- **Sequencing and genome assembly**
    - Advantages of paired end sequencing
    - Graph representations for genome assembly
        - Know how to build a deBruijn from reads
- **Alignment**
    - Dynamical programing for global and local alignments
    - Scoring matricies
        - Log-odds scoring: What do the numbers in BLOSSUM PAM represent?
    - Multiple alignments
        - Scoring: sums of pairs and weighted sums of pairs
        - Progressive alignments
        - Clustalw gap penalty adjustments
- **Molecular evolution**
    - Relationship between time and observed number of changes
    - Saturation
    - Jukes Cantor and Kimura model—know the basic assumption—equations will be provided if needed
        - Transitions vs transversion
    - Variability in substitution parameters
        - Different genomic regions have different evolutionary rates
        - Changes in nucleotide distribution: GC content varies over evolutionary time
    - Synonymous vs non-synonymous
        - Remember: Ka/Ks and dN/dS are different names for the same thing!
        - What does dN/dS>1 mean? How often is it observed in practice?
        - When is it not possible to computer dN/dS accurately, what does it mean for dS to be saturated?
- **Phylogeny**
    - UPGMA: know how it works and what the limitations are
    - Neighbor Joining: know the theoretical guarantees, formulas will be provided if needed
    - Maximum parsimony: know the procedure and understand its weakness--long branch attraction
    - Which methods produce rooted and unrooted trees. How do we pick a root?
- **Gene expression**
    - What do microarrays measure? What is the relationship between microarray quantification and transcript abundance?
    - Normalization methods: quantile and loess normalization (discussed in the classification lecture), quantile-quantile plots

    - Statistical testing for differential expression: T-test, linear models
    - Moderated T-test: understand basic principle—no need to memorize formulas
- **Multiple hypothesis corrections**
    - P-value distributions
    - FWER and FDR definitions
    - Bonferroni correction
    - Methods for computing FDR: Benjamini-Hochberg, q-value, permutation based
- **Clustering**
    - Goals for clustering gene expression data
    - Clustering algorithms, pros and cons of each
        - Hierarchical clustering
            - different linkage methods (complete, average, single)
        - k-means clustering
            - limitations
        - Mixture of Gaussians
            - What parameters are estimated by different models
            - What are the limitations
- **Statistics for sequence based data**
    - ChIPseq
        - How does the experiment work
        - What is the basic output
        - How do we find peaks? What do these peaks represent?
        - Poisson distribution.
    - RNAseq
        - What does the number of reads from a transcript depend on?
        - Understand RPKM normalization
        - Negative binomial distribution—how is it different from Poisson?
- **Genetic variants in population**
    - Definitions of genotypes, alleles, haplotypes, linkage disequilibrium
    - Algorithms for haplotype inference, Clark's algorithm, PHASE
        - Key idea: leverage the similarities across individuals in genomes as can be seen in LD blocks
- **Population structure**
    - HWE, genetic drift
    - STRUCTURE, PCA for detecting population structure
- **Understanding the link between genotypes and phenotypes**
    - Pros and cons in family-based and population-based methods
    - Linkage analysis, single-locus

    - GWAS
        - case-control studies (discrete valued phenotypes): chi square test
        - quantitative trait locus analysis (continuous valued phenotypes): regression analysis
- **Structural variants and how to detect them**
    - Approach based on paired-end sequencing
    - Approach based on read depths
    - Approach based on split reads
- **Motif detection and discovery**
    - PWM, PSSM, scanning the genome with PSSM
    - MEME for motif discovery
- **Gene regulatory networks**
    - Cis/trans regulatory elements and their variation
    - Cis/trans eQTLs
    - Allele-specific expression
- **Probabilistic graphical models: Module networks**