Sanger sequencing. Steps: Mix a low amount of chain-terminating ddNTP with normal dNTP in PCR reaction, causing random termination of replication; use gel electrophoresis to separate chain-terminated oligonucleotides by size; each ddNTP has a unique fluorescent label, allowing the sequencer to read the sequencing results based on color. Speed is too **slow** using one Sanger sequencing machine. (In each tube, only a single DNA template sequence and a specific primer are contained) Compared with NGS, the error rate is **lower**.

Illumina sequencing. Library preparation: In NGS library preparation, the DNA sample is firstly fragmented. Next, the specialized adapters are ligated to both fragment ends. **Cluster Amplification:** The library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a colonel cluster through bridge amplification (a type of PCR act on flow cell). **Sequencing:** Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases. **Alignment and Data Analysis:** Reads are aligned to a reference genome with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

How can we **measure diverse ranges of genomic molecules** with illumina sequencing? A DNA library is created to represent the signal of interest, such as mRNA, epigenetic markers, or chromatin conformation.

RNA-seq. The fragments are reverse transcribed to cDNA, and the cDNA fragments are then amplified by PCR. 2 types of DNA libraries can be generated for illumina sequencing: single-end library and paired-end library. Gene expression levels can be quantified by counting the aligned reads mapped to the gene annotations.

Profile epigenetic modifications: Chip-seq experiment -> Alignment -> Signal construction -> Peak calling. The histone binding DNA fragments are enriched with immuno-precipitation.

Nanopore sequencing. Read sequences in real time. Nanopore sequencing directly measures single DNA/RNA molecules without PCR amplification. The technique measures voltage changes as the molecule passes through a nanopore transmembrane protein. Maximum read length can be up to 100 kbp. Base modifications can be detected, albeit with some noise, by analyzing signal alterations. **Workflow:** An enzyme unzips the DNA helix into two strands; a protein creates a nanopore in the membrane which holds the adapter molecule; a flow of ions creates an electric current through the nanopore; the adapter molecule keeps DNA bases in place long enough to be identified electronically. **Oxford Nanopore Sequencing.** (electric signal) The electric signal of a nucleotide is predicted by the 5 nucleotides upstream of the current position, which is called the 5-mer sequence.

NGS applications: Basic: DNA-Seq, RNA-seq (Sanger-seq, illumine-seq, nanopore-seq); Epigenetic: Bisulfite-seq, ATAC-seq; Single cell: scRNA-seq.

Bisulfite sequencing: Bisulfite sequencing is used to profile DNA m5C methylation. The DNA fragments are treated with bisulfite, which convert the C into T. However, methylation on C can protect the nucleotide from conversion. Methylation on reads is inferred by tracking nucleotide conversion events.

ATAC-seq. ATAC-seq is a technique for epigenetic profiling that can detect open chromatin regions in a genome. The DNA sample is treated with Tn5 transposase, which introduces sequencing adapters into the accessible regions of the genome. The adapter-ligated fragments are then sequenced, and the sequenced library can be mapped to the accessible regions of the genome.

scRNA-seq. Feature: obtain genomic data from individual cells rather from a mixture of cells; by using the cell specific library preparation techniques (e.x. cellularly unique barcodes), each sample in scRNA-Seq represents a single cell; There are also single cell assays to measure DNA sequences (scDNA-Seq), DNA methylation (scBS-Seq) and chromatin conformation (scATAC-Seq). **Compared to RNA-seq:** RNA-seq obtains average expression level, homogeneous in expression signals. scRNA-seq separates cell populations and detects heterogeneity, identify rare cell populations.

Premapping QC. Raw data problem: fragment length, position bias, fragment sequence bias, read start bias, and library preparation can introduce technical biases from multiple sources. **Fragment GC content bias.** PCR amplification of DNA/cDNA fragments introduces bias in 2nd generation sequencing-based techniques (This is typically the most severe type of technical bias for illumina sequencing) (e.g. DNA-seq, RNA-seq, Chip-seq). **Adaptor contamination.** illumina sequencing uses adaptors, which are repeated sequences attached to both ends of DNA/cDNA fragments; Adaptors facilitate hybridization with probes (on the flow cell) and primers (in bridge PCR); Short fragments can lead to adaptor contamination at the 3' end of reads, especially when the real length exceeds the insert length.

Fastq format. Fastq is a text-based format. It represents each raw read with 4 lines: (1) A sequence identifier with information about the sequencing run and the cluster; (2) The sequence or base calls in the order of 5'-3'; can be A, C, T, G and N; (3) A separator of a plus (+) sign; (4) Characters encoded base call quality scores (Phred scores). The Phred scores or scores have the following definition: $Q = -10 \times \log_{10}(e)$, where e is the estimated probability of the base call being wrong.

Fastqc is a command line tool on Linux/Unix system to generate quality report on fastq files. Output html report, contains multiple QC statistics. **Example QC metrics: Per base sequence quality.** Interpretation: A box plot of Phred scores for every positions of read; The y-axis on the graph shows the Phred scores; The background of the graph divides the y-axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red); Warning will be issued if the lower quartile for any bases fall bellow the red region; If the IQR drop below the read line (<20) near the 3' end, then quality trimming is needed. **Adaptor Content.** Interpretation: The plots shows a cumulative percentage count of proportion of your library which has seen each of the adapter sequences at each position; This module will issue a warning if any sequence is presented in more than 5% of all reads; Problematic if read 3' end contain adaptor contents; Adaptor trimming can be used to remove adaptors.**GC content distribution.** Interpretation: The graph displayed a histogram of GC content over all reads; Warning is issued when observed read GC content distribution (red) is significantly deviant from the expected normal distribution (blue).

Trimming software: Trim Galore. The adaptor sequences and low quality ends can be removed via trimming. Trim Galore can automatically scan & remove adaptors and low quality base calls from the read 3'end. Normalization methods are required to address other types of technical biases, such as GC content biases, in downstream analysis.

Algnment-based: Bowtie2, Tophat2. Bowtie2. align short reads to genome efficiently; Bowtie 2 extracts seed substrings from the read and its reverse complement; Seeds are aligned to the reference genome with the help of the genome index; The precise locations of seeds on the reference genome are calculated from the index; Seeds are extended into full alignments on the genome. **Tophat2.** Pipeline: reads are aligned against the transcriptome (defined in GTF); unmapped reads from the previous step are aligned against the genome; reads are split into smaller segments, and these segments are aligned to the genome using spliced alignment strategy. The alignment tool used by Tophat2 is Bowtie2.

Alignment-free: Kallisto/Salmon. Alignment free: map reads to transcripts without (precise) alignment (without needing to know the exact location of the reads on the transcripts). **Kallisto.** (Pseudo-alignment with TDB graph.) The input for Kallisto includes a reference transcriptome and RNA-seq reads; Kallisto constructs a transcriptome de Bruijn graph (T-DBG) using k-mers as nodes; The T-DBG allows for the efficient identification of compatibility relationships between reads and transcripts, without requiring precise read mapping to the transcripts; Kallisto is able to quantify transcript expression levels based on these compatibilty relationships. **Tool comparisons.** For DNA-seq based assays, botwtie2 is recommended. For RNA-seq based assays, Hisat2 or Tophat2 is recommended. **Limitation.** Alignment-free and traditional alignment-based quantification methods have similar performance for common gene targets such as protein-coding genes; However, alignement-free methods have limitations in analyzing and quantifying lowly expressed genes and small RNAs, particularly when these small RNAs have biological variations; Therefore, sliding windows in peak calling cannot be reliably quantified using alignement-free methods due to their small feature (bin) size.

NGS pipeline: Raw reads (fastq) -> quality control -> trimming -> genome alignment -> quantification -> data analysis.

Read count methods over genomic ranges. 3 major modes are implemented in HTSeq Count (or equivalently R summarizeOverlaps): Union, a read belongs to the feature if any overlap exist between read & feature (Can ensure sensitivity, should be used for bin count in peak calling.); Intersection $\times$, a read belongs to the feature if it falls "within" a feature. i.e. only compatible reads are counted (Can ensure specificity, should be used for transcript quantification.); IntersectionNotEmpty, a loosely defined union mode, reads mapped to > 1 features are still counted to the compatible feature.

Fragment count vs. read count Illumina paired-end sequencing library generates reads from both ends of a DNA/cDNA fragment; The paired reads are expected to be aligned concordantly by genome mapping software,which allows the determination of the range of the fragment on the genome; To quantify PE NGS library, fragment count is often used instead of read count, as it better reflects the underlying biology; In practice, fragment count is approximately half of the corresponding read count.

Isoform level quantification. The challenge of transcript isoform: Alternative splicing can result in genes expressing multiple transcript isoforms; The read coverage of such genes can be convolved by signals originating from multiple transcript isoforms; To estimate isoform-specific expression levels, an EM (Expectation-Maximization)algorithm can be used. **Em algorithm** is an iterative procedure for estimating the expression levels of transcripts, given a compatibility matrix between reads and transcripts. The **goal** of the EM algorithm is to estimate the "probability" of reads coming from each transcript. **Algorithm:** (1) Initialize with some random expression level estimates; (2) E-step: Estimate the probability of reads being assigned to different transcripts, given the compatibility matrix and the current expression level estimates; (3) M-step: Update the expression level estimates by summing the read probabilities (column sums); (4) Repeat steps 2 and 3 until the expression level estimates converge. **Usage:** The EM algorithm is commonly used to estimate transcript expression levels and is implemented in many RNA-Seq quantification software such as Kallisto, salmon, and alpine.

Ratio based quantities. The log of ratio between read counts is often used in functional genomics and epigenetics to represent meaningful quantities. For instance, the log fold change estimate is used to measure how much a gene's expression level has changed across two conditions. log odds ratio estimate is used to measure the abundance of an epigenetic site in a given condition.

Log Odds Ratio: M-level= $\log(\frac{\text{methylated}}{\text{un-methylated}}) = \log(\frac{4}{3})$ over methylation sites in bisulfite sequence;

DBP enrichment level= $\log \frac{\text{IP read count}}{\text{input read count}}$ over peaks in CHIP-seq, where DBP is "DNA binding protein", IP is "immuno-precipitation". **Log Fold Changes: differential gene expression effect size**= $\log \frac{\text{treatment read count}}{\text{control read count}}$ over genes in RNA-seq.

Shrinkage estimator for ratio. One critical **challenge** of log fold change estimates is the high estimation noise (standard error) when counts are small (typically <= 10); Therefore, low-count genes or epigenetic sites are often filtered out or treated as missing values in down stream analysis; A Bayesian solution to reduce statistical noise in low count regions is empirical Bayes shrinkage, which is implemented by R packages such as DESeq2, ashr, andapeglm ...

Sequencing depth. Sequencing depth can be understood as the mean read coverage over the genome transcriptome of an aligned NGS library; Sequencing depth changes a lot across sequencing samples; As a type of technical variation, sequencing depth is often estimated in order to normalize read count. **Causes of sequencing depth variation.** initial # of cells in the sample (NGS library is constructed with different amount of starting cells.); PCR amplification efficiency (Variation in PCR temperature and cycle # can affect the fragment amplification rate.); NGS platform (The fragment detection rate varies across sequencing lanes and platforms.)**Normalize sequencing by depth.** Sequencing depth is often estimated by the location estimators (e.g. mean or median) over read counts in a sequencing sample; A commonly used estimation is by summing up all counts within a sample; A natural way to adjust sequencing depth is to divide counts by the size factors.

RPKM, FPKM, TPM. Effect of feature length. Longer genes express longer transcripts, thereby producing more RNA fragments to be sequenced; The gene lengths (calculated over exonic regions) also need to be normalized when quantifying gene expression. **Feature specific normalization factors.** Normalize over multiple size factors at once by dividing the product of size factors (in this case the sequencing depth and the feature length). **RPKM:** reads per kilobase of transcript per million reads mapped, RPKM= $\frac{\text{Read Count}}{\text{Gene Length} \times \sum_{\forall \text{genes}} \text{Read count}} \times 10^9$, essentially, the RPKM liked measures are making empirical estimation on the probabilities of getting each facet of a biased dice.; **FPKM:** Fragments per kilobase of transcript per million reads mapped, FPKM= $\frac{\text{Fragment Count}}{\text{Gene Length} \times \sum_{\forall \text{genes}} \text{Fragment count}} \times 10^9$, where $\forall$genes is the sum over all genes within a sample; **TPM:** Transcripts per million, TPM= $\frac{\text{Read Count}}{\text{Gene Length} \times \sum_{\forall \text{genes}} \text{Read count/Gene length}} \times 10^6$, where $\forall$gene is sequencing depth estimated on the length normalized count, ensuring sample wised sum of TPM = constant. **Disadvantage:** The 2 samples can be different in both means and variances, normalizing (e.g. RPKM) only over sequencing depths (means) cannot account for the dispersion level difference.

Normalization. The **z-score normalization** is defined by: $z = \frac{X - mean(X)}{sd(X)}$ ($z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$, i is row, j is column); The process transforms any data variable into 0 mean and unit variance (sd = 1); Rescaling is often crucial for downstream analysis, such as clustering and PCA. **Quantile normalization.** Quantile normalization (QN) can enforce identical distributions across any sequencing samples; QN steps: 1. order column (sample) values. 2. substitute values with row (gene) averages. 3. return to the original order; The procedure can effectively remove batch effect in genomic data. **Importance of QN.** Perform QN across biological groups may distort meaningful biological signal; QN should be ideally performed within major biological conditions (e.g. tissues and cell types); Run QN within each tissue or biological condition, not across them; (should apply QN) Large variability within groups, small variability across groups; (should not apply QN) Small varability within groups, large variability across groups. **MA-normalization.** Check for reproducibility: correlation coefficient. **Ma-plot** is a graphic technique for reproducibility assessment; its x axis is (log(E1) + log(E2))/2 (average of the log expressions), its y axis is log(E1/E2) (expression log fold change). One can correct the genomics data by MA-normalization: Choose a reference sample, typically computed by gene-wise averages; Generate an MA-plot for each sample by comparing it to the reference sample, and fit a linear regression to each plot; Normalize each sample by subtracting the fitted values to account for deviations from the expected horizontal line passing origin. **Log transformation.** Count and ratio data types are often beneficial from log transformation; log(count+1) and log fold changes are commonly used in genomic data visualization and data analysis; log is also a mathematically natural transformation for ratio and count. **Reminder:** No single normalization pipeline is guaranteed to perform well for all data; A suitable normalization procedure need to be selected for the specific genomic data type and end application.

Batch effect. Unexpected sources of variations between groups of experiments. Batch effect adjustment by **more feature specific size factors**. **Read genome mappability.** The idea is that some regions along the genome are harder to be (uniquely) mapped due to the presence of repetitive sequences; One can use specialized tool to estimate mappability across any genomes. **GC content bias** is the dependence between fragment count (read coverage) and GC content found in Illumina sequencing data. **Correction for GC content bias.** Estimate GC content bias ($f_j(gc_i)$) with smooth linear regression. **Batch effect factors may beyond accountable technical artifacts.** Batch effects in genomics can be caused by both technical factors and untracked biological factors; Technical factors are easier to adjust after understanding the generation mechanism of technical artifacts; Untracked biological factors, such as age, ethnicity, environmental factors, and epigenetic differences, can confound with the factor of experimental design; Adjusting for bio-based confounding factors is harder since they affect the true biological signals.

Supervised batch effect modeling: combat. Combat is a method used to correct for batch effects when we know the key confounding factors that are causing the batch effects; It works by fitting a multiple linear regression model to the gene expression data, where both the known confounding factors and the experimental design factors are used as covariates in the model; The model then estimates the effect of each covariate on the gene expression data and removes the unwanted variation due to the known confounding factors. **Unsupervised batch effect modeling: SVA.** Unsupervised methods are used when batch factors are unknown and cannot be directly accounted for; These methods estimate the "latent" batch factors using techniques like PCA or other factor analysis algorithms; Surrogate Variable Analysis (SVA) is a sophisticated form of PCA that can estimate batch effect factors while also isolating the influence of experimental treatment factors.

Control experiment. A strong approach to identify artifact is to run a control experiment; When object is known, we can learn artifact $f()$ by observing the deviations in data. E.g. Calibration of antibody unspecific binding by control experiment; using spike-in control to estimate exact sequencing depth (when the same degree of change happens everywhere on the genome, normalizing total sequencing reads to the same number hides the change, whereas normalizing spike-in reads to the same number reveals the global change of read density). Most NGS experiments don't have control. When lacking control, the optimal correction pipeline is often discovered by trial and error; True θ and f are often not identifiable in such cases, as different combinations of θs and fs can generate the same data X; As a result, the optimal correction methods are often different by different downstream applications, since they have different tolerances to different types of errors. **Remainder:** (1) In practice, choosing the right normalization and batch effect removal methods often lead to the most significant performance boost among all steps; (2) The normalization procedures introduced are generally useful for most types of genomic assays. (E.g. DNA-Seq, RNA-Seq, scRNA-Seq, metagenomic sequencing, and CHIP-Seq can all benefit from GC bias correction and quantile normalization.)

PCA/Matrix factorization. In genomics, PCA and matrix factorizations will return the factors of "eigengenes"; The eigengenes can be understood as the characteristic gene expression pattern of a gene module; eigengenes are low dimensional representations of the gene expression matrix. **PCA applications.** Visualizing genomoic assay in 2D; estimation and correction for the batch effect (The idea is that, in heterogenous data set, the top eigengenes are often batch factors). **The principal component correction (PCC)** is a method used to correct for batch effects in gene expression data. The PCC is an effective way to correct for batch effects and other unwanted technical variation in gene expression data, and is widely used in genomic research. The PCC involves two main steps: (1) Per-

form a principal component analysis (PCA) on the normalized expression matrix to obtain the principal components (PCs). The number of top PCs (p) to use is usually determined by a method in the SVA (surrogate variable analysis) package; (2) For each gene, regress the top p PCs using multiple linear regression. The corrected expression values are the residuals of the fitted models. **Nonlinear dimensional reduction.** tSNE/UMAP: non-linear embedding that keep close-by points close using a probabilistic objective. Advantage: Can learn complex non-linear relationships, disadvantage: Axes have no meanings. PCA: finding low dimensional projections that spread data as much as possible. Advantage: High interpretability as factor analysis, disadvantage: work less well for non-linear patterns.

K-means clustering algorithm: (1) Randomly initialize cluster centers; (2) E steps: Assign data points to nearest clusters; (3) M step: Recalculate cluster centers; (4) repeat until convergence.

Classification: random forest algorithm. Create V bootstrap samples, which are training sets resampled with replacement Build a (randomized) decision tree on each bootstrap sample Average the predictions made by the V randomized decision trees(averaging the predictions of multiple models is called ensemble.

Soft clustering: Gaussian mixture model. (1) Randomly initialize Gaussian distribution parameters (μ, σ^2); (2) E step: Assign data points to each Gaussian distribution by probabilities; (3) M step: Recalculate Gaussian distribution parameters (using weighted estimators). (4) Repeat until converge.

Application: Cell clustering in scRNA-seq (The dimensional reduction techniques are doing the "feature extraction for clustering."); batch effect correction in scRNA-seq (Harmony): Original gen expression matrix -> PCA & clustering -> correction (Correction by shrinking data points toward clustering centroids (in a way grouped by batches)) -> Factor loadings -> corrected matrix.

Adjusted p-values for multiple hypothesis testing. Address multiple hypothesis testing problem. **Family wised error rate (FWER)** controlled by **Bonferroni correction**; **False discover rate (FDR)** controlled by **Benjamini-Hochberg correction**.

Bonferroni corrected p-vlaue is defined by $m \times$p-value, where m is the total number of tests conducted (e.g. the # of genes in differential expression analysis). Filtering Bonferroni corrrected p-value at 0.05 ensures FWER < 0.05.

Fail to define the randomness accurately. In practice, the distribution of read counts across biological replicates follows a **negative binomial (NB) distribution** rather than a Poisson distribution; Many classic statistical models (e.g. Poisson/binomial models) fails to account for the over-dispersed nature of genomic count data. **Solution:** Selecting suitable statistical distributioon for your data: it is important to use a statistical model that **specify** the data, i.e. the model used should be able to gerenate the observed data under some parameterization; This can be done by examining the **goodness of fits** of different distribution families on the data, statistical test should be constructed using the best fitting distribution family.

Limited sample size estimating gene variances. Tests integrating multiple replicates require the estimation of dispersion parameters (e.g Gaussian o' and NB over-dispersion parameter). Many experiments only have 2 or 3 replicates, this is too few for accurate dispersion parameter estimation.One solution is to use a smooth curve to predict gene dispersions from gene means.which shares information between all genes. This approach is commonly used by DGEA packages such as Limma, EdgeR, and DESeg2.

Functional annotations. Annotations are stored knowledge from previous biological experiments; Functional annotations are essential for the interpretation of gene sets obtained from the upstream analysis; Gene set enrichment is calculated via the statistical association between gene functions and gene sets.

Gene Ontology (GO) describes our knowledge of the biological domain with respect to three aspects: **Molecular Function:** Molecular-level activities performed by gene products; **Cellular Component:** The locations relative to cellular structures in which a gene product performs a function; **Biological Process:** The larger processes, or "biological programs" accomplished by multiple molecular activities. **For example**, the gene product "cytochrome c" can be described by the molecular function oxidoreductase activity, the biological process oxidative phosphorylation, and the cellular component mitochondrial matrix.

GO Graph. The Gene Ontology (GO) is represented as a graph with terms as nodes and relationships between terms as edges; GO is hierarchical, with more specific child terms and more general parent terms; Terms can have multiple parent terms.

KEGG. Gene annotation via signaling pathway. KEGG is a database for understanding biological systems; KEGG pathway maps are molecular interactions/reaction networks represented in terms of KEGG Orthology groups; These maps can help generalize experimental evidence from one organism to other based on genomic information.

Fisher's exact test. is often used to calculate p-value of association between gene sets and functional terms(Calculating statistical association between annotations). The p-value is calculated by the hypergenometric distribution: (1) Enumerate all possible 2 by 2 tables that are as or more associated than the observed given fixed margins (column and row sums); (2) Use hypergeometric distribution to calculate the probabilities of each table, sum them up and you will get the p-value.

Range based annotations. Transcript annotation Gene & Transcript annotations from GTF/GFF files are often used to annotate range based genomic experiments (e.g. peaks from CHIP-Seq). **Epigenetic markers: ENCODE** (ENCyclopedia Of DNA Elements) It's a database that collects high-quality data about epigenetic markers, expressed transcripts, and epitranscriptomic markers; ENCODE uses strict and well-documented data processing pipelines to ensure data quality; researchers can use the epigenetic markers from ENCODE to annotate their own experiments.

Correlational graphs (undirected graph) Represent the positive / negative correlation between genes. The significantly correlated genes are linked by an undirected edge. **Example:** PPI network, gene co-expression network.

Cause-effect graphs (directed graph) Describe the relationship of causality between genes, such as a gene is changed upon the action of another gene. The direction of the arrowed edge represents cause and effect. **Example:** Cell signaling network, epigenetic regulatory network.

Random v.s. Scale-free network. The distribution of degrees over a graph reveals essential network properties In **random network**, edges are added to node pairs with equal probabilities The degree distribution for random network is Poisson distribution In **scale-free network**, the probability of adding a new edge from node i to a new node increases as the degree of node i increases The degree distribution for scale free network is power distribution

Scale-free network. Average steps between a random pair of nodes in a graph of size n: (1) For a random network, the average path length is log(n); (2) For a scale-free network, the average path length is log(log(n). There by, information transfer is more efficient on a scale-free network; When "attacks" are made by removing nodes from the graph: (1) If the failures happened randomly, the scale-free network is more likely to survive than the random network; (2) If the failures are targeted toward the hub nodes (the nodes with highest degree), then the scale-free network is more vulnerable than the random network.

Hub-nodes: essential proteins. A protein is essential if its knock-down is lethal; In yeast PPI network, the proteins with higher degree (more direct interactions with other proteins) are more likely to be essential proteins; 2240 edges are formed among 1870 nodes (proteins) in yeast PPI network; 93% of proteins have degrees < 3, among them, 21% are essential to yeast survival. 0.7% of proteins have > 15 degree, and 62% of those are essential. The overall correlation coefficient between lethality and connectivity is 0.76.

Co-expression network analysis. Workflow: (1) Pairwise correlation used to construct network; (2) Clustering identifies modules; (3) Differential co-expression analysis identifies regulatory genes; (4) Guilt-by-association approach identifies potential disease genes. **GINIE3:** a high performing network inference algorithm. To **create** a gene regulatory network in GINIE3: (1) For each gene, train Random Forest predictors (f_j) with its expression levels as output and other genes' levels as input; (2) For each predictors, rank all input genes by feature importance; (3) Combine the rankings of all predictors to get the edge scores for network's regulatory links.

Motif discovery: finding repetitive patterns. **Genomic predictive modeling:** predict genomic markers & conservation scores directly from sequences.

Sequence motif. The motif can be discovered from: (1) Sequences of common function (e.g. Zinc- Finger DNA binding domain, phosphorylation sites); (2) From antibody pull down experiments (e.g. CHIP-Seq); (3) Comparative genomics by multiple-sequence alignment. **Function of sequence motif:** (1) Predict DNA / RNA binding protein binding preferences; (2) Predict covalent-modification sites on protein/DNA/RNA; (3) Recover the network of gene expression regulation. (Know which protein/RNA/DNA is regulated by which regulator at what residue)

Computational representation of motif. Motif is often described by PPM (position probability matrix, per base probabilities calculation, consensus matrix), which summarizes the probabilities of observing different nucleotides (rows) at each positions (columns) of the motif sequences.

Discover motifs over a set of long genomic sequences: (1) Known set of functional relevent sequences (e.x. context of single based resolution epigenetic modification sites) -> directly calculate motif PPM; (2) Set of longer sequences that contain potential motifs (e.x. Peaks from Chip-seq experiment) -> Discover potential motifs using EM algorithm.

MEME: motif discovery software. a web based tool to identify unknown short motifs over long input sequences (e.g. > 10000 bp). MEME is a web based tool to identify unknown short motifs over long input sequences (e.g. > 10000 bp). Its core method is based on the following EM algorithm: (1) Randomly initialize motif PPM; (2) Iterate: E-step: Infer expected counts of the motif over long sequences, given the current motif PPM; M-step: Calculate updated motif PPM from the expected counts; (3) Repeat until convergence.

Epigenetic markers prediction from DNA sequence automatically. **Motif based prediction:** (1) Functional relevant DNA sequences (e.g. CHIP-seq peaks) -> Discover motifs -> Given a new DNA sequence, scan for motif as candidate prediction. **Supervised machine learning modeling:** (1) Positive sequences (e.g. flanking region of epigenetic markers & negative sequences (e.g. genome background) -> HMM or Deep learning -> Inference over new sequence using the trained prediction model (Often more accurate and flexible than the motif based method).

Hidden Markov model for CpG island. HMM is a commonly used machine learning model for biological sequences; Considering 2 unfair dices, each with 4 faces of A, T, C, G; one is for genome background and another is for CpG-island; At each roll, we will either keep the current dice, or switch to the other one. The initial roll is selected evenly between the 2 dices; After rolling a series of outcomes, we have generated a DNA string, in which the CpG island properties are encoded by the transition and emission parameters. **State inference (prediction).** After estimating the transition & emission parameters from the data, one can compute the state posterior along the genome using Bayesian inference; State posterior=P(state at position i | the entire observed sequence); Two inference algorithms are often used: Viterbi algorithm and forward backward algorithms. **Viterbi algorithm (Return binary classification).** Classify the regions of CpG island from background on genome; predict protein coding genes. **Forward backward algorithm (Return probabilities).** Estimating a score for evolutionary conservation along the genome (e.g. phastScons score in phylo-HMM).

AUROC: classification evaluation metric. x-axis: FPR=False positives/(TRUE negatives+False positives); y-axis: TPR=True positives/(True positives+False negatives)

Workflow of sequence based supervised learning. (1) A dataset should be randomly split into training, validation and test sets. The positive and negative examples should be balanced for potential confounders (for example, sequence content and location) so that the predictor learns salient features rather than confounders; (2) The appropriate machine learning algorithm is selected and trained on the basis of domain knowledge. For example, CNNs (Convolutional Neural Networks) capture translation invariance, and HMMs capture more flexible spatial interactions; (3) True positive (TP), false positive (FP), false negative (FN) and true negative (TN) rates are evaluated. When there are more negative than positive examples, precision and recall are often considered; (4) The learned model is interpreted by computing how changing each nucleotide in the input affects the prediction.

De Bruijn graph-based genome assembly algorithm: (1) Short reads broken into small pieces (k-mers) and de Bruijn graph constructed; (2) Genome derived from de Bruijn graph by finding the longest possible path (Eulerian walks).

SPAdes. SPAdes is a de-brujin graph based genome assembler; By default SPAdes assembles using kmers of lengths 21, 33, and 55 and chooses the assembly with the best N50 score; N50 can be understood as the median contig length in the assembly.

Variant calling pipeline: (1) Raw reads; (2) Reads pre-processing (Quality check [FastQC], Adapter trimming [Cutadapt]) -> Read alignment/Mapping [Bowtie,, BWA, Novoalign, SOAP, MOSAIK] -> Alignment post processing (Removal of PCR duplicates [PiCard Tools]) -> Variant calling [GATK, SAMTools, FrwwBayes, DeepVariant] -> SNVs and inDels (VCF format)

Variant Call Format (VCF) is the standard file format for storing genetic variant and was developed as part of the 1000 Genomes project.

Copy number variation detection. CNVs are regions of the genome with variable number of copies; DNA-Seq can detect CNVs by analyzing the number of sequencing reads that map to a genomic region; Higher reads suggest a duplication, while lower reads suggest a deletion; CNV detection requires careful normalization and calibration, as read depth can be affected by factors such as GC content, mappability, and sequencing bias.

GWAS. Population structure: (1) Systematic differences in allele frequencies between subgroups in a population due to non-random mating between individuals; (2) Can be estimated from data using statistical methods such as PCA. **Kinship:** (1) Describes the genetic relatedness between individuals in a population; (2) Kinship matrix is often modeled as the covariance of the random effect term.

Expression quantitative trait loci (eQTL). An expression quantitative trait locus (eQTL) is a genetic locus that affects gene expression; eQTL mapping studies use RNA-Seq data to identify eQTLs; Variants are called either from DNA-Seq / RNA-Seq; expression levels are quantified via the regular pipeline, and differential analysis is performed between genotypes.

cis-eQTL. Variants affecting expression of local genes; Found in promoter and gene body of the effect genes. **trans-eQTL.** Variants affecting expression of distal genes; found in other regions.

Accounting for hidden batches. (1) P values of eQTL association are calculated from the linear regression tests with the following regression equation.

In-silico mutation. (1) $f()$ is a sequence based predictive model, it accepts an input of a DNA string and output a probability of the string being a functional epigenetic modification or protein; (2) Calculate the probabilities of WT sequence and mutated sequence (e.g. caused by a SNP): f(WT sequence -> prob1 f(mutated sequence) -> prob2; (3) Inference of SNP function: prob1 » prob2: **loss of function** mutation, prob1 « prob2: **gain of function** mutation.

Homology Modeling Homology modeling is a procedure that generates a previously unknown protein structure by "fitting" its sequence (target) into a known structure (template), given a certain level of sequence homology (at least 30%) between target and template. **Steps:** (1) Identify related structure (template) using database searching tools (E.g. NCBI BLAST protein-protein; database: PDB); (2) Align target (query) sequence with template sequence to check query coverage; (3) Model building for the target using information from the template structure; (4) Model evaluation (Ramachandran plot; similarity; energy simulation).

Molecular dynamics (MD) simulation Molecular dynamics (MD) is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic "evolution" of the system. **Steps:** (1) Initial coordinates (assign a 3D box to the protein), fill with water molecules; (2) Ionization, add Na^+ or Cl^- to stabilize the system; (3) Energy minimization; (4) Equilibration, temperature/pressure; (5) MD production.

How does MD simulation help study proteins? Molecular Dynamics (MD) simulations is a computational method that employs Newton's laws to evaluate the motions of molecules. MD simulations allow protein motion to be studied, by following their conformational changes through time. Proteins are typically simulated using an atomic-level representation, where all or most atoms are explicitly present. (Record protein motion, conformational changes, RMSD value, and effect of pH/temperature/denaturant/solute)

How would you estimate how long it would take to run an MD simulation? The MD simulation time depends on the number of time steps, total number of atoms in the system, and the average number of non-bonded interactions to be computed for each atom at each time step, types of algorithm, and time to compute each non-bonded interaction.

Is it possible for two proteins with the same amino acid count to have different simulation times? Yes, two different proteins with the same number of amino acids have different simulation timescales. The order of the amino acids (primary structure) in the two proteins may be different. This can result in ionic, hydrogen and disulphide bonds to form in different locations in each protein. Such differences may cause variations in the three dimensional structures of the proteins (tertiary structure).

High-throughput virtual screening (HTVS). High-throughput virtual screening (HTVS) is a leading biopharmaceutical technology that employs computational algorithms to uncover biologically active compounds from large-scale collections of chemical compound libraries. **Software**: Autodock, Autodock vina, PyRx, Glide (Schrödinger) etc.

Explain diagram. Selecting the best molecule that inhibits SARS-CoV-2 protease from drug database through HTVS to design some drugs ... (1) Drug database (E.g. antiviral compound database); (2) Filter: Molecular docking (binding energy), PAINS filter, ADMET; (3) Drug-likeliness molecules; (4) Perform MD simulation (purpose: record drug behavior, binding stability); (5) Select few important inhibitors.

Pan-Assay INterference compounds (PAINs) is a term used to describe a broad range of compounds that interfere with biological screening assays by acting through a range of mechanisms. **Chemical absorption, distribution, metabolism, excretion, and toxicity (ADMET)**, play key roles in drug discovery and development.

Mobile domain of lipase. Lipases are important industrial enzymes. Most of the lipases operate at lipid–water interfaces enabled by a mobile lid domain located over the active site. Lid protects the active site and hence responsible for catalytic activity. In pure aqueous media, the lid is predominantly closed, whereas in the presence of a hydrophobic layer, it is partially opened. Hence, the lid controls the enzyme activity. Lids of lipases are amphipathic structures; in the closed conformation, their hydrophilic side faces the solvent, while the hydrophobic side is directed toward the catalytic pocket. As the enzyme shifts to the open conformation, the hydrophobic face becomes exposed and contributes to the substrate-binding region. Therefore, not only the amphipathic nature of the lid but also its specific amino acid sequence is important for activity and specificity of lipases.