1. **Introduction**

* There are many organizations whose primary goal is to assemble novel genomes. The Earth Biogenome Project (EBP), Global Invertebrate Genomics Alliance (GIGA), and the Genome 10K Project (G10K) to name just a few.
* Define whole genomic sequence as a hypothesis. Explain that a genome is only of use with the proper downstream applications.
* Explain briefly the importance of genomics in conservation and population genomics.
* Make a clear definition of genomics vs genetics.

1. ***De novo* genome sequencing**

There are three main steps to reach a high-quality whole genome sequence. 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. 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 whole 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.

1. **Adaptive Regions of the Genome**

Through whole genome resequencing methods, such as low coverage sequencing and pool-seq, many conservation questions can be addressed. With genome-wide data, there is new potential 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 variant call is correct and indicative of diversity. 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.

* Identification of signatures of selection often depend on reduced nucleotide diversity (pi), linkage disequilibrium, and high homozygosis (WGSconservation). It is important, especially when looking at nucleotide diversity, to create a null model which accounts for diversity under natural processes. (Ellegren).
* Maybe mention the concept of genetic architecture behind adaptive loci.
* Reverse genetics suggests that the study first aims to find signatures of selection within the genome, then look to define the adaptive phenotypic trait expressed by that region. Reverse genetics methods are more efficient in natural populations and can detect small, medium, and large effect loci. These strategies typically include genome scans and genome-environment association (GEA) analysis. Genome scans will be discussed here and GEA is discussed in the “Landscape Genomics” section. Typically, whole genome resequencing for non-model organisms is classified under reverse genetics because we often do not know what adaptive phenotype we are looking for (WGSconservation).

**Genome Scans**

* Uses a window-based approach for scanning.
* Genetic hitchhiking = When a locus undergoes selection, linked loci will also be affected, which can lead to fixation of neutral loci. The closer the two loci are to each other, the larger the effect. This results in a loss of genetic diversity in areas around the site of adaptation.
* Selective sweep
  + Hard sweeps = mutations that arise and quickly progress into fixation. Leave more distinct footprints in the genome and stronger signals of selection. Large effect.
  + Soft sweeps = naturally selected allele was already present in the population diversity prior to sweep. Much weaker signal of selection and can blend in with background selection. Multiple small effect.
* Locations of selective sweeps can be detected using linkage disequilibrium. Linkage disequilibrium is the non-random association between alleles at multiple loci. In an area under selection, LD will be increased.
* Over time, signals of selection will weaken and be much less detectable.

1. **Landscape Genomics**

Landscape genomics offers, “greater power to disentangle adaptive from neutral genetic

divergence and identify environmental factors acting as selective agents,” (Grummer, et al. 2019). There is an important relationship to be seen between landscape genomics and conservation efforts. Prediction can be made about the adaptive potential and future health of a population, which is a major influence on what measures need to be taken and to what extent.

**Environmental Association Analysis**

* Two types of data: environmental factors and genetic polymorphisms

1. **Conservation Measures**

Conservation genomics forms a relationship between a population’s genetic diversity and persistence in local environments. Development of genomic tools allows for more efficient management of endangered species by identifying specific environmental stressors. “Genomics has the potential to dramatically improve our ability to identify, monitor and predict the effects of stressors, as well as to mitigate their impacts through the use of marker-assisted selection…” (The Bee Paper)

**Candidate Genes for Conservation Monitoring**

**Species-Specific Microsatellite Markers**

* One important thing I have run into in almost every paper is the ability to use whole genome information to find an abundance of species-specific microsatellite markers that can be more effectively and inexpensively produced. These markers can be made to be conservation specific and be an extremely useful management tool.

**Treatment Development**

**Breeding Program**

* Use of large-scale breeding programs that select for resistance to stressors and can result in the production of a more sustainable population.
* The large scale is used to lessen the concern for “genetic dilution”. (Bee)

1. **Conclusion**