Lecture 1: Introduction

1. What is a p-value? What is multiple test correction?
2. What are the three primary nucleotide sequence databases?
3. What's the difference between a hard link and neighbor?

Lecture 2: Unix and Command Line Interface

1. What are the two main functioning units of a computer? What are the main divisions of each functioning unit?
2. What is an operating system? What are the three main parts of a UNIX operating system? Describe what they do.

Lecture 3: DNA Sequencing

1. Describe the three generations of sequencing. What are their pros and cons?
2. What is a Phred score? How is it calculated?
3. What is the Bermuda standard?
4. What is the current build of the human reference genome?
5. What are the methods for clonal amplification?
6. Describe the process of library construction.
7. Describe the method of first-generation sequencing.
8. Describe the three methods of next-generation sequencing.
9. Describe the two methods of third-generation sequencing.
10. Describe the process "Sequencing by Synthesis".
11. Describe the process of "base calling". What are its advantages and disadvantages?
12. How are sequences indexed with barcodes?
13. Describe phasing/pre-phasing. Why does this occur? How does this relate to the limitations of next-generation sequencing? How is this problem overcome?
14. What is chastity in base calling? What is it used for? How many failures are allowed?
15. How are Phred scores encoded in a FASTQ file?
16. Describe the lines of a FASTQ file. What information is present in the header?

Lecture 4: Sequence Alignment

1. What are the two different classes of alignments? How do they differ?
2. What is a 'heuristic' method? What are the four acts being balanced?
3. Explain the BWA algorithm.
4. What is the difference between BWA aln and mem?
5. What is a mapping quality? What if a read aligns exactly to two different locations?
6. What is the md5sum, and why is it important?
7. What is the significance about the use of read pairs?
8. What is the maximum fragment size of next-generation sequencing?
9. What information is present in the SAM/BAM header?
10. What is the distribution of insert sizes?
11. What does it mean to be 'properly' paired when referring to read pairs?

Lecture 5: ChIP-seq

1. What are the 6 histone modifications and their effects?
2. What are the considerations with using ChIP-seq?
3. What kind of quality control can be done with ChIP-seq data?
4. Describe the ChIP-seq workflow.
5. Explain the MACS2 algorithm. What kind of distribution is assumed? What kind of multiple test correction is used?
6. What are the inputs and outputs for MACS2?

Lecture 6: RNA-seq

1. Describe prokaryotic vs eukaryotic transcription.
2. What are the 4 main methods of measuring gene expression? What is their throughput? How do they work?
3. Describe the RNA-seq workflow.
4. Describe the RNA-seq analysis workflow.
5. What are the 5 reasons for sequencing RNA instead of DNA?
6. What are the considerations with RNA-seq experimental design?
7. What are the three principles of GOOD experimental design?
8. What are the two main sources of variation?
9. What are batch effects?
10. What challenges is RNA-seq faced with?
11. How do you judge good vs bad RNA?
12. How is mRNA isolated?
13. How are RNA-seq libraries constructed?
14. Describe stranded vs unstranded library prep. What are the pros and cons of each?
15. What is RNA-seq used to analyze?
16. Why are PCR duplicates removed?
17. How is RNA-seq library depth chosen?
18. Compare and contrast RNA-seq vs micro-arrays. What are their distributions and detection methods?

Lecture 7: RNA-seq Alignment

1. What do you map RNA-seq reads to? What are these used for? What type of aligner is used?
2. Describe genome mapping of RNA-seq reads. What is the final output?
3. Describe transcriptome mapping of RNA-seq reads. What is the final output?
4. Describe de novo assembly of RNA-seq reads. What is the final output?
5. What are the two methods of spliced mapping?
6. Explain the STAR algorithm.
7. Describe how to run STAR.
8. How do you conduct quality control analysis on RNA-seq alignments?
9. Explain the input/output/options of HTSeq.
10. How are sequence alignments converted into expression values? What are the three common normalization strategies?

Lecture 8: Differential Expression

1. What are the concerns for comparing expression between and within samples? What are instances of the two types of variation?
2. Why conduct differential expression analysis?
3. Explain the DESeq2 normalization algorithm. Which multiple test correction method is used? Why do library size scaling?
4. Explain how to run DESeq2.
5. Explain the output files of DESeq2.
6. What is pathway analysis?
7. What is analysis at the functional level useful for?
8. What are the three main Pathway Databases?
9. What are the three methods of Functional Enrichment Analysis?
10. Describe Over-Representation Analysis workflow. What are its limitations?

Lecture 9: Genome Assembly

1. How many clones are required for a probability of 0.99 that a desired gene in the library and an insert size of 40kb for the E. coli genome? H. sapiens genome?
2. How much did the human genome project cost per base pair? What was the total cost? Total cost per gene?
3. How was the human genome project completed? What was the competing effort?
4. Why perform de novo assembly?
5. What is an assembly?
6. What is reference-guided de novo assembly?
7. How many reads are required to sequence the E. coli genome at 10X coverage with read lengths of 700bp, and a pass rate of 0.8?
8. How do you measure the probability that a base is not sequenced? What distribution is assumed?
9. What are the problems faced by assembly?
10. What does the sequencing depth for completion depend on?
11. What is a graph?
12. What is a Hamiltonian cycle?
13. What is the overlap layout consensus? What are some short-read assemblers based on this method? When is this method used?
14. Describe the Celera workflow.
15. What are challenges faced by the OLC method?
16. What are De Bruijn Graphs? What are some short-read assemblers based on this method? When is this method used?
17. What does changing the size of the k-mers do?
18. Describe the Velvet workflow.
19. How are graphs compressed?
20. Describe SPAdes.
21. How do you evaluate an assembly? What metrics are used? What tools are used?

Lecture 10: Metagenomic Analysis

1. Describe the metagenomic analysis workflow.
2. What are the challenges associated with the data and computation?
3. What design considerations need to be taken into account computationally?
4. What is the Gordian Knot?
5. How do you calculate metagenome size?
6. What is the "Minimal Information" standard?
7. What are the genome reporting standard for metagenomic assemblies?
8. What are the quality control metrics for metagenomic assemblies?
9. Why are sequence bins generated?
10. What are phylogenetic anchors?
11. What is the Critical Assessment of Metagenome Interpretation (CAMI) challenge?

The final exam although cumulative will primarily cover material presented around the RNA-Seq Project through to the end of the course.

There will be questions related to assembly and metagenomic binning. Please review the assembly and metagenomics lectures and be prepared to explain both the challenges and benefits of using metagenome assembled genomes for metabolic pathway reconstruction and taxonomic assignment.

In this light, you should also be familiar with the concept of distributed metabolism as it related to metabolic pathway reconstruction and binning. You should also review the workflow for Project 2 and be prepared to apply it (on paper) in a novel context.

I have also been receiving inquires related to the sequencing platform specifications sheet that I handed out in class and that we used to calculate coverage, etc under different scenarios both in class and during a tutorial with Sean. We will not be using this information on the final and there is no need to try and memorize the sheet. Place your emphasis instead on assembly methods and metrics, workflows and key concepts that we covered during lectures and in group projects.

Lecture 1

1. What is a p-value?
   - p-value is the probability of getting a positive result under the null hypothesis by chance.
     e.g. what is the probability of getting a $\log_2FC > 2$ when a gene is not differentially expressed, simply by chance / at random.
   - multiple test correction adjusts the p-value according to how many tests were conducted
     - Bonferoni – divide the p-value by the number of tests conducted
       e.g. if the p-value threshold is 0.05, if 100 tests were conducted, then the FDR (p-value adjusted) would be 0.0005
     - Benjamin-Hochberg – rank the p-values to choose a threshold, divide rank by number of tests, and multiple by a chosen FDR, then find the highest critical value that is greater than the original p-value → new FDR, and anything ranking higher is significant

2. What are the three primary nucleotide sequence databases?
   - NCBI Genbank (USA)
   - DDBJ (Japan)
   - EMBL ENA (Europe)

   } INSDC

3. What's the difference between a hardlink and a neighbour?
   - hardlink: a direct and objective connection
     - nucleotide sequence → protein sequence
   - neighbour: an indirect and subjective/relative connection
     - nucleotide sequence → 98% identity nucleotide sequence

Lecture 2

1. What are the main functioning units of a computer? What are the main subdivisions of each functioning unit?
   - CPU: central processing unit
     - arithmetic and logic unit: where all the arithmetic and logic operations take place
     - control unit: the computer's "nerve centre"; controls order of operations; accesses, interprets, and directs instructions
   - Memory
     - Primary storage: program storage during execution (RAM)
     - Secondary storage: hard disk, USB, DVD, etc.

2. What is an operating system? What are the three main parts of a UNIX operating system? What do they do?
   - operating system: the suite of programs that maker a computer work
   - Kernel: the "hub" of the OS; allocates time and memory to programs; handles the filestore; handles the communications in response to the system calls
   - Shell: interface between user and kernel; command line interpreter
   - programs: the processes

Lecture 3

1. Describe the three generations of sequencing. What are their pros and cons?

- first-generation: analog sequencing; chain termination method
  - pros: fast, little data
  - cons: not scalable
- next-generation: digital sequencing; clonal amplification; sequencing by synthesis; short reads
  - pros: high throughput, accurate, cheaper
  - cons: basecalling with fluorescence
- third-generation: single molecule sequencing; real time; long reads
  - pros: fast, long range information
  - cons: high error rate, more expensive

2. What is a Phred score? How is it calculated?
- Phred Score: base quality score (higher is better) → confidence score (Phil Green)
- $Q = -10\log_{10}(P)$

3. What is the Bermuda standard?
- Bermuda standard: 6X coverage; Q = 20
- the Human Genome Project was sequenced at the Bermuda standard

4. What is the current build of the human reference genome?
- GRCh38.p12 (hg38): Genome Reference Consortium Human Build 38 patch release 12

5. What are the methods for clonal amplification?
- Oil/aqueous emulsion: e.g. 454, ion torrent
  - microreactor
  - put DNA fragments into each bead (assume Poisson distribution)
  - oligos used to make clonal copies of one DNA strand
- Solid surface: e.g. Illumina flow cell
  - microfluidics slide
  - bridge amplification
- Rolling circle amplification: e.g. Complete Genomics
  - nanoball
  - thousand copies of the same DNA

6. Describe the process of library construction.
1. Isolate the DNA.
2. Shear the DNA.
3. End repair and ligate adapters or molecular barcodes
4. PCR amplify the fragments

7. Describe the method of first-generation sequencing.
- Sanger sequence: chain termination method
  - use of labelled primer, dNTP and ddNTP (terminator)
  - labelled with P32 radioactive tags
  - the oxygen is removed from the hydroxyl group necessary for elongation making that dNTP into ddNTP (irreversible terminator)
  - gets ~250 bp off each sequence at the time
  - the fluore labelling of each base allowed the use of a single lane
  - shifted from a gel to a capillary

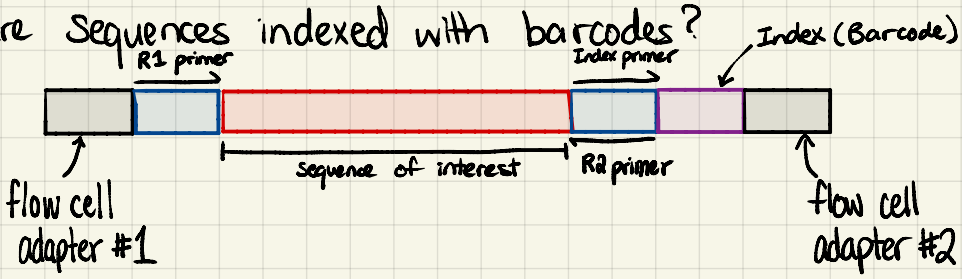
8. Describe the three methods of next-generation sequencing.
- GS-FLX 454 Roche sequencer: oil/aqueous emulsion
  - when a dNTP is incorporated, and an inorganic phosphate is released
  - this phosphate is captured by luciferase, emitting light (pyrosequencing)
  - in each well, there is a bead
  - flow one type of dNTP over the plate into the wells, and record which wells emitted light
  - wash and flow another type of dNTP and repeat
  - with homopolymers, how to quantify how many bases were incorporated using light emission → DRAWBACK
- Ion Torrent Sequencer: oil/aqueous emulsion
  - similar concept as the 454, but instead the H^+ ion released upon base incorporation is measured via current
  - number of H^+ ions released corresponds to how many bases incorporated but is still hard to quantify when it comes to homopolymers → DRAWBACK
- Illumina HiSeq sequencer: solid surface - microfluidic slide
  - clonal arrays generated on a solid surface
  - bridge amplification
  - cluster generation
  - both the forward and reverse strand are sequenced
  - dNTPs are fluorescently labelled
  - reversible terminators

9. Describe the two methods of third-generation sequencing.
- PacBio: SMRT (Single Molecule Real Time)
  - at each emission, time and fluorescence intensity is recorded
  - high error rate for homopolymers → indels
- Oxford Nanopore: Nanopore Amperage
  - single stranded DNA run through a nanopore
  - each base blocks the pore differently, with a unique current signature
  - shifts in current are recorded in real time
  - high error rate for homopolymers → indels

10. Describe the process "Sequence By Synthesis"
- sequencing with reversible terminators instead of the irreversible terminators used in Sanger sequencing

11. Describe the process of "base calling". What are its advantages and disadvantages?
- base calling: calling the base of the sequence depending on its X and Y coordinates from the image
- advantages: create a focal map for read 1 that is used for read 2.
  - high accuracy when it comes to sequencing homopolymers
- disadvantages: requires library diversity to differentiate between clusters.

12. How are sequences indexed with barcodes?



13. Describe phasing/prephasing. Why does this occur? How does this relate to the limitations of next-generation sequencing? How is this problem overcome?
- phasing: running behind in sequencing → synthesizing nucleotide #3 when others are on nucleotide #4
- prephasing: running ahead in sequencing → synthesizing nucleotide #3 when others are on nucleotide #2
- this occurs due to indels, depending on the efficiency of the fluidics and sequencing reactions
- solution: apply statistical averaging over many clusters and sequences to estimate the correlation of signal between different cycles
- the prephasing/phasing becomes so significant that it limits the fragment size/read lengths

14. What is chastity in base calling? What is it used for? How many failures are allowed?
- chastity is the ratio of the brightest intensity to the sum of the brightest and second brightest intensity
- used to detect polyclonal clusters (clusters formed from more than one piece of DNA)
- ideally want chastity to be ≥ 0.6, that a larger difference between the brighest and second brightest intensity is desired
- over the first 25 bases, one failure is allowed

$$\frac{\text{brightest intensity}}{\text{brightest intensity} + \text{second brightest intensity}} \geq 0.6$$

15. How are Phred scores encoded in a FASTQ file?
- they are encoded in base 33, represented by their ASCII character

ASCII char − 33 = Q score

16. Describe the lines of a FASTQ file. What information is present in the header?

1: @HEADER
2: SEQUENCE
3: + OPTIONAL IDENTIFIER
4: BASE QUALITIES IN ASCII BASE 33

→ @INSTRUMENT:RUN:FLOWCELL:LANE:TILE:X:Y:READ:FILTERED:CONTROL:INDEX

Lecture 4

1. What are the two different classes of alignments? How do they differ?
   - local alignment: optimizes the sequence similarity of local regions of the two sequences
     - e.g. BLAST
   - global alignment: optimizes the sequence similarity of the whole sequence and of all sequences
     - e.g. CLUSTAL

2. What is a heuristic method? What are the four acts being balanced?
   - heuristic method: a method that trades time for optimality, that produces a "good enough" solution in a reasonable time frame
   - Optimality: If a read maps to multiple regions in the genome with 1 base mismatch, do we need to know which is correct?
   - Completeness: If a sequence aligns to hundreds of positions in the genome, do we need to know them all?
   - Accuracy and Precision: Provide a way to rank quality of an alignment against other possible solutions
   - Compute Time and Resources: A heuristic may only marginally improve compute time with significant costs (an unworthy compromise)

3. Explain the BWA algorithm.
   1. Index the reference sequence into a hash table of k-mers of size 32
   2. Extract a 32-mer seed from the 5' of the sequence read and use the index to find the genomic location of the seed on the reference
      - 2 mismatches allowed in the seed region
   3. Extend the seed until the mismatch / base quality threshold is reached
   4. A mapping quality is assigned for each alignment.

4. What is the difference between BWA aln and mem?
   - aln: for reads less than 75 nucleotides
   - mem: for reads greater than 75 nucleotides

5. What is a mapping quality? What if a read aligns exactly to multiple locations?
   - how well a sequence (read) aligns to another sequence (reference)
   - if exactly maps to multiple regions, MQ = 0

6. What is the md5sum, and why is it important?
   - md5sum: a digital fingerprint of a file
     - calculates and verifies the 128-bit MD5 hashes of a file
   - used to verify the identity and completeness of a file
   - important when reference genomes are often very large files

7. What is the significance about the use of read pairs?
   - used to resolve multiple alignments to the genome
     - if read 1 aligns to two locations, see where read 2 aligns

8. What is the maximum fragment size of next-generation sequencing?
   - 600 bp

9. What information is present in the SAM/BAM header?
   - @SQ
     - SN: reference sequence name
     - LN: reference sequence length
     - AS: genome assembly identifier
     - MD: MD5 checksum
     - SP: species
     - UR: URI of sequence
   - @PG
     - ID: program record identifier
     - PN: program name
     - CL: command line

10. What is the distribution of insert sizes?
    - normal distribution

11. What does it mean to be "properly" paired when referring to read pairs?
    - read pairs are aligned within the insert size distribution

Lecture 5

1. What are the 6 histone modifications and their effects?
   - H3K4me3: active promoters
   - H3K4me1: enhancers
   - H3K27ac: active enhancers
   - H3K36me3: transcription elongation
   - H3K9me3: repressive / heterochromatin
   - H3K27me3: repressive

2. What are the considerations with using ChIP-seq?
- antibody specificity and sensitivity
- which histone modifications / markers to look at
- sequencing depth
  - 50M read pairs for punctate marks (H3K4me3)
  - 100M read pairs for broad marks (H3K9me3)
- biases

3. What kind of quality control can be done with ChIP-seq data?
- FASTQC
  - insert size distribution
  - diversity of fragments in library
- Library Diversity
  - PCR duplicates occur within library amplification (not clonal)
    - reads with identical +/- read pair coordinates
    - more PCR duplicates = poor library quality
- Immunoprecipitation Quality

4. Describe the ChIP-seq workflow.
1. Sequence IP and Input reads.
2. Align to reference genome.
3. Mark PCR duplicates.
4. Call peaks with MACS2.

5. Explain the MACS2 algorithm. What kind of distribution is assumed? What kind of multiple test correction is used?
1. Find 1000 regions randomly with enriched reads.
2. Calculate the maximum of the distribution of positive and negative reads. The difference of these two maximums is d.
3. Scan the genome with $2d$ window, where the average number of IP reads that align in that bin is $\frac{\#\text{reads}}{2d}$.
4. Scan the genome with a 1kb, 5kb, 10kb and whole genome window. The average of the average number of input reads that aligns is λ.
5. Assuming a Poisson distribution, the area under the curve at the number of IP reads that align in the bin is the FDR, using Benjamini-Hochberg correction.

6. What are the input/outputs for MACS2?
- input: treatment reads aligned to reference BAM/BED file
  - optional: control BAM/BED file, effective genome size
- output: BED and tabular file of peaks, two bedGraph files of scores, for treatment and control lambda

Lecture 6

1. Describe prokaryotic vs eukaryotic transcription.
   - prokaryotic: genes organized in operons
     - tend not to have intronic structures
   - eukaryotic: double stranded DNA template
     - contain exons and introns, as well as UTRs
     - polyadenylation and post transcriptional modifications
     - single stranded mRNA with 5' cap and 3' polyA tail
     - splicing to remove introns
     - mature RNA is then exported to be translated

2. What are the 4 methods of measuring gene expression? What is their throughput? How do they work?
   1. Quantitative PCR (qPCR)
      - throughput: 1-10 genes (low)
        - more can be done but is noisy and expensive
      - design qPCR primer pairs
      - PCR from RNA or cDNA template with fluorescent dye incorporated
      - the amount of intercalating dye fluorescence is proportional to the number of DNA molecules
      - PCR cycle at which fluorescence enters exponential phase (i.e. the cycle threshold (Ct)) is correlated to the number of template DNA molecules
        - on fluorescence vs cycle plot, have housekeeping genes and genes of interest
   2. Microarrays
      - throughput: 100s to 1000s of genes
      - DNA probes (for exons or gene regions) printed on a solid array
      - cDNA or RNA is fluorescently labelled and hybridized to the array
      - array is imaged, a mask applied, and probes quantified
      - fluorescent signal = bound, quantification
      - each DNA spot contains picomoles (10-12) of a specific DNA sequence
        - red: control
        - green: test
        - yellow: overall
   3. Serial Analysis of Gene Expression (SAGE)
      - throughput: 100s to 1000s of genes
      - any transcript with a known sequence can be identified by a short signature sequence or "tag" (uniquely identified)
      - SAGE tags are then cloned into cloning vector before Sanger sequencing
   4. RNA-Seq

3. Describe the RNA-seq workflow.
    1. Get samples of interest (infected/uninfected)
    2. Isolate the RNAs using the polyA tails
    3. Generate cDNAs, select for fragment size, add linkers (unique barcodes for pooling)
    4. Sequence the ends (SBS)

4. Describe the RNA-seq analysis workflow.
    1. Get the RNA-seq reads
    2. Map the reads → Post-alignment QC
    3. Count the reads
    4. Differential expression analysis
    5. Functional Enrichment Analysis

5. What are the 5 reasons for sequencing RNA instead of DNA?
    - functional studies
        - understand gene expression changes in response to an experimental condition (whereas genome is constant)
    - some molecular features can only be observed at the RNA level
        - alternative isoforms
        - fusion transcripts
        - RNA editing
    - predicting transcript sequence from genome sequence is difficult
        - alternative splicing
        - RNA editing
    - interpret mutations that do not have an obvious effect on protein sequence
        - "regulatory" mutations that affect what mRNA isoform is expressed and how much
    - prioritize protein coding somatic mutations (often heterozygous)
        - if the gene is not expressed, a mutation in that gene would be less interesting
        - if the gene is only expressed from the wild type allele, this might suggest loss of function (haploinsufficiency)
        - if the mutant allele is expressed, this could be candidate drug target

6. What are the considerations with RNA-seq experimental design?
    - number of replicates
        - always set up experiments with more replicates than you think you need (not all samples will pass QC)
        - minimum number of 3 biological replicates for statistical analysis
    - library type
        - what kind of RNA?
    - sequencing depth
    - rRNA or globin (for blood samples) removal method
    - ways to minimize batch/confounding effects

7. What are the three principles of GOOD experimental design?
   1. Replication
      - measurements are usually subject to variation and uncertainty
      - replication allows us to estimate the true effects of treatments to further strengthen the experiment's reliability and validity
   2. Randomization
      - assign individuals at random to groups or to different groups in an experiment to reduce bias
   3. Blocking
      - reduces known but irrelevant sources of variation between treatments

8. What are the two main sources of variation?
   1. Biological variation
      - intrinsic to all organisms
      - may be influenced by genetic or environment factors
   2. Technical variation
      - variability in measurements (i.e. the uncertainty in the abundance of each gene in each sample that is estimated by sequencing technology
      - rRNA / library prep method / extraction
      - variations between flow cells / lanes within the same flow cell

9. What are batch effects?
   - sources of variation that are "unrelated" to the biological or scientific variables in a study
   - technical variabilities that potentially contribute to batch effects
      - different personnel / lab
      - different experimental / sample processing dates
      - different sample processing methods / reagents / equipment

10. What challenges is RNA-seq faced with?
   - sample
      - purity
      - quantity
      - quality
   - RNAs consist of small exons that may be separated by large introns
   - relative abundance of RNAs vary wildly
   - RNAs come in a wide range of sizes
   - RNA is fragile compared to DNA (easily degraded)

11. How do you judge good vs bad RNA?
   - RNA quality assessed via bioanalyzer (uses ratio of rRNA band intensities)
   - output = RNA integrity number (RIN)
   - Good RIN → 10
   - Bad RIN → 0

12. How is mRNA isolated?
- Eukaryotic = polyA tail selection
- Prokaryotic: ribosomal depletion

13. How are RNA-seq libraries constructed?
1. PolyA RNA captured (using polyT bead)
2. RNA fragmented and primed
3. First strand cDNA synthesized
4. Second strand cDNA synthesized
5. 3' ends adenylated and 5' ends repaired
6. DNA sequencing adapters ligated (e.g. index)
7. Ligated fragments PCR amplified (5-10 cycles)

14. Describe stranded vs. stranded library prep. What are the pros and cons of each?
- stranded: the second strand is synthesized using dUTP instead of dTTP, and those with uracil will be degraded with uracil-N-glycosylase
  - leftover strands are complementary strand to template
- unstranded: both strands use dTTP
  - unclear which strand reads originated from
  - cannot accurately determine gene expression from overlapping genes

15. What is RNA-seq used to analyze?
- gene expression and differential expression
- transcript discovery or annotation
- allele specific expression (in relation to SNPs or mutations)
- mutation discovery
- fusion detection
- RNA editing

16. Why are PCR duplicates removed?
- duplicates may correspond to biased PCR amplification of particular fragments
- for highly expressed, short genes, duplicates are expected even if there is no amplification bias
- removing them may reduce the dynamic range of expression estimates
- if removed, assess duplicates at the level of paired-end reads (fragments) and not single end reads

17. How is RNA-seq library depth chosen?
- research question
  - gene expression changes
  - alternative splicing
  - mutation calling
  - ↳ rare transcripts require more depth
- tissue type, RNA prep, RNA quality, library construction
- sequencing type (read length, paired/unpaired
- identify publications with similar goals
- pilot experiment

18. Compare and contrast RNA-seq vs micro-arrays. What are their distributions and detection methods?
- Microarrays: use fluorescence for detection
  - has detection limit
  - background noise at low levels
  - saturation at high levels
  - decent correlation for genes with medium level expression
  - poor correlation for genes with high/low levels of expression
  - assumed normal distribution
- RNA-seq
  - no prior assumptions made
  - count what you "see"
  - assumed Poisson distribution

Lecture 7

1. What do you map RNA-seq reads to? What are these used for? What type of aligner is used?
(2) - map reads to genome
  - use a gapped aligner (e.g. STAR, TopHat)
  - used for transcript identification and counting (with annotation)
  - used for transcript discovery and counting (without annotation), and functional annotation (e.g. HTSeq)
(3) - map reads to transcriptome
  - use an ungapped aligner (e.g. BOWTie)
  - used for transcript identification and counting (e.g. RSEM, Kallisto)
(4) - de novo assembly
  - assemble with De Bruijn Graph (e.g. Trinity)
  - align reads to assembled transcripts
  - count using HTSeq-count, RSEM
  - functional annotation with homology, BLAST2GO

5. What are the two methods of spliced mapping?
- exon-first approach (e.g. TopHat)
  - exon read mapping
    - align reads with complete alignment
  - spliced read mapping
    - spliced reads chopped to bits then realigned
- seed-extend approach
  - seed matching
  - seed extend

6. Explain the STAR algorithm.
   1. seed searching
      - search for longest sequence that matches one or more location in the reference genome
      - differently aligned parts of the read are separate "seeds"
      - STAR searches for the unmapped portion only
      - extend MMP to accomodate for mismatches
        - penalizing gaps end up selecting for pseudogenes (introns unspliced)
        - if extension does not give a good alignment → soft clip
   2. Clustering, Stitching, Scoring
      - separate seeds are stitched together
        - stitched based on best alignment using scores based on mismatches, indels, gaps, etc.
        - cluster seeds based on how close they are to a set of "anchor" seeds (uniquely mapped seeds)

7. Describe how to run STAR.
   1. Create a genome index
   2. Map reads to genome

8. How do you conduct quality control analysis on RNA-seq alignments?
   - QoRTs (Quality of RNA-Seq Toolset)
     - 5'-3' bias (gene body coverage)
       - a 3'-bias may be due to RNA degradation or stem from polyA enrichment
     - "correct" stranded protocol used?
   - look for...
     - 3' and 5' bias (in coverage)
     - nucleotide content
     - base/read quality
     - sequencing depth
     - base distribution
     - insert size distribution

9. Explain the input/output options of HTseq.
   - input: BAM file, annotation
   - output: counts file
   - modes: union (default), intersection-strict, intersection-nonempty
   - options: --stranded no/reverse/yes (un/first/second)

10. How are sequence alignments converted into expression values? What are the three normalization strategies?
    - normalize for varying sequencing depth and gene length
        - more sequencing depth = more reads per gene
        - longer genes = more reads for that gene
    - RPKM: Reads per Kilobase Mapped per Million Sequence Reads
        - for single end RNA-seq $\frac{\text{reads on gene}}{10\text{ M reads}} \div \text{length of gene in kb} = \text{RPKM}$
    - FPKM: Fragments per Kilobase Mapped per Million Sequence Reads
        - ~ $\frac{\text{RPKM}}{2}$ for paired-end reads (fragment = 2 reads)
    - TPM: Transcripts Per Million

Lecture 8

1. What are the concerns for comparing expression between and within samples? What are instances of the two types of variation?
    - within-sample comparison: at the same expression level, longer transcripts have more read counts
    - between-sample comparison: higher counts at higher sequencing depth
    - technical: adjusting for differences in library size
        - sequencing depth differences
    - biological: adjusting for differences in library composition
        - library composition differences due to expression differences in tissue type, genetic differences

2. Why conduct differential analysis?
    - need to be able to account for library size and composition
    - count data has a unique distribution, typically a negative binomial distribution
    - large data, small number of biological data
    - variance of the measured data is dependent on the mean
        ↳ "heteroscedasticity"

3. Explain the DESeq2 normalization algorithm. Which multiple test correction method is used? Why do library scaling?
    1. ln(counts)
    2. average each row (average across samples for each gene)
    3. remove genes with infinite values
    4. ln(counts for each gene) − ln(average for each gene)
    5. Calculate the median for each sample
    6. Convert medians to normal numbers to get the final scaling factor
    7. Divide the original read counts by the scaling factor

    ↳ uses Benjamini-Hochberg correction
    - eliminates genes that are only transcribed in one sample type
    - smooth over outlier read counts (via Geometric Mean)
    - median further downplays genes that soak up a lot of genes (emphasizing moderately expressed genes)

4. Explain how to run DESeq2.
   - uses raw counts

   DESeqDataSet() → DESeq() → results()

5. Explain the output file of DESeq2.
   - baseMean: average of normalized count values
   - log2FoldChange: expression changes comparing treated vs. untreated
   - padj: FDR

6. What is pathway analysis?
   - simplify analysis by grouping long lists of individual genes into smaller sets of "related functions" reduces the complexity of analysis

7. What is analysis at the functional level useful for?
   - grouping genes by pathways they are involved in reduces the complexity from thousands of genes to just hundreds of pathways
   - finding a pathway that differs between two conditions can provide (testable) biological explanation than just a gene list

8. What are the three main pathway databases?
   - Gene Ontology: formal representation of a body of knowledge within a given domain
   - KEGG: computer representation of the biological system
   - Reactome: database of signalling and metabolic molecules and how they are organized into biological pathways and processes

9. What are the three methods of Functional Enrichment Analysis?
   - Over Representation Analysis: input is a list of differentially expressed genes
   - Functional Class Scoring: input is the entire data matrix
   - Pathway Topology: use the number and type of interactions between gene and product

10. Describe Over-Representation Analysis Workflow. What are its limitations?
   - statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression, using the 2x2 table method
   1. Create an input list using a threshold
   2. For each pathway, input genes that are part of the pathway are counted
   3. Repeat this for the "background" list of genes (all the genes with an RNAseq count)
   4. Each pathway is tested for over-representation in the list of input genes with statistical tests (e.g. hypergeometric test)
   - limitations
      - considers number of genes alone but ignores the log_2FC change associated with them
      - uses the most significant genes and discards others → information loss
      - assumes each gene is independent of other genes
      - assumes each pathway is independent of other pathways

Lecture 9

1. How many clones are required for a probability of 0.99 that a desired gene in the library and an insert size of 40kb for the E. coli genome? H. sapiens genome?

$$n = \frac{\ln(1-P_b)}{\ln(1-f)} = \frac{\ln(1-0.99)}{\ln\left(1-\frac{40\times10^3}{4.6\times10^6}\right)} = 5293 \text{ clones} \;/\; \text{H. sapiens} = 3.7\times10^6 \text{ clones}$$

2. How much did the human genome project cost per base pair? What was the total cost? Total cost per gene?
   - $1 per bp
   - $3 billion total
   - $146,000 per gene (20,500 predicted genes)

3. How was the human genome project completed? What was the competing effort?
   - low throughput method → Sanger sequencing
   - human DNA (13 individuals) → ligated to vector → transformed yeast to clone → Sanger
   - competing effort: Celera Genomics used a more scalable small insert clone approach that cost $100 million at $4,878 per gene (but many more gaps in assembly)

4. Why perform de novo assembly?
   - assembly of a non-model or uncultivated organism, where there is no reference genome
   - interest in novel genomic elements not present in reference
   - update old references with newer sequencing technologies

5. What is an assembly?
   - hierarchal data structure that maps the sequence data to a putative reconstruction of the target
   - contigs provide a multiple sequence alignment of reads plus the consensus sequence
   - scaffolds define the order and orientation of contigs, as well as the size of the gaps

6. What is reference-guided de novo assembly?
   - use a related reference to aid in assembly process
   - method still involves de novo steps

7. How many reads are required to sequence the E. coli genome at 10X coverage with read lengths of 700bp, and a pass rate of 0.8?

$$\text{reads} = \frac{(\text{coverage})(\text{genome size})}{(\text{read length})(\text{pass rate})} = \frac{(10)(4.6\times10^6)}{(700)(0.8)} = 82143 \text{ reads}$$

8. How do you measure the probability that a base is not sequenced? What distribution is assumed?

$$n_0 = e^{-c} = e^{-\frac{LN}{n}} = e^{-\frac{\text{read length} \times \text{number of reads}}{\text{genome length}}}$$

- poisson distribution

9. What are the problems faced by assembly?
   - read length shorter than even smallest genomes
     - overcome by oversampling the target genome
   - repeat sequences, sequencing error, non-uniform coverage, computational complexity

10. What does the sequencing depth required for completion dependent on?
   - genome size
   - G+C content
   - repeat content
   - sequencing platform

11. What is a graph?
   - representation used to draw inferences between related nodes and the links between them (edges)

12. What is a Hamiltonian Cycle?
   - visit each node exactly once, returning to the start
   - NP-hard problem in CS

13. What is the overlap layout consensus? What are some short-read assemblers based on this method? When is this method used?
   - overlap between sequence reads is used to create a link between them, resulting in a directed graph based on all versus all alignment that completes a Hamiltonian cycle
   - the genome sequence is then assembled by aligning sequences of adjacent contigs and calculating a path that will produce a non-redundant sequence → Tiling Path
   - assemblers: Celera, SSAKE, VCAKE, Newbler
   - nodes are whole reads → computationally expensive → for long reads

14. Describe the Celera workflow.
   1. Mask repeats and low complexity regions, trimming to remove low quality bases
   2. Identify overlaps between reads at user defined length and identity thresholds
   3. Assemble high-confidence sequences from overlaps into contigs
   4. Order and orient contigs
   5. Attempt to resolve sequencing errors

15. What are challenges faced by the OLC method?
   - polymorphisms, errors, repeats and other ambiguities result in forking paths
   - RAM intensive

16. What are De Bruijn Graphs? What are some short-read assemblers based on this method?
    - Euler path that crosses every edge exactly once
    - Velvet, MEGAHIT, SPAdes, ABySS, ALLPATHS, SOAPdenovo

17. What does changing the size of the k-mers do?
    - too small k: graph becomes tangled
    - too large k: graph becomes fragmented

18. Describe the Velvet workflow.
    1. Create a k-mer hash table with sequence coverage information (multiplicity)
    2. Construct the DBG
        - compress graph based on unambiguous edges
        - simplify using tip removal, bulge removal, and removal of erroneous graph connections
        - resolve repeats and extract contigs

19. How are graphs compressed?
    - remove tips, bulges, and frayed ropes
      (errant base calls)

20. Describe SPAdes.
    - use multiple k DBGs } multi-layer
    - use paired k-mer DBG }

21. How do you evaluate an assembly? What metrics are used? What tools are used?
    - contiguity: N_{50}
    - completeness: BUSCOs
    - tools: QUAST

Lecture 10

1. Describe the metagenomic analysis workflow.

trim reads → assemble into contigs → filter out duplicates → bin contigs → assemble into scaffolds → check contamination and completeness → classify taxonomy → predict ORFs → assign KOs → RPKM → pathway analysis

2. What are the challenges associated with the data and computation?
   - a lot of data
     - volume: billions of sequences across many samples
     - variety: environmental conditions, taxonomy and function
     - complexity: multiple hierarchal levels of organization
     - dynamism: lateral gene transfer and viral reprogramming
   - heterogeneous computation
     - software: many different software have to work together
     - computation: local resources or high performance computing
     - parallelism: local (CPU) and cloud (distributed)

3. What design considerations associated with data and computation?
   - modularity: rapid swap-in/swap-out and integration of new modules
   - input/output: archive, retrieve raw and compare
   - memory and performance issues: fast and cache friendly

4. What is the Gordian Knot?
   - multiple coexisting genotypes represented in one sample
   - cross-section of naturally occurring heterogeneity (microdiverse clusters)
   - on average, no two clones derived from same genome
   - sequence space is complex and interwoven
     - lateral gene transfer
     - symbiosis and syntrophy
     - ecological context and natural history

5. How do you calculate metagenome size?

$$G_{im} = \sum_{i=1}^{l} n_i G_i$$

6. What is the "Minimal Information" standard?
   - inclusion of sampling location, environmental conditions

7. What are the genome reporting standards for metagenomic assemblies?
   - Finished: single continuous sequence; $Q \geq 50$
   - HQ: multiple fragments, rRNA and tRNAs; >90% completeness, <5% contamination
   - MQ: many fragments, no metrics; ≥50% completeness, <10% contamination
   - LQ: many fragments, no metrics; <50% completeness, <10% contamination

8. What are the quality control metrics for metagenomic assemblies?
   - N_{50}
   - completion
   - contamination

9. Why are sequence bins generated?
   - group taxonomically using phylogenetic anchors, G+C content, k-mer profiling

10. What are phylogenetic anchors?
    - translated nucleotide sequences mapped onto reference phylogeny
    - trees for rRNA, 18S rRNA, and 40 universal COGs

11. What is the Critical Assessment of Metagenome Interpretation (CAMI) challenge?
    - encourages benchmarking programs