**NOTE: \*\*\* I have mainly been reading over review papers to get acquainted with terminology and existing themes, I will now focus on finding literature that examines specific cases of genomics used in conservation, like the Scandinavian Wolf Paper. While looking through this, keep in mind that I will go back and plug real world examples into the material that I outline here. This outline is basically me trying to set up a flow of subjects that I want to hit on\*\*\***

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.

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

* Touch on the importance of a reference genome (briefly explain whole genome resequencing importance too). The construction of a reference genome is expensive; however, it is paid off in the downstream applications.
* Define *de novo* genome sequences when a reference is not available.

* **Sequencing:** this is the done using high quality extracted DNA and high-throughput technology like Illumina.
  + **Read Length and Platforms**: 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.
* **Assembly** is the computational process of constructing longer sequences through piecing together shorter sequence reads.
  + Discuss some software that is used to assemble and trim the sequences. SOAPdenovo, GapCloser, ConDiTri = (Ekblom, et al. 2018)
  + To publish a whole genome as a reference, there are certain quality guidelines that journals require. It is essential to understand statistics like N50 and L50. 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.
* **Annotation** is done by assigning biological information to the genomic data (Ekblom and Wolf, 2014).
  + Give a couple of examples, but really let this part lead into the rest of the paper since the remaining parts are all annotation.

\*\*Most of the technical information will come from (Ekblom and Wolf, 2014), (Fuentes-Pardo and Ruzzante, 2017)

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.

* **Adaptive Loci and Signatures of Selection. Adaptive Potential.**
  + Adaptive potential is a major concern when considering a population’s health, whether they can respond to a change in environment or not. Many conservation organizations use adaptive potential when making listing or delisting decisions. -Population Genomics for Conservation). High adaptive potential (the ability to change according to the environment) is in part determined by the amount of genetic diversity in a population.
  + No longer requires controlled breeding or common garden experiments. – Population Genetics for Conservation
  + 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 adaptive regions of the genome related to conservation concerns.
  + Maybe mention the concept of genetic architecture behind adaptive loci.
  + Forward genetics focuses on a specific phenotype and uncovers the genomic foundation. Reverse genetics suggests that the study first finds signatures of selection within the genome, then looks to define the adaptive phenotypic trait expressed by that region. This typically reveals environmental stressors. Reverse genetics methods are more efficient in natural populations and can detect small, and large effect loci. These strategies typically include genome scans and genome-environment association (GEA) analysis. 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 ((Fuentes-Pardo and Ruzzante, 2017)
* **Genome Scans**
  + Uses a window-based approach for scanning.
  + Identification of signatures of selection often depend on reduced nucleotide diversity (pi), linkage disequilibrium, and high homozygosis (Fuentes-Pardo and Ruzzante, 2017). It is important, especially when looking at nucleotide diversity, to create a null model which accounts for diversity under natural processes. (Ellegren,2014).
  + 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.
  + Problem: Over time, signals of selection will weaken and be much less detectable.
* **Genome-Environment Association (GEA) Analysis**
  + This studies adaptive traits to the environment *a priori,* which means that the phenotype is not identified beforehand. It measures how an environment affecting a population. It also allows us to view genomic differences between populations and relate that to environmental differences.
  + Two types of data: environmental factors and genetic polymorphisms. For this analysis, measure environmental factors and polymorphic data independently, then use a combination to link genome to environment.

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.
* “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…” (Grozinger and Zayed, 2020)
* Once a high-quality reference genome is assembled and annotated, low quality DNA from non-invasive collection (scat, hair, eDNA, museum samples) can be used. It is then mapped against high quality genome. (Hohenlohe, et al. 2020)
* Identification of conservation units. (Hohenlohe, et al. 2020)
* **Genetic Monitoring**
  + 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.
  + Example: Scandinavian wolf: 96 markers were found useful for genetic monitoring (Ekblom, et al. 2018)
  + Markers can be used to measure the general genetic diversity within a population, but can also be made more specific to genes that are of interest to conservation (resistant genes in Tasmanian devils). (Brandies, et al. 2019)
* **Controlled Breeding Programs, Captive Populations, Translocation**
  + Use of large-scale breeding programs that select for resistance to stressors and can result in the production of a more sustainable population. (Grozinger and Zayed, 2020)
  + Insurance population maintenance of Tasmanian devil to release back into captivity to maintain a high enough level of population genetic diversity. (Brandies, et al. 2019)
  + Use specific microsatellite markers (discussed above) that a good amount of genetic diversity (especially related to known environmental stressors) is maintained in case of reintroduction.
* **Mitigating stressors**
  + **Disease stressors:** We can use genomic information to create treatment plans or vaccinations for specific diseases that affecting a population. Tasmanian devil. (Brandies, et al. 2019)
  + **Environment stressors:** Use EAA to understand underlying environmental concerns and steer conservation efforts to mitigate those effects. This could even be dealing with things like habitat loss or runoff into aquatic habitats. Bee population.
  + **Find information about more kinds of stressors with examples.**

1. **Conclusion**

Go back to broader discussion on conservation genomics.

**I will make a chart to put all of the studies like this:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Organism** | **Type of reads (short, long, hybrid)** | **Resequencing method** | **Conservation recommendation** | **Source** |
| **Study A** |  |  |  |  |
| **Study B** |  |  |  |  |

**I may also make a table that lists the softwares used in the sequencing, assembly, annotation process:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Organism** | **Type of reads (short, long, hybrid)** | **Assembly software** | **Annotation and Quality Control software** | **Source** |

**These tables will really depend on the accumulation of studies that I look at and what kind of information is offered in them.**