to Profile 3D Chromatin Structure

Hi-C Maps (Simplified)

- **What is it?** Hi-C is a technique that finds out which parts of the genome are physically close to each other in 3D space inside a cell's nucleus.
- **How does it work?** First, cells are fixed to "freeze" the DNA in its 3D shape. Then, the DNA is cut into pieces using enzymes and reconnected (ligated) in such a way that pieces close to each other are joined. These joined DNA fragments are sequenced to identify which regions are interacting.
- **What do you get?** The sequenced data is used to create a Hi-C map, showing how often different parts of the genome interact.
- **Extra info** Many loops on the map mark regions of regulation and are often anchored by proteins like **CTCF** or **cohesin complex components (RAD21/SMC3)**.

Other Methods:

Hi-ChIP

- **What is it?**: Hi-ChIP combines **Hi-C** (used to study 3D chromatin structure) with **ChIP-seq** (used to identify DNA regions bound by specific proteins). This method focuses on 3D interactions involving a particular protein, such as transcription factors or histone modifications.
- **How it works**: Cells are fixed to preserve chromatin interactions, and DNA is fragmented. Antibodies specific to the protein of interest pull down DNA regions associated with that protein (ChIP step). The captured DNA is processed like Hi-C, where ligated DNA fragments are sequenced to detect interactions mediated by the protein.
- **Output**: Hi-ChIP generates high-resolution maps of protein-mediated chromatin interactions, focusing on regions brought into proximity by the activity of the chosen protein. It is efficient and requires less sequencing than traditional Hi-C.

ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing):

- **What is it?**: ChIA-PET combines **chromatin immunoprecipitation (ChIP)** with **chromatin conformation capture (3C)** to detect chromatin loops mediated by specific proteins. This

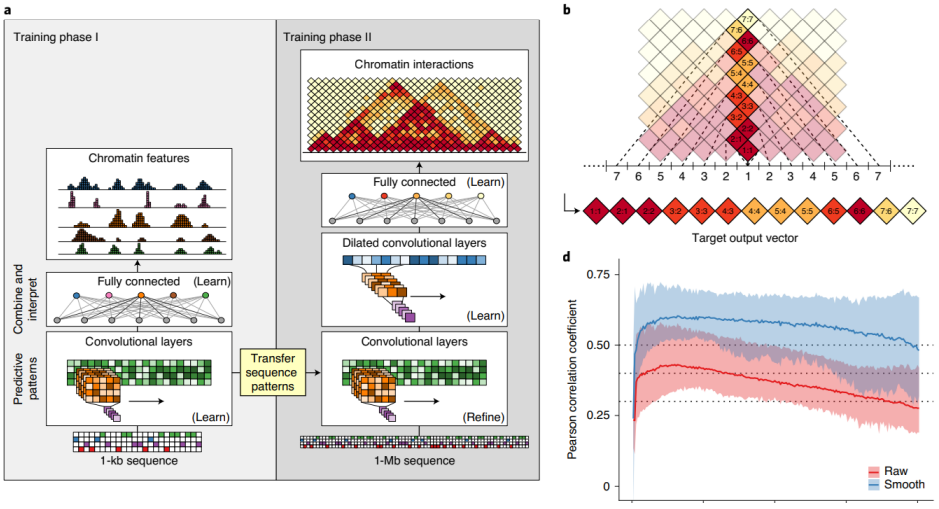

Figure 8: DeepC Model Architecture and Training

method is particularly useful for identifying long-range interactions, such as those involving transcription factors or histone modifications.

- **How it works**: Cells are cross-linked to maintain 3D chromatin structure, and DNA is fragmented. Antibodies pull down DNA fragments bound to a specific protein of interest. Linker sequences are added to DNA ends for tagging, allowing paired-end tags (PETs) corresponding to interactions to be identified. These PETs are sequenced and mapped to the genome.
- **Output**: ChIA-PET produces detailed maps of long-range, protein-mediated chromatin interactions, showing which genomic regions are brought together by the activity of a specific protein.

Comparison: Both methods study protein-mediated chromatin interactions. Hi-ChIP is more streamlined and requires less sequencing, whereas ChIA-PET offers extremely high resolution and is ideal for capturing distant chromatin interactions.

Prediction of 3D Chromatin Interactions

Most methods propose using CNNs, RNNs or attention mechanisms to predict 3D chromatin interactions from DNA sequence.

TODO: Explain DeepC, DeepTACT, TransEPI.

Recall **chromatin interactions** refer to physical contacts between different regions of the genome within the 3D structure of the nucleus. These interactions are essential for gene regulation because they bring distant enhancers close to promoters and other regulatory elements through looping, enabling or repressing gene transcription.

Some Prediction Methods

DeepC

- **Method**: The authors used a dilated CNNs to predict chromatin interaction (scores) for pairs of genomic regions, using DNA sequence only. The training of the model is split into 2 phases: (i) the upstream CNN layers are first trained to predict chromatin features, and (ii) the downstream dilated CNN layers are trained to predict chromatin interactions, refining the weights of the pre-trained upstream layers (transfer learning). Hi-C data was used as the ground truth for the chromatin interactions.
- Biggest **drawback**: the model cannot be extended to a different tissue/cell type without retraining the Phase II with the corresponding Hi-C data.

DeepTACT

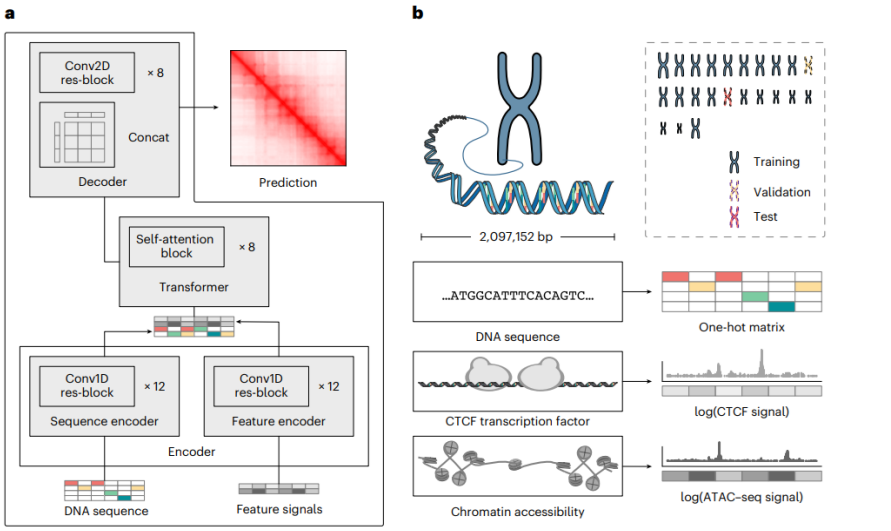

Fig. 1 | De novo, cell-type-specific prediction of 3D chromatin organization with C.Origami. a, A schematic of C.Origami architecture. **b**, C.Origami integrates DNA sequence, CTCF ChIP–seq and ATAC–seq signals as input features to predict Hi-C interaction matrix in 2-Mb windows. Concat, concatenation; Conv1D, one-dimensional convolution; Conv2D, two-dimensional convolution; res-block, residual network block.

Figure 9: C. Origami Model Architecture

- **Method**: Uses LSTMs, CNNs and attention mechanisms to predict chromatin contacts using sequence features and epigenomic information.
- **Drawback**: Controversial results and evaluation procedure. The autors selected the negative set of interactions in a way that the evaluation was altered. One of the reasons for low performance can be that the model does not take into account the distance between two regions in the genome. Also it does not consider global DNA+chromatin context around regulatory elements.

TransEPI

- **Method**: The input to the model is a 2.5Mb region of the genome centered on the E-P pair. They modeled chromatin features in the input (binning in 500bp windows), including: CTCF-binding sites, DNase signals, 5 histone-marks. The model feds the input to a block of CNN layes + MaxPool. Then, the activations are refined by a Transformer encoder, and then the final prediction is made by a fully connected layer.

SotA Model: C. Origami

It is a multimodal deep NN that predicts cell-type-specific chromatin organization using DNA sequence and epigenomic information (CTCF binding and chromatin accessibility). It enables experiments designed to evaluate the impact of genetic variation on chromatin interactions.

The method uses a 2-stage encoder + decoder architecture, where the output of the model is the Hi-C contact matrix (modelates chromatin interactions between pairs of regions in the genome).

The encoder starts with a CNN block that uses residual cnns to extract features from the one-hot encoded DNA sequence and the epigenomic signals. These condensed sequence and genomic representations are subsequently processed by a Transformer encoder to capture long-range dependencies in the data. The decoder is a 2D CNN with a large receptive field that predicts the Hi-C contact matrix.

C. Orignami predicts chromatin interactions within a 2Mb window, to cover typical TAD sizes. The final Hi-C contact matrix uses a bin size of 8,192 base pairs.

Protein Structure Prediction

There are large experimental databases available for protein structures, such as: **Protein Data Bank (PDB)**, which contains experimentally determined 3D structures of proteins resolved using techniques like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM).

In protein structure prediction, the input is usually the **primary amino acid sequence** of the protein (linear sequence of amino acids encoded by the gene). The output consists of:

- A **contact map**, which is a matrix (often binary or probabilistic) indicating whether two amino acids are spatially close in the folded 3D structure (typically within a threshold distance, such as 8 Å). An angstrom (Å) is a unit of length used in chemistry to measure bond lengths and interatomic distances (equivalent to 10^{-10} meters).
- The **3D coordinates of the residues**, representing the spatial positions (x, y, z) of the atoms in each amino acid residue that make up the protein.

Modern computational approaches, such as **AlphaFold**, predict protein structures by directly generating 3D atomic coordinates with high accuracy, bypassing intermediate steps like contact map prediction. These predictions are often benchmarked against experimental structures from databases like PDB.

Other older methods such as **RaptorX** use the contact matrix instead, and then use the contact matrix to predict the 3D structure of the protein. Two residues are said to form a contact if they are spatially proximal in the native 3D structure (i.e. if their euclidean distance of their C-beta atoms is less than 8Å). → less than 2% of the residues in a protein form contacts.

RaptorX uses a CNN-based framework to predict protein contact maps, that will then be fed to the CNS (Crystallography and NMR System) software to predict the 3D structure of the protein.

RaptorX takes as input: (i) 1D sequence of amino acids, (ii) pairwise 2D input that encodes co-evolutionary information between pairs of residues, using mutual information and pairwise contact potential, derived from MSA.

Multiple Sequence Alignment (MSA): MSA is a technique used to align homologous sequences of proteins or nucleic acids to identify conserved regions and structural motifs. MSA is used to infer evolutionary relationships between sequences and to predict the structure and function of proteins.

Self-Check Questions

1. *What is an enhancer? What is a promoter? What is chromatin?* An **enhancer** is a DNA sequence that increases the transcription of a target gene, often by interacting with the promoter. A **promoter** is a DNA region near the transcription start site where RNA polymerase binds to initiate transcription. **Chromatin** is the complex of DNA and proteins (mostly histones) that organizes and compacts the genome within the nucleus.
2. *Is chromatin folding the same across different cell types?* No, chromatin folding varies across cell types because the 3D structure of chromatin reflects cell-type-specific gene regulatory programs. However, some features like TADs (topologically associating domains) are conserved to an extent across cell types.
3. *What is more conserved across cell types: TAD structure (determined by CTCF binding) or enhancer-promoter interactions? Why?* TAD structure is more conserved because it is largely driven by architectural proteins like CTCF, which establish stable genomic domains. Enhancer-promoter interactions are more dynamic and vary across cell types based on specific regulatory needs.
4. *How does the knowledge about (or prediction of) enhancer-promoter interactions (for specific cell types) for wild-type (i.e., non-mutated) and mutated genomes help in understanding potential effects of genomic variants?* Understanding enhancer-promoter interactions helps reveal how genomic variants might disrupt regulatory connections, altering gene expression. This insight is crucial for linking noncoding variants to phenotypic changes or diseases.
5. *What size of the receptive field should one use when predicting enhancer-promoter interactions and why?* The receptive field should span up to 1-2 Mb because most enhancer-promoter interactions occur within this genomic range. Smaller fields may miss distal interactions, while larger ones increase computational complexity unnecessarily.
6. *What information, in addition to the DNA sequence, may be helpful to better predict enhancer-promoter interactions? Why?* Epigenomic data, such as chromatin accessibility (ATAC-seq), histone modifications, and 3D chromatin interaction data, can greatly enhance predictions. These data provide functional and spatial context that complements sequence information.

7. *What deep learning architectures could be useful for the task of predicting enhancer-promoter interactions?* Graph neural networks (e.g., GATs) are useful for leveraging 3D interaction data, while convolutional and transformer models can capture sequence features. Hybrid approaches, such as GraphReg, integrate spatial and sequence information for improved accuracy.

8. *What information does MSA provide for protein folding?* Multiple Sequence Alignments (MSA) highlight evolutionary conservation and covariation between residues, which inform structural constraints. This information is critical for predicting residue contacts and inferring 3D protein structures.

9. *What is a contact map for protein folding?* A contact map is a matrix representation indicating which pairs of residues in a protein are spatially close in the folded structure. It serves as a simplified and interpretable abstraction of the protein's 3D conformation.

Week 6 - Deconvolution of mixed transcriptomics signals

Learning Goals:

- **What?**: How can we deconvolve mixed transcriptional signals.
- **Why?**: To understand the cellular composition of complex tissues.
- **How?**: Using machine learning models to infer cell type proportions from bulk RNA-seq data.
- **Underlying biological intuition**: Different cell types express different sets of genes, which can be used to infer their proportions in a mixed sample.

Recap:

- Recall gene expression determines types and amounts of proteins that ought to be produced by a cell. Thus, it is directly related to the cell phenotype and function.
- **Bulk RNA-seq** measures the average gene expression in a sample, which can be a mixture of different cell types.
- **Deconvolution** is the process of estimating the relative proportions of different cell types in a mixed sample based on their gene expression profiles.
- **Cell type-specific gene expression** can be used to infer the presence of different cell types in a mixed sample.

There are 2 ways to measure gene expression: bulk RNA-seq and single-cell RNA-seq. **Bulk RNA-seq** measures the average gene expression in a sample, which can be a mixture of different cell types. **Single-cell RNA-seq** measures gene expression at the level of individual cells, allowing for the identification of distinct cell types and their gene expression profiles.

Main Differences:

- Bulk RNA-seq provides an average gene expression profile for a mixed sample, while single-cell RNA-seq provides gene expression profiles for individual cells.
- Bulk RNA-seq is more cost-effective and requires less sequencing depth than single-cell RNA-seq, but it lacks the resolution to identify individual cell types.
- Bulk samples can be stored for a long time before preparing for sequencing.

Deconvolution of Mixed Transcriptomics Signals

Deconvolution solves the "Cocktail Party Problem", where the goal is to separate the individual signals from a mixture of signals. In the context of transcriptomics, deconvolution aims to estimate the relative proportions of different cell types in a mixed sample based on their gene expression profiles.

Why is deconvolution important? For example, the composition of tumor microenvironment cells has been found to correlate with patient outcomes and response to therapy. Deconvolution can help identify cell types that are associated with disease progression or treatment response.

There are 2 main types of deconvolution methods: **blind** unsupervised methods, and **reference-based** supervised methods.

Unguided Deconvolution

The problem is modeled as a matrix factorization problem, where the gene expression matrix (X) is factorized into two matrices: a cell type-specific gene expression matrix (S) and a cell type proportion matrix (P): $X = SP$.

Independent Component Analysis (ICA) is a popular method for blind deconvolution. ICA assumes that the gene expression profiles of different cell types are statistically independent, allowing it to separate the mixed signals into independent components. **FastICA** is an efficient and popular implementation of ICA.

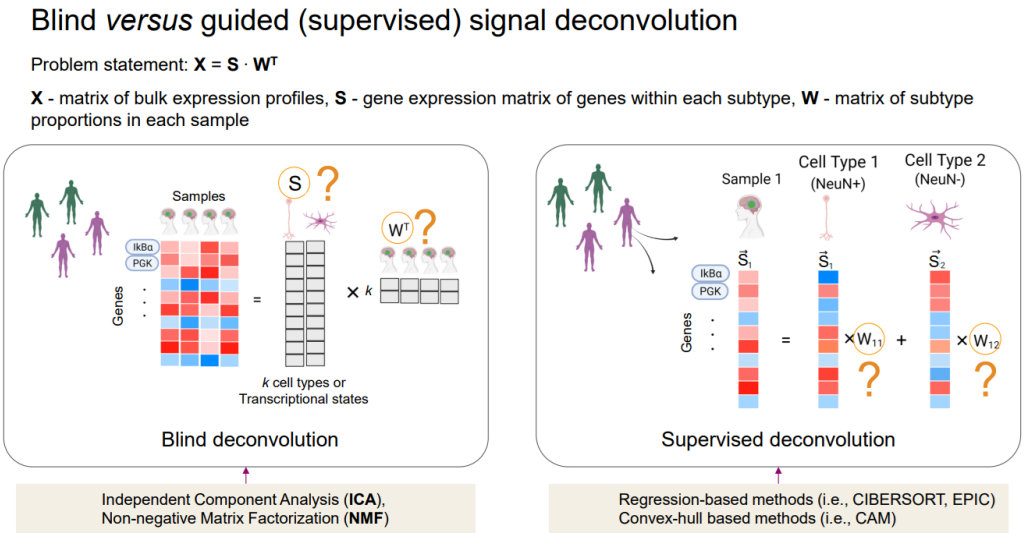

Figure 10: Deconvolution of Mixed Transcriptomics Signals

Non-negative Matrix Factorization (NMF) is another method used for blind deconvolution. NMF assumes that the gene expression matrix can be approximated as the product of two non-negative matrices, which represent the cell type-specific gene expression profiles and the cell type proportions. This approach makes more sense from the biological perspective, as both gene expression and cell type proportions are non-negative. Usually, solved using **Alternating Least Squares (ALS)**.

Supervised Deconvolution

In supervised deconvolution, the cell type-specific gene expression profiles are known (often derived from single-cell RNA-seq data or other reference datasets), and the goal is to estimate the cell type proportions in a mixed sample.

Important methods in this field include: **CIBERSORT**, **MuSiC**, **EPIC**, etc.

CIBERSORT is a widely used method for deconvolving bulk RNA-seq data. It applies support vector regression (SVR) to estimate cell type proportions based on a reference matrix of gene expression profiles from known cell types. CIBERSORT has been extensively used in cancer research to analyze the tumor microenvironment and immune cell composition, providing robust results when the reference profiles are well-matched.

MuSiC is another popular method that estimates cell type proportions from bulk RNA-seq data by leveraging single-cell RNA-seq data as a reference. It uses a weighted least squares framework, to account for inter-sample variability and the heterogeneity of cell types across tissues.

EPIC is a deconvolution method specifically designed for cancer samples. It estimates cell type proportions by modeling bulk RNA-seq data using reference expression profiles that include immune cells and cancer-associated cell types. EPIC does not rely on deep learning but uses a probabilistic model optimized for tumor microenvironment studies.

Self-Check Questions

1. *What do sources inferred by deconvolution methods on bulk RNA-seq correspond to?*

   The sources correspond to cell type-specific gene expression profiles or components representing the distinct transcriptional signatures of different cell types. These inferred profiles are used to estimate the proportions of cell types contributing to the bulk RNA-seq sample.

2. *Why does a linear decomposition make sense?*

   Linear decomposition makes sense because the observed bulk RNA-seq signal is assumed to be a weighted sum of gene expression profiles from individual cell types, with weights corresponding to

their proportions. This linearity reflects how gene expression levels from different cell types combine in a mixed sample.

3. *With blind source deconvolution methods, should we decompose raw values (e.g., read counts normalized by the gene length and library size (FPKM or RPMK)) or log-transformed values (i.e., log2(FPKM + 1) or log2(RPKM + 1))?*

   It is better to use log-transformed values because they stabilize variance and reduce the impact of highly expressed genes, making the data more suitable for statistical modeling. Log-transformation also helps handle the skewness often present in raw RNA-seq read counts.

4. *If we experimentally mixed three cell types and performed bulk RNA-seq on them, which value of k should we use for the blind deconvolution?*

   The value of k should be set to 3 because the number of components (k) in blind deconvolution should match the number of cell types in the mixture.

5. *Do we generally know how many components to expect in real-life settings for blind deconvolution?*

   No, the number of components is often unknown in real-life settings due to the complexity and variability of biological samples. Methods like non-negative matrix factorization (NMF) or independent component analysis (ICA) can help infer k during the deconvolution process.

6. *Can you think about a biological setting when applying blind deconvolution would be a better choice than using reference-based deconvolution such as CIBERSORT?*

   Blind deconvolution is preferable when no reference gene expression profiles are available, such as in rare diseases, novel tissues, or uncharacterized experimental systems. It is also useful when the sample contains unknown or unexpected cell types.

7. *Can you describe in your own words strategies available for deconvolution of bulk transcriptomics signals?*

   Two main strategies are **reference-based deconvolution**, which uses known cell type-specific profiles to estimate proportions (e.g., CIBERSORT, MuSiC), and **blind deconvolution**, which infers both the sources and their proportions without prior knowledge (e.g., NMF, ICA). Additionally, hybrid methods may integrate single-cell RNA-seq data to improve reference profiles or combine supervised and unsupervised approaches.

Week 7 - Introduction to single cell transcriptomics and dropout imputation

Learning Goals:

- **What?**: Understand the basics of single-cell transcriptomics and the challenges of dropout events.
- **Why?**: Analysis of scRNA-seq data is a great playground for ML methods. Analysis of scRNA-seq data helps address many biological questions linked to the heterogeneity of normal and tumor tissues.
- **How?**: Autoencoder-based methods, Markov Affinity-based Graph Imputation of Cells (MAGIC), etc.
- **Underlying biological intuition**: Single-cell transcriptomics captures cell-to-cell variability and reveals rare cell types.

Single cell revolution in 2018 (Breakthrough of the Year by *Nature Methods*): availability of high-throughput single-cell RNA sequencing (scRNA-seq) technologies.

Recall that, in contrast to bulk RNA-seq, scRNA-seq measures gene expression at the level of individual cells, allowing for the identification of distinct cell types and the analysis of cell-to-cell variability within a population.

To allow for quantification of gene expression for each cell, we need toa add a **cell-specific barcode** to each mRNA molecule during the library preparation step. This barcode, called a **Unique Molecular Identifier (UMI)**, allows us to distinguish between mRNA molecules that originate from the same gene but are counted separately due to technical artifacts like PCR amplification bias.

To get a clean final dataset: (i) reads with unique UMIs are counted, (ii) low-quality cells are filtered out, (iii) genes with low expression are filtered out. The output is a 2D matrix, where rows represent cells and columns represent genes.

Visualization of scRNA-seq Data

Visualization is one of the most important steps in any data analysis. In the context of scRNA-seq data, visualization techniques are used to explore the structure of the data, identify cell types, and visualize the relationships between cells.

Common techniques for visualizing scRNA-seq data include:

- **t-SNE (t-distributed Stochastic Neighbor Embedding)**: t-SNE is a dimensionality reduction technique that is commonly used to visualize high-dimensional data in a lower-dimensional space. It is a non-linear transformation that preserves the local structure of the data. It is particularly useful for visualizing the structure of scRNA-seq data and identifying clusters of cells that share similar gene expression profiles.
- **UMAP (Uniform Manifold Approximation and Projection)**: UMAP is another dimensionality reduction technique that is similar to t-SNE but offers several advantages, including faster computation and better preservation of global structure. UMAP is becoming increasingly popular for visualizing scRNA-seq data due to its speed and accuracy. This one also applies a non-linear transformation to the data.
- **PCA (Principal Component Analysis)**: PCA is a classical dimensionality reduction technique that is often used to reduce the dimensionality of scRNA-seq data and visualize the main sources of variation in the data. PCA is useful for identifying the most important features that drive the differences between cells.
- **Hierarchical Clustering**: Hierarchical clustering is a clustering technique that organizes cells into a tree-like structure based on their gene expression profiles. It is useful for identifying groups of cells that share similar expression patterns and for visualizing the relationships between different cell types.

Challenges in scRNA-seq Data Analysis

One of the main challenges in scRNA-seq data analysis is the presence of **dropout events**. Dropout events occur when the expression of a gene is not detected in a cell due to technical limitations of the sequencing process. This leads to zero values in the gene expression matrix, which can complicate downstream analysis (Sparse matrices, zero-inflated data).

To account for this, there are several methods to **impute** the missing values in the gene expression matrix. These methods aim to estimate the true expression levels of genes in cells where they were not detected due to dropout events. A commonly used measure to asssess the accuracy of the imputation is the **correlation** between bulk RNA-seq data and the "bulkified" scRNA-seq data (i.e., the average expression of each gene across all cells), for data extracted from the same tissue.

Imputation Methods

Autoencoder-based Imputation

Autoencoders are a type of neural network that learns to encode data into a lower-dimensional representation and then decode it back to the original input. They are commonly used for dimensionality reduction and feature learning. Among the most common applications of autoencoders we have: anomaly detection, denoising, and data imputation.

In the context of scRNA-seq data imputation, an autoencoder is trained to learn a compressed representation of the gene expression data and then reconstruct the original expression values. The autoencoder is trained on the observed gene expression values and is used to predict the missing values in the gene expression matrix.

However, the latent space of a vanilla autoencoder is not well suited for a generative process, as points sampled from the 1-d latent space can generate meaningless data once decoded. To address this issue, **Variational Autoencoders (VAEs)** are used. In VAEs, the latent space is constrained to follow a specific distribution (e.g., Gaussian), allowing for meaningful sampling and interpolation between data points. We encode the inputs into a distribution over the latent space, and then sample from this distribution to generate new data points. The latent space distribution is usually regularized using the Kullback-Leibler (KL) divergence to ensure that it follows the desired distribution.

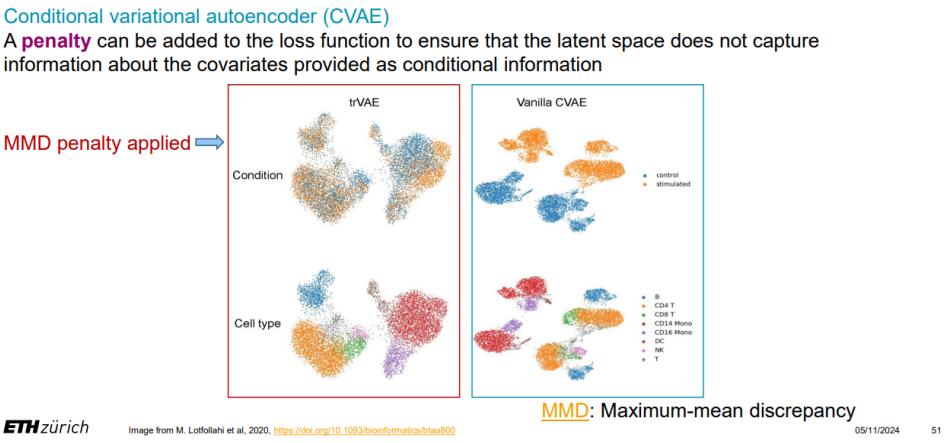

Figure 11: Conditional Variational Autoencoder (CVAE) for scRNA-seq Data Imputation

In the particular case of **Conditional Variational Autoencoders (CVAEs)**, the model is conditioned on an additional input (e.g., cell type or batch information). A penalty (*Maximum-Mean Discrepancy (MMD)*) is added to the loss function to ensure that the latent space does not capture information about the conditioning variable.

Imputation with diffusion: the MAGIC method

The algorithms works as follows:

- **Step 0**: The input to the algorithm is the gene expression matrix with dropout events.
- **Step 1**: The algorithm transforms the read counts into CPM, CP10K, ... to remove library size bias.
- **Step 2**: PCA is applied to the data to reduce the dimensionality of the data.
- **Step 3**: The algorithm computes distances between cells in the reduced space. And transforms these distances using a Gaussian kernel, to retrieve a similarity matrix. Recall a gaussian kernel applies a transformation to the distances, so that the closer cells have a higher similarity score: e.g. $K(x, y) = \exp(-\frac{||x-y||^2}{2\sigma^2})$. These similarities (affinities) are then row-normalized to construct a Markov transition matrix M.
- **Step 4**: The matrix M is raised to a power t, to allow for the diffusion of information across the cells. The resulting matrix represents the probability of transitioning from one cell to another in t steps.
- **Step 5**: The algorithm uses the transition matrix to impute the missing values in the gene expression matrix. The imputed values are computed as a weighted average of the expression values of neighboring cells, with the weights determined by the transition probabilities: $X_{imputed} = MX$.

Overall, MAGIC is a powerful method for imputing dropout events in scRNA-seq data and has been shown to improve the accuracy of downstream analyses, such as clustering and visualization.

TODO: maybe summarise SAVER here.

Self-Check Questions

1. *How does information about UMI help get rid of PCR duplicates (i.e., identical reads resulting from the PCR amplification step prior to DNA sequencing)?*

   Unique Molecular Identifiers (UMIs) are random sequences added to each RNA molecule before amplification, enabling the identification of PCR duplicates. By counting UMIs instead of reads, only unique RNA molecules are measured, eliminating duplicate bias caused by PCR.

2. *What is the definition of the library size (of each cell profiled with scRNA-seq)? How can we normalize the raw read counts for differences in library sizes?*

   Library size refers to the total number of reads or UMIs captured from a cell. Normalization is typically performed by dividing raw counts by the library size and scaling to a common reference, often followed by log-transformation.

3. *What is the dropout effect in the single-cell RNA-seq data and how (in a couple of words) can we correct for it?*

   The dropout effect occurs when lowly expressed genes are randomly undetected, resulting in zeros in the data matrix. It can be corrected using imputation methods or probabilistic models like zero-inflated distributions.

4. *What distribution is generally used to model scRNA-seq raw read counts? How many parameters does it have and what are they?*

   The negative binomial distribution is commonly used, which accounts for overdispersion in the data. It has two parameters: the mean (expected expression level) and the dispersion (variability across cells).

5. *What is the purpose of adding a penalty based on the Hilbert Schmidt independence criterion (HSIC) and Maximum-mean discrepancy (MMD) to the loss function of a CVAE when looking for a biologically interpretable latent space of scRNA-seq data?*

   These penalties encourage disentanglement and alignment of the latent space with biologically meaningful features by reducing dependence on confounding or irrelevant factors. HSIC ensures independence, while MMD aligns distributions to improve interpretability and clustering quality.

Week 8 - Batch Correction, Clustering, Differential Gene Expression and Cell Type Annotation in scRNA-seq Data

Learning Goals:

- **What?**: Understand the importance of batch correction, clustering, differential gene expression, and cell type annotation in scRNA-seq data analysis.
- **Why?**: These steps are crucial for identifying cell types, characterizing cell populations, and understanding gene expression changes in single cells.
- **How?**: Using different batch correction methods, clustering algorithms, differential expression analysis tools, and cell type annotation resources.
- **Underlying biological intuition**: Cells with similar gene expression profiles are likely to belong to the same cell type or state.

Batch Correction

Batch effects refer to the **technical variability** that arises from the separate culturing and sequencing of several groups of cells, resulting in variability that is correlated with technical factors. This is a common issue in scRNA-seq data due to the high sensitivity of the technology to experimental conditions. Some common sources of batch effects include: laboratory conditions, time of the day, instruments used, reagents, etc.

Batch effects can confound downstream analyses, leading to spurious associations between gene expression and technical factors rather than biological differences. To address this issue, batch correction methods are used to remove the technical variability and harmonize the data across different batches.

Regression-based methods

Regression-based methods are commonly used for batch correction in scRNA-seq data. These methods model the gene expression levels as a function of the batch variable and remove the batch-specific effects from the data. The most common approach is to include the batch variable as a covariate in a linear regression model and adjust the gene expression values accordingly.

ComBat is a popular batch correction method that uses an empirical Bayes framework to adjust the gene expression values for batch effects. ComBat estimates the batch-specific effects and shrinks them towards the overall mean, effectively removing the technical variability introduced by different batches.

Harmony is another batch correction method that uses a graph-based approach to align the data across different batches. Harmony constructs a neighborhood graph based on the gene expression profiles of the cells and then optimizes the embedding of the cells to minimize the batch effects while preserving the biological variability. It basically applies soft-clustering to the data and then finds the correction that minimizes the batch effects across the clusters (by maximizing batch-diversity within a cluster). It applies this concept iteratively, until the batch effects are minimized.

Dimensionality reduction methods

Dimensionality reduction methods, such as PCA and t-SNE, can also be used to remove batch effects by projecting the data into a lower-dimensional space where the technical variability is minimized. By focusing on the main sources of variation in the data, these methods can help reduce the impact of batch effects on downstream analyses.

Canonical Correlation Analysis (CCA) is a technique that can be used to align the data across different batches by identifying common sources of variation between the datasets. CCA finds linear combinations of the gene expression profiles that are maximally correlated across batches, allowing for the removal of batch effects and the integration of multiple datasets.

Integrative Non-negative Matrix Factorization (iNMF) is another method that can be used to integrate scRNA-seq data from different batches. iNMF decomposes the gene expression matrix into non-negative factors that represent the cell type-specific gene expression profiles and the batch effects.

By jointly optimizing the factors, iNMF can separate the biological and technical sources of variation in the data.

Deep Learning methods

Similar to the task of data imputation, conditional VAEs can be used for batch correction. By conditioning the model on the batch variable, the CVAE can learn a latent representation of the data that is independent of the technical factors. This allows the model to remove the batch effects from the gene expression profiles and harmonize the data across different batches.

In the end, the goal of batch correction is to mix batches well within the same cluster, but conserve the biological signal across different clusters.

Takeaway messages for batch correction:

- Batch effects can confound downstream analyses in scRNA-seq data.
- Batch correction methods aim to remove technical variability, and avoid spurious associations between gene expression and technical factors.
- Regression-based methods, dimensionality reduction techniques, and deep learning models can be used for batch correction.

Clustering in scRNA-seq Data

Main Challenges in Clustering scRNA-seq Data:

- **High-dimensional data**: scRNA-seq data is high-dimensional, with thousands of genes and cells, making it challenging to identify meaningful clusters.
- **Rare cell types**: scRNA-seq data often contains rare cell types that may be difficult to detect using traditional clustering methods.
- **Batch effects**: Batch effects can introduce technical variability that may confound clustering results, leading to spurious clusters.
- **Dropout events**: Dropout events can result in zero values in the gene expression matrix, complicating the clustering process.

There are different strategies when it comes to cell clustering:

- **K-means clustering**: K-means is a simple and widely used clustering algorithm that partitions the data into k clusters based on the mean of the data points. It requires specifying the number of clusters (k) in advance. Some papers, such as *SC3*, *RaceID*, and *SIMLR*, use K-means as a base clustering algorithm.
- **Hierarchical clustering**: Hierarchical clustering organizes cells into a tree-like structure (dendrogram) based on their gene expression profiles. It initializes all cells as individual clusters and iteratively merges the most similar clusters until a stopping criterion, such as a pre-specified number of clusters, is met. Examples of hierarchical clustering methods include *CIDR*, *BackSPIN*, and *pcaReduce*.
- **Community detection algorithms**: Community detection algorithms identify densely connected subgraphs (communities) within a network. In scRNA-seq data, cells are treated as nodes and similarity measures based on their gene expression profiles are used to define edges. Examples of community detection algorithms include *Louvain* and *Leiden*.

Louvain Clustering

The **Louvain algorithm** is a modularity-based community detection method that identifies clusters by optimizing the modularity score, which measures the density of edges within clusters compared to between clusters. It works iteratively, first assigning each node to its own cluster, then repeatedly moving nodes to neighboring clusters if this improves modularity. Finally, it aggregates nodes within each cluster into a "supernode" and repeats the process until no further modularity improvement is possible.

Leiden Clustering

The **Leiden algorithm** improves upon Louvain by ensuring more robust and well-connected clusters through an additional refinement step. After moving nodes to improve modularity, Leiden partitions clusters further by splitting disconnected components, ensuring each cluster is internally well-connected. This algorithm addresses limitations of Louvain, such as the tendency to produce disconnected or poorly structured clusters, and is often more efficient computationally.

Take-away messages for clustering in scRNA-seq data

- Clustering is a key step in scRNA-seq data analysis for identifying cell types and characterizing cell populations.
- Different clustering algorithms, such as K-means, hierarchical clustering, and community detection methods, can be used to identify clusters in scRNA-seq data. However, traditional methods generally do not transfer well without dimentionality reudction or consideration of single-cell specific factors.
- Community detection algorithms, such as Louvain and Leiden, are popular for identifying clusters in scRNA-seq data due to their ability to capture densely connected subgraphs.

Differential Gene Expression Analysis

What is the difference between **cell types** and **cell states**?

- **Cell types** are distinct categories of cells that share common characteristics, such as morphology, function, and gene expression profiles. Cell types are often defined based on the expression of marker genes that are specific to each type.
- **Cell states** refer to the different functional states that a cell can adopt, such as quiescent, proliferative, or differentiated states. Cell states are more dynamic than cell types and can change in response to environmental cues or developmental signals.

Differential gene expression analysis is a key step in scRNA-seq data analysis for identifying genes that are differentially expressed between cell types or states. It helps us find genes that are characteristic of specific cell types or states and provides insights into the molecular mechanisms underlying cellular heterogeneity.

Gene Set Enrichment Analysis (GSEA) is a widely used method for identifying enriched gene sets in scRNA-seq data. It tests whether a predefined set of genes shows statistically significant differences between two biological conditions. GSEA is particularly useful for interpreting the biological significance of differential gene expression results and identifying pathways or processes that are enriched in specific cell types or states.

How does GSEA work?

1. **Ranking genes**: Genes are ranked based on their differential expression between two conditions (e.g., cell types or states).
2. **Computing enrichment score**: An enrichment score is calculated to measure the degree to which a gene set is overrepresented at the top or bottom of the ranked list. Normalize the enrichment score by the size of the gene set.
3. **Estimating significance**: The significance of the enrichment score is assessed through permutation testing to determine if the observed enrichment is statistically significant. Adjust for multiple hypothesis testing to control the false discovery rate.
4. **Interpreting results**: Enriched gene sets provide insights into the biological processes or pathways that are differentially regulated between cell types or states.

Cell Type Annotation

There are two main strategies for cell type annotation in scRNA-seq data:

- **Marker-based annotation**: Marker-based annotation relies on the expression of known marker genes that are specific to different cell types. By comparing the gene expression profiles of cells to a reference database of marker genes, we can assign cell types based on the expression of these markers. This approach is useful when the cell types are well-characterized and have known marker genes. For instance one can cluster cells and then use marker genes to assign cell types to each cluster.
- **Reference based annotation**: Reference-based annotation where the gene expression profiles of cells are compared to a reference dataset of known cell types. By finding the closest match between the query cells and the reference dataset, we can assign cell types to the query cells. This approach is useful when the cell types are not well characterized or when the marker genes are not known.

Marker-based method example: **GSVA** (Gene Set Variation Analysis) is a method that estimates the activity of gene sets in individual cells based on their gene expression profiles. By comparing the activity of marker gene sets across cells, we can assign cell types based on the expression of these markers.

Self-Check Questions

1. *What is batch effect and what are the three main ways to correct for it?*

   Batch effect is unwanted technical variability introduced during sample processing across different experimental batches. Correction methods include regression-based approaches (e.g., ComBat), mutual nearest neighbors (MNN) alignment, and deep learning methods like scVI.

2. *How can one probabilistically model the counts of a (sc-)RNA-seq experiment? Which batch correction method uses this modeling?*

   Counts from RNA-seq experiments are often modeled using a negative binomial distribution to account for overdispersion. Methods like scVI use this probabilistic modeling as part of their batch correction process.

3. *Why can't traditional clustering methods be applied to scRNA-seq directly? What methods exist to efficiently cluster cells from scRNA-seq?*

   Traditional methods struggle with the sparsity and high dimensionality of scRNA-seq data. Efficient methods include graph-based approaches like Louvain or Leiden and dimensionality reduction techniques like PCA followed by clustering.

4. *How does Louvain clustering work (in a few words)? Explain how one can get clusters using Louvain starting from a gene expression matrix.*

   Louvain clustering identifies densely connected subgraphs by optimizing modularity. Starting from a gene expression matrix, one computes a similarity graph (e.g., k-nearest neighbors) and applies the Louvain algorithm to find clusters.

5. *What purpose does differential gene expression serve? How can one compute the significance of the difference in gene expression between two groups?*

   Differential gene expression identifies genes that are upregulated or downregulated between conditions, providing insights into biological processes. Statistical tests like the Wilcoxon rank-sum test or DESeq2's negative binomial-based test are commonly used for significance.

6. *What purpose does gene set enrichment analysis serve?*

   Gene set enrichment analysis determines whether a predefined set of genes shows significant differences in expression between conditions. It provides biological interpretation by linking differential expression to known pathways or processes.

7. *What are the two types of methods that exist to annotate single cells? What are the main caveats of these two methods?*

   Annotation methods include reference-based approaches (e.g., mapping cells to a known atlas) and unsupervised clustering-based methods. Reference-based methods require high-quality references, while clustering-based methods may misclassify rare cell types or subpopulations.

Week 9 - Trajectory Analysis from scRNA-seq

Learning Goals:

- **What?**: Understand the concept of trajectory analysis in scRNA-seq data.
- **Why?**: To study the developmental processes, cell differentiation, and cell fate decisions.
- **How?**: Manifold learning, graph algorithms...
- **Underlying biological intuition**: Cells progress through different states during development or in response to external signals.

Cell reprogramming means the conversion of one cell type into another. Well-differentiated cell types can be reprogrammed into pluripotent stem cells or other cell types. A **pluripotent stem cell** is a cell that has the potential to differentiate into any cell type in the body. The most well-known type of pluripotent stem cell is the **embryonic stem cell**, which is derived from the inner cell mass of the early embryo.

An example of an application of cell reprogramming is the "in vivo" reprogramming of tumor cells into dendritic cells, to fight cancer. This is achieved by introducing a cocktail of transcription factors that induce the expression of dendritic cell-specific genes in the tumor cells.

PHATE is a visualization tool that allos us to visualize high-dimensional data in a low-dimensional space. It is particularly useful for visualizing the trajectories of cells in scRNA-seq data.

Graph Based Trajectory Analysis

Trajectory analysis is a computational method used to infer the developmental trajectories of cells based on their gene expression profiles. It aims to reconstruct the paths that cells take as they differentiate from one cell type to another or in response to external signals.

Pseudotime is a continuous measure that represents the progression of cells along a trajectory. It is often used to order cells along a developmental trajectory, even if the trajectory is not linear.

Monocle is a popular tool for trajectory analysis that uses graph-based methods to infer the developmental trajectories of cells. It constructs a minimum spanning tree (MST) based on the gene expression profiles of the cells and then orders the cells along the branches of the tree to define the pseudotime.

The algorithm starts by projecting the cells onto a low-dimensional space using dimensionality reduction techniques, such as ICA. Then it creates a graph that connects the cells based on their similarity in the low-dimensional space. The MST is constructed by finding the shortest path that connects all cells in the graph without forming cycles. The pseudotime is then defined as the shortest path along the tree from the root to each cell.

Self-Check Questions

1. *Why are we interested in trajectories?*

   Trajectories allow us to study dynamic biological processes, such as cell differentiation or disease progression, by ordering cells based on their gene expression profiles. This helps reveal intermediate states and identify key genes involved in transitions.

2. *What does pluripotent stem cell mean?*

   A pluripotent stem cell is a type of cell that can differentiate into any cell type of the body. These cells have the ability to self-renew and give rise to all three germ layers (ectoderm, mesoderm, and endoderm).

3. *What is a minimum spanning tree and how to construct it?*

   A minimum spanning tree (MST) is a subgraph that connects all nodes in a graph with the minimum total edge weight and no cycles. It can be constructed using algorithms like Kruskal's or Prim's, which iteratively add the smallest edge that does not form a cycle.

4. *What are the reasons for calculating pseudotime instead of the age of the cell?*

Pseudotime reflects the relative progression of cells along a biological process, such as differentiation, rather than absolute time. Unlike age, it allows ordering of cells based on gene expression similarities, which is useful when true timestamps are unavailable.

5. *A recorded video consists of 5000 individual frames. Someone has shuffled them, so that temporal order is not preserved. How would you reconstruct the original video?*

   To reconstruct the original video, you could compute pairwise similarities between frames (e.g., using visual features) and build a graph where frames are nodes and edges represent similarities. Applying a minimum spanning tree or trajectory inference algorithm would allow you to order the frames based on their visual continuity.

Week 10 - Spatial Transcriptomics

Learning Goals:

- **What?**: Understand the concept of spatial transcriptomics and its applications.
- **Why?**: To study the spatial organization of cells in tissues and identify spatially regulated genes.
- **How?**: GNNs, Optimal Transport, Manifold Alignment...
- **Underlying biological intuition**: Cells in tissues are organized in a spatially specific manner, and their interactions are crucial for tissue function.
- **Applications**: Spatial transcriptomics can be used to study tissue development, disease progression, and the tumor microenvironment.

Spatial omics enables a new level of understanding by providing and analyzing the spatial context and structure of molecular data. We want to know where the cells are located in the tissue, and how they interact with each other.

The old-school traditional approach was to stain the tissue using H&E (Hematoxylin and Eosin) and then look at the tissue under a microscope. This method provided information about the nucli and the cytoplasm of the cells, but not about the gene expression profiles of the cells (for example).

Spatial domain analysis is one of the most important aspects of spatial transcriptomics. It allows us to study the spatial organization of cells in tissues and identify spatially regulated genes. This is crucial for understanding the interactions between different cell types and their role in tissue function.

Spatial Domain Analysis

Motivation: spatial domains delineate different anatomical regions within a tissue. Being able to have a discrete labeling (segmentation) of the tissue is crucial for understanding the spatial organization of cells, and gives us a vocabulary to express which changes occur in tissues with different conditions. For instance, in mice with alzheimer's disease, a study showed that the spatial organization of the different cell types in the brain changes, when compared to healthy mice.

SpaGCN is a graph convolutional network that can be used to analyze spatial transcriptomics data. It creates a graph representation of the tissue, where nodes represent cells and edges represent relationships between cells, based on both their spatial proximity and gene expression profile similarity. The network is trained to predict the spatial domain by performing iterative clustering and classification on the graph.

Spatial Mapping of scRNA-seq Data

scRNA-seq data is not super expensive to generate. Spot-resolution spatial data is also not that expensive. However, **single-cell resolved spatial data** is expensive to generate.

We want to be able to map scRNA-seq data to their corresponding spatial locations. This would give us a high-resolution spatial map of individual cells in the tissue, with their coordinates and gene expression profiles.

Optimal Transport is a mathematical framework that can be used to map scRNA-seq data to spatial locations. In transcriptomics, different type of sc assays are destructive, meaning that the cells are destroyed in the process. Thus we can't do repeated measurements on the same cells. Therefore, to understand and reconstruct the full picture of the population, we would want to have a mapping that relateds two sets of cells.

...

Self-Check Questions

1. *What is spatial omics data? What kind of information does it provide about cells in the profiled samples? How is it different from scRNA-seq data?*

   Spatial omics data provides information about gene expression or other molecular features within their spatial context in a tissue. Unlike scRNA-seq, which loses spatial information, spatial omics preserves the physical location of cells or molecules in the tissue.

2. *What scales of spatial omics data exist? Can you name an example technology for each of the scales?*

   Spatial omics exists at spot resolution (e.g., 10x Visium) and single-cell resolution (e.g., MERFISH). Spot resolution captures groups of cells, while single-cell resolution profiles individual cells with precise spatial locations.

3. *What are some drawbacks of spot-resolution data? Is the entire tissue profiled?*

   Spot-resolution data aggregates signals from multiple cells, making it difficult to distinguish cell types within a spot. Additionally, the entire tissue may not always be profiled due to limited spatial coverage of the technology.

4. *Does single-cell resolution spatial transcriptomic data with a full genomic resolution (20,000 genes profiled) exist?*

   No, current single-cell resolution methods typically profile a subset of genes, often focusing on the most informative ones, rather than all 20,000 genes.

5. *Why is it useful to identify spatial domains when working with spatial data?*

   Identifying spatial domains reveals regions of similar molecular profiles, aiding in understanding tissue organization and functional specialization. It also helps link molecular signals to histological or anatomical structures.

6. *Why are GNNs applicable to a strictly broader class of data than CNNs?*

   Graph Neural Networks (GNNs) can operate on graph-structured data, which does not require a regular grid-like structure as in CNNs. This makes GNNs versatile for spatial omics, where relationships between cells or spots can be represented as a graph.

7. *Why would one want to map scRNA-seq data to physical space?*

   Mapping scRNA-seq data to physical space integrates rich molecular profiles with spatial information, providing a more complete understanding of tissue architecture and function. It enables linking cellular states or types with their locations and interactions within the tissue.

8. *Why is optimal transport a good choice when trying to align data from diverse assays and distinct cell populations? What are simpler alternatives and why don't we want to use them?*

   Optimal transport aligns distributions while preserving relationships between data points, making it effective for aligning data across assays or populations. Simpler methods, like nearest neighbor mapping, may ignore global structure and fail to capture complex correspondences between datasets.

Week 11 - Multi-omics Integration

Learning Goals:

- **What?**: Understand the concept of multi-omics integration and its applications.
- **Why?**: To study the interactions between different molecular layers and identify regulatory mechanisms.
- **How?**: Multi-view learning, data fusion, Autoencoders...
- **Underlying biological intuition**: Different molecular layers (e.g., genomics, transcriptomics, proteomics) interact to regulate cellular processes.
- **Applications**: Multi-omics integration can be used to identify biomarkers, study disease mechanisms, and predict drug responses.

...

Self-Check Questions

1. *Why do we want to integrate multiple -omics layers?*

   Integrating multiple -omics layers provides a more comprehensive view of biological processes by combining complementary molecular information. This can uncover cross-layer interactions, such as linking gene expression (transcriptomics) with protein abundance (proteomics).

2. *What is the difference between matched and unmatched data?*

   Matched data comes from the same set of cells or samples across all modalities, preserving direct correspondences. Unmatched data originates from different sets of cells or samples, requiring computational alignment methods to integrate.

3. *What advantages does middle integration have over early and late integration of different modalities?*

   Middle integration combines the benefits of both early (raw data-level) and late (final results-level) integration by working at the feature level. This approach captures shared patterns while preserving modality-specific information, improving alignment and interpretability.

4. *What kind of representation of multi-omics data can the Seurat package build?*

   Seurat builds a shared latent space that aligns different modalities, allowing integrated analyses across multi-omics data. This enables clustering, visualization, and identification of shared and modality-specific features.

5. *What is the intuition behind the way that Seurat determines the importance of modalities?*

   Seurat weights modalities based on their shared information and relevance to the integrated structure. This ensures that dominant modalities do not overshadow those with weaker but biologically important signals.

6. *What are the roles of anchors in Seurat?*

   Anchors link corresponding cells or features across modalities, serving as reference points for alignment. They help build a consistent shared space by capturing pairwise correspondences between datasets.

7. *How does the framework by Yang et al. ensure that the latent representations of the different modalities are aligned?*

   Yang et al.'s framework uses shared loss functions and regularization terms to align the latent spaces of different modalities. This ensures that corresponding cells or features are mapped to similar points in the shared space.

8. *How can we translate from one modality to another using the framework by Yang et al.?*

   The framework learns mappings between modalities in the latent space, allowing conversion by projecting data from one modality to another. This enables predictions or imputation of missing data in one modality based on another.

Week 12 - Survival Analysis

Learning Goals:

- **What?**: Understand the concept of survival analysis and its applications in genomics.
- **Why?**: To study the relationship between molecular features and patient outcomes.
- **How?**: Cox proportional hazards model, Kaplan-Meier estimator, LASSO...
- **Underlying biological intuition**: Molecular features can be used to predict patient survival and guide treatment decisions.
- **Applications**: Survival analysis can be used to identify prognostic biomarkers, stratify patients, and predict treatment responses.

...

Self-Check Questions

1. *Provide an example of data with right-censoring not mentioned during the lecture.*

   An example of right-censoring is the time until a user unsubscribes from a subscription service, where some users remain subscribed at the end of the study period. The "event" (unsubscribing) has not occurred for all individuals within the observation window, leaving their event times unknown.

2. *Without looking at the slides, could you tell what the value of the step of the Kaplan-Meier estimate will be when an event (or events) happen(s) at time t?*

   The step size of the Kaplan-Meier curve at time t corresponds to the survival probability decrement, calculated as $1 - \frac{\text{Number of Events at } t}{\text{Number at Risk at } t}$. This step accounts for the proportion of individuals who experienced the event relative to those still at risk.

3. *What is the null hypothesis of the log-rank test?*

   The null hypothesis of the log-rank test states that there is no difference in the survival distributions between two or more groups. It assumes that the hazard functions are equal across the groups being compared.

4. *What does the hazard function model?*

   The hazard function models the instantaneous rate at which an event occurs at a given time t, conditioned on survival up to that time. It provides insights into how the risk of the event changes over time.

5. *Imagine two Cox PH models, each model estimates the relative hazard ratio $\frac{h(t,\mathbf{X})}{h_0(t)}$, but values of the first model are twice as large as those of the second model. Will the Cox loss be different? Will the concordance index on the train set (or test, for that matter) be different for these two models?*

   The Cox loss will not be different because it depends on the rank order of the hazards, not their absolute values. Similarly, the concordance index, which measures how well the predicted hazards agree with observed outcomes, will also remain unchanged.

6. *List at least one advantage and one disadvantage of DeepSurv relative to the linear Cox PH model.*

   An advantage of DeepSurv is its ability to model non-linear relationships between features and survival outcomes, enabling it to capture complex patterns in the data. A disadvantage is its increased computational complexity and the need for larger datasets to prevent overfitting, as well as its lack of interpretability compared to linear models.

7. *What is the easiest way to perform multi-omics data integration in survival models and what are more sophisticated alternatives?*

   The easiest way to integrate multi-omics data is to concatenate features from all omics layers into a single input matrix for a survival model. More sophisticated alternatives include using dimensionality reduction techniques, graph-based methods, or specialized architectures like multi-modal neural networks to capture interactions between omics layers.