**Introduction**

“The ‘Omics Era”

* Model organisms 🡪 non-model organisms to create novel genomes. Mention some organizations that are active in this.
* Explain briefly the importance of genomics in conservation and population genetics

**The ‘omics era**

***De novo* genome sequencing**

There are three main parts to the whole genome sequencing. The first part is library preparation and sequencing, this is the done using high quality extracted DNA and high-throughput technology like Illumina. Following this, the assembly is the computational process of constructing longer sequences through piecing together shorter sequence reads. Finally, the annotation is done by assigning biological information to the genomic data (Ekblom and Wolf, 2014). *De novo* assembly is the construction of a continuous sequence without the availability of a reference genome. There are many factors to consider before sequencing a whole genome, however, two of the most important are the length of reads and sequence coverage. These decisions depend highly on the budget for the genomic project and the resources available.

Short reads made their appearance as the "second generation” of sequencing between 2005 and 2010. Using high-throughput systems, like Illumina, we can obtain short sequence reads that can be joined to other short reads via paired-ends. Through this, contigs are built and eventually merge into longer scaffolds. One drawback is that by joining short reads only, you can run into problems with repetitive regions and gaps. They are also lacking in the ability to detect large structural variants (SVs). Between 2011-2014, long reads made their debut as the "third generation" of sequencing technologies. Long reads produce much longer continuous sequences than typical short read technologies. This is done in one of two ways: single-molecule real time (SMRT) sequencing and synthetic long read sequencing. The SMRT method produces long reads of a single DNA molecule and is primarily being done by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies. Synthetic long reads (SLR-seq) are made by using a computational program to assemble barcoded short reads to make the longer read, performed by Illumina. The SMRT method is, currently, the more popular option in long read sequencing. PacBio's SMRT cell services are expensive and increase in price with the length of the genome. Many research teams are not able to use this resource due to cost limitations. Additionally, the Nanopore technology has shown to be more sensitive to error, making it harder to use those long reads effectively. When it is possible to obtain high enough coverage and low error, long reads are used to assemble complex regions, repetitive stretches, and structural variants. Long reads are not mandatory for assembly, but they do improve the overall quality and comprehensiveness of the genome. It is possible to do short reads, long reads, or a combination of long and short which typically results in the highest genome quality. It is expected that the price of long read sequencing technology will improve as more technological advancements are made, becoming more available to research groups around the world.

Coverage, or the read depth, is the average number of reads per locus. The level of coverage is most important in downstream applications. Increasing the sequence’s coverage results in a higher confidence in the base composition. It is recommended that regardless of read length, greater that 50-60x coverage should be applied throughout the genome (Fuentos-Pardo and Ruzzante, 2017). This is especially important when trying to call variants within a population with a high degree of confidence in that base call. To publish a whole genome as a reference, there are certain quality guidelines that journals require. It is essential to understand N50 and L50 statistics. N50 is a statistic that describes a set of contigs or scaffolds and measures the continuity of the assembly. It is defined as an estimation of a contig/scaffold length such that any sequence equal to or longer accounts for at least 50% of the bases in the assembly. L50 is the actual number of contigs and scaffolds that are equal to or greater than the N50 value. For example, the Journal of Heredity explains in their author guidelines that the minimum assembly requirements are contigN50 > 200Kb and scaffoldN50 > 5Mb.

**Application to Conservation**

Through whole genome resequencing methods, such as low coverage sequencing and pool-seq, many conservation questions can be addressed. Historically, neutral markers have been used to investigate genetic variation. With genome-wide data, there is a new ability to study loci under selection and reveal adaptive potential of a population. This is where having a high-quality reference genome becomes important. Any analysis of the genomic data, such as identifying SNPs, requires a high confidence that the call is correct. While there are endless possibilities for the application of a whole genome, my objective is to focus on signatures of selection and adaptive regions of the genome. There are two strategies, for understanding loci under selection, that I will discuss: forward genetics and reverse genetics.

**Forward Genetics**

The forward genetics = you already have the phenotype, but you want to know the underlying genomic region.

Use genome-wide association study (GWAS) and quantitative trait loci (QTL) mapping. Both identify association between specific phenotypic traits and fitness.

GWAS = uses linkage disequilibrium. The strength and accuracy of the association study depends on phenotypic variance of the given population. This is determined by effect size and frequency in a sample. Therefore, GWAS is more efficient when studying large-effect loci. The need for detecting large-effect loci can be reduced by using very large sample size. This is typically not an option when working with conservation biology, as most populations are smaller or more likely on a downward trend.

Ex: mimicry supergene in swallowtail butterfly (*Papilio polytes)* 🡪used association mapping to find that a single gene *doublesex* is at the functional level of this adaptive phenotype.

QTL mapping = is another method in forward genetics that relates the adaptive trait to its genomic basis.

**Reverse Genetics**

**Conclusion**

**Review article = read x number of papers, methodical way of collecting information “ I reviewed 20 articles, this is where I got it from” set goals and objectives of the review. Need to use tables and figures… probably a summary table with sources in that. Think of it as a meta-analysis. Harvest info🡪 summary or analysis or look at trends-🡪 show some gaps 🡪 create table and figures**

**Literature Review= read paper, make some notes (like annotated bibliography)🡪 less organized and less focused. Just introduction to what is already in the literature**

**Think about scale and scope of review. Do not go too big. Don’t set the scope too large… narrow it to 20-25 papers. Get specific about objectives. Make table be harmonious with the objectives.**

**Looking for pattern and trends between all of the works… identify**

**Let my own situation dictate formatting**

**What is the scope? How wide a net am I casting? Don’t cast too wide.**

**Review paper should self-organize**

**Intro**

**Methods section that goes over how we selected articles.**

**Talk about what you systematically pull out**

**Make sense of that 🡪 pulled info, set an objective, how to organize and highlight patterns or gaps**

**IMRaD format**

**At least 12 articles cited in paper, 10 pages of writing**

**For a review article, generally about twice that… 20-25 papers for this class and thesis. Start with the idea of a couple dozen and go with that.**

**Ex: I want to review papers that reveals alleles under selection… less taxonomic**

**We all live downstream: applications from whole genome sequencing within adaptive loci**