**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**

**See how people are preserving specimens for high quality DNA extraction🡪 reach out to PacBio**

**Try gill, leg, and abdomen tissue for extraction**

**The rest can be in liquid nitrogen and store in -80 C**

**I think I will delete:**

**Forward Genetics (“Top-Down Approach”) 🡪 keep discussion brief**

Forward genetics is a popular strategy for finding the genetic basis of an already known phenotype. For example, the Tasmanian devil is a conservation case that many people use to study the methodology of conservation genomics. The Tasmanian devil was severely endangered due to a transmissible facial cancer, DFDT. Some individuals displayed a resistance to DFDT and even showed evidence of tumor regression. Using genome-wide association studies, they were able to locate the loci responsible for that resistant phenotype and use that information for implementing conservation strategies and potential treatment. Genome-wide association studies (GWAS) and quantitative trait loci (QTL) mapping and two tools often used in forward genetics approaches. GWAS is used mainly for extreme phenotypic variation and large effect loci.

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= Not as used not and not suitable for fisheries = is another method in forward genetics that relates the adaptive trait to its genomic basis. Many times, RAD-seq is used. Restriction is “not obtaining enough recombinant offspring…due to the nature of many organisms”. Not very suitable for small sample sizes = ‘Beavis effect’. Cannot detect small and medium effect size.

Add examples of using the genome in forward genetics, and cases of conservation.

Outline:

Overall topic: How can genomics advance conservation measures

Sub topics:

de novo sequencing with endangered populations

* Importance of having a reference genome
* Sequencing
* Assembly
* Annotation 🡪 let annotation lead into next part.

Landscape genomics

* Population genomics + landscape genetics
* Explanation of adaptive regions in the genome (maybe discuss pos vs neg)
* Genome scans
* Genome-Environment Association

Conservation Measures

* Candidate Genes for Monitoring
* Species-specific microsatellite markers
* Treatment development
* Breeding Programs

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.

The Importance of a Reference Genome

* Discussion about the availability of a reference genome and what kinds of applications that can lead to. Then explain the need for *de novo* sequencing if that reference is not available.
* Sequencing
  + Get tissue minimally invasive
  + Need high weight molecular DNA extracted
  + Library Prep
  + Read depth (coverage)
  + Length of reads
* Assembly – computational process of connecting reads.
  + Statistics like N50 or L50.
  + Touch on a few recommended software and commom problems
  + Look at Scandinavian wolf paper to talk about software to mitigate repeat concern and stuff like that
* Annotation – Assigning biological information to the genomic data.
  + Give a couple of examples, but really let this part lead into the rest of the paper since the remaining parts are all annotation.

Landscape Genomics

* discussion about adaptive regions of the genome and loci under selection.
* landscape genomics as the relationship between environment and adaptive regions of the genome. Population genomics + landscape genetics.
* Positive vs negative genetics (VERY brief… really just a preface for further discussion)
* Genome scans
* Genome-Environment Association (GEA) analysis – this will be the longest subsection

Specific Conservation Measures

* Species specific microsatellite markers (genetic monitoring)
* Breeding programs
* Mitigating stressors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | De novo seq | Genome Scans | GEA analysis | Specific conservation |
| A | A |  |  |  |
| B |  | B |  |  |
| C | C | C |  |  |
| D |  | D | D | (candidate gene) |
| E |  |  |  |  |
| F |  | F | F | F |
| G |  | G | G | G |
| H |  |  |  |  |
| I | I |  |  | I |
| J | J | J | J | J |
| K | K |  |  | K |

A = A field guide to whole-genome sequencing, assembly

and annotation, Robert Ekblom and Jochen B. W. Wolf

B = A practical guide to environmental association analysis

in landscape genomics, CHRISTIAN RELLSTAB,\* FELIX GUGERLI , \* ANDREW J . ECKERT,† ANGELA M. HANCOCK‡ and ROLF HOLDEREGGER\*§

C = Genome sequencing and population

genomics in non-model organisms, Hans Ellegren

D = Towards the identification of the loci of adaptive

Evolution, Carolina Pardo-Diaz1\*, Camilo Salazar1 and Chris D. Jiggins2

E = Genomics in Conservation: Case Studies and Bridging the Gap between Data and Application

Brittany A. Garner,‡, Brian K. Hand,,‡,\* Stephen J. Amish, Louis Bernatchez, Jeffrey T.

Foster, Kristina M. Miller, Phillip A. Morin, Shawn R. Narum, Stephen J. O’Brien, Gretchen

Roffler, William D. Templin, Paul Sunnucks, Jeffrey Strait, Kenneth I. Warheit, Todd R., Seamons, John Wenburg, Jeffrey Olsen, and Gordon Luikart

F = Improving bee health through

Genomics, Christina M. Grozinger 1 and Amro Zayed

G = Aquatic Landscape Genomics and Environmental Effects on Genetic Variation

Grummer, Jared A.\*1, Beheregaray, Luciano B.2, Bernatchez, Louis3, Hand, Brian K.4, Luikart,

Gordon5, Narum, Shawn R.6,7, Taylor, Eric B.

H = Population genomics for wildlife conservation and

Management, Paul A. Hohenlohe1 | W. Chris Funk2 | Om P. Rajora3

I = The Value of Reference Genomes in the Conservation

of Threatened Species

Parice Brandies, Emma Peel, Carolyn J. Hogg and Katherine Belov \*

J = Whole-genome sequencing approaches for conservation

biology: Advantages, limitations and practical

recommendations

Angela P. Fuentes-Pardo | Daniel E. Ruzzante

K = Genome sequencing and conservation genomics in

the Scandinavian wolverine population

Robert Ekblom ,1 ∗ Birte Brechlin,1 Jens Persson,2 Linn´ea Smeds,1 Malin Johansson,1

Jessica Magnusson,1 Øystein Flagstad,3 and Hans Ellegren