Next Generation Sequencing Technologies

- These technologies represent a vast improvement over chain termination sequencing(Sanger)
- Use DNA synthesis through a template to determine sequences.
- Include:
    - Roche/454
    - Illumina
    - ABI SOLiD
    - Ion Torrent
    - Pac Bio
    - Oxford Nanopore(technically '3rd generation')
- Phred quality scores: a quality score that helps resolve discrepancies during sequencing. The score represents confidence in each position during sequencing.
- FASTQ format: allows for each nucleotide in a sequence to also have a represented phred score using ASCII.
    - Similar to the FASTA format but the 1st header line begins with the "@" character, and FASTQ includes a second header line for quality scores that begins with the "+" character.
    - Each FASTQ file follows the same format: Header line with "@", DNA sequence line, Second header line with "+", and the quality line.
- Sequence Read Archive: international project from NCBI, EMBL, and DDBJ that stores raw sequencing data from projects or studies, and is designed to prevent the loss of that data.
- Metagenomics: study of all genetic material in an environmental sample, provides a NGS way to determine which species/DNA/RNA samples are present.

Genome Sequence Assembly:

- NGS technology reduces cost by sequencing smaller fragments, becomes challenging to align all of the small pieces into one cohesive sequence.
- De Novo Assemblers:
    - Split DNA into small fragments and then either sequence those pieces by single-end (both fragments sequenced at one end) or paired-end (both fragments sequenced from each end). Knowing which method was used important in assembling those sequences.
    - Some assemblers better from prok genomes (SPAdes/Velvet) and others better for euk(ABySS).
    - **Velvet:** uses the de Bruijn graphical approach to make contigs that can then be assembled into a sequence.
- Reference based assemblers: possible when a reference sequence is available to be used as a guide.
    - Bowtie/Bowtie2: most known and used, with a method of inexact matching to align mistmatches
    - BWA: uses Burrows-Wheeler aligner, preferred by researchers
    - HISAT: newer than BWA, becoming a quick favorite.

Variant Calling

- Mismatches: reasons that a sequence may not match the reference include: sequencing error, heterozygous SNPs, homozygous SNPs, or PCR amplification errors.
- Variant vs SNP: how to tell if it was an error or there is an actual SNP in the sequence, use MORE COVERAGE. More reads of the sequence = more confidence.
- Five Variant Calling Approaches:
    - FreeBayes
    - SAMtools
    - GATK Unified Genotyper
    - GATK Haplotype Caller
    - Platypus
- Variant Calling Output: BAM files generated from the aligned reads/reference genome. These are in a VCF(variant call format) where each line shows chromosomal position, expected ref. seq nucleotide, and other statistics.
- Suggested Variant Call Workflow:
    - FASTQ Groomer: reformat to Sanger/illumina 1.9 encoding
    - FASTQ Trimmer: trim low quality
    - Bowtie/BWA/HISAT: align FASTQ to reference
    - SAM-to-BAM conversion
    - Filter SAM/BAM to limit to a genomic region
    - Sort BAM
    - FreeBayes: use aligned BAM file and reference genome to call variants.
-