INTRO: Advancement of Sequencing Technologies

Rapid development of high-throughput sequencing technologies over the past few decades has led to an era of genome-centric research through the ability to assemble high-quality chromosome-level reference genomes across the tree of life. These genomic resources contribute to two broad categories: medicine development and the preservation of biodiversity. Medicine leverages access to highly contiguous genomes using comparative methods to identify conserved loci essential to life or different classes of organisms, understanding individual effects of medicinal treatments, as well as identify genetic variants associated with disease, disease susceptibility and other phenotypic traits (Claussnitzer et al., 2020). For wild populations genetic diversity is related to the evolutionary capacity for adaptation to environmental change and is an essential element of biodiversity to which protections focus on preserving. Genomic resources allow for a broad investigation of genomic motifs in endangered animals (Zoonomia Consortium, 2020), the identification of physical threats to endangered species (Wasser et al., 2015; Wright et al., 2015) and identification of vulnerable populations susceptible to increasingly variable climatic conditions (TKT INSERT RACHEL BAY PAPER). However, genomic studies involving reference genomes are limited by the completeness of the assembled resource.

Hybrid genome assembly uses multiple sequencing technologies to assemble genomes through capitalizing on the various strengths of each technology. This combination of creating one assembly from multiple sequencing technologies has led to more contiguous and accurate assemblies [TKTK], and as a result has become the gold standard for *de novo* genome assembly. Hybrid assemblies may use any combination of long-read sequencing, short-read sequencing, optical mapping, and interaction mapping. Each technology type has different biases, errors, costs, and uses. In general, hybrid assemblies use long error prone reads to generate scaffolds, correct base calling errors with short reads which have high accuracy but cannot span highly repetitive sequences, and anchor scaffolds into chromosomes using interaction mapping which shows physical associations to span and link proximal scaffolds. Biases, errors and pricing are consistently being minimized as sequencing research continues to generate more, longer, and increasingly accurate reads from a single sequencing machine.

Next generation sequencing (NGS) and third generation sequencing (TGS) technologies have made it relatively easy to generate low cost, high-throughput sequencing data. Of the numerous new methods to obtain sequencing data for assembly, long-read sequencing, short-read sequencing, and interaction mapping have transformed the quality and contiguity of genome assemblies today. Current *de novo* genomes require collaboration between relatively few scientists and reach completeness standards which took decades for the first human genome to reach at a fraction of the cost and by the collaboration relatively few individuals (International Human Genome Sequencing Consortium et al., 2001). A recent example showcasing the power of hybrid *de novo* assembly was the sequencing and assembly of the domestic goat genome in 2017. Through the use of high-quality DNA extraction and combining recently stablished sequencing technologies the (Bickhart et al., 2017) used long-read sequencing, short-read sequencing, and interaction mapping to produce a *de novo* assembly that was over two orders of magnitude more contiguous than the previously published goat. The paper also proposed it was the most continuous *de novo* mammalian assembly of its time. Since the publication of the *de novo* goat assembly, hybrid assembly publications are commonplace. As such, hybrid assembly is an accepted and reliable way to achieve a chromosome-scale high-quality reference genome (Bickhart et al., 2017; Rhie et al., 2021). Since 2017, over half of all vertebrate chromosome-level assemblies submitted to GenBank implemented a hybrid assembly approach to genome assembly (Hotaling et al., 2021).

Chapter 1 – Genome

We used PacBio HiFi long-reads, 10X Genomics linked-reads, and Phase Genomics chromatin confirmation capture to assemble one male genome and one female genome. In addition to these three sequencing technologies, we sought to further increase the contiguity of the genome by incorporating information from a previously published linkage map (Lew et al., 2015). Finally, because species with different numbers of chromosomes can hybridize and it is unclear if progeny from a wakasagi and delta smelt cross are fertile, we independently validated the number of expected chromosomes in the final delta smelt assembly by karyotyping the species.

evolutionary genomic research.

OLD INTRO

Rapid development of high-throughput sequencing technologies over the past few decades has led to an era of genome-centric research through the ability to assemble high-quality chromosome-level reference genomes across the tree of life. These genomic resources contribute to two broad categories: medicine development and the preservation of biodiversity. Medicine leverages access to highly contiguous genomes using comparative methods to identify conserved loci essential to life or different classes of organisms, understanding individual effects of medicinal treatments, as well as identify genetic variants associated with disease, disease susceptibility and other phenotypic traits (Claussnitzer et al., 2020). For wild populations genetic diversity is related to the evolutionary capacity for adaptation to environmental change and is an essential element of biodiversity to which protections focus on preserving. Genomic resources allow for a broad investigation of genomic motifs in endangered animals (Zoonomia Consortium, 2020), the identification of physical threats to endangered species (Wasser et al., 2015; Wright et al., 2015) and identification of vulnerable populations susceptible to increasingly variable climatic conditions (TKT INSERT RACHEL BAY PAPER). However, genomic studies involving reference genomes are limited by the completeness of the assembled resource.

Hybrid genome assembly uses multiple sequencing technologies to assemble genomes through capitalizing on the various strengths of each technology. This combination of creating one assembly from multiple sequencing technologies has led to more contiguous and accurate assemblies [TKTK], and as a result has become the gold standard for *de novo* genome assembly. Hybrid assemblies may use any combination of long-read sequencing, short-read sequencing, optical mapping, and interaction mapping. Each technology type has different biases, errors, costs, and uses. In general, hybrid assemblies use long error prone reads to generate scaffolds, correct base calling errors with short reads which have high accuracy but cannot span highly repetitive sequences, and anchor scaffolds into chromosomes using interaction mapping which shows physical associations to span and link proximal scaffolds. Biases, errors and pricing are consistently being minimized as sequencing research continues to generate more, longer, and increasingly accurate reads from a single sequencing machine.

Next generation sequencing (NGS) and third generation sequencing (TGS) technologies have made it relatively easy to generate low cost, high-throughput sequencing data. Of the numerous new methods to obtain sequencing data for assembly, long-read sequencing, short-read sequencing, and interaction mapping have transformed the quality and contiguity of genome assemblies today. Current *de novo* genomes require collaboration between relatively few scientists and reach completeness standards which took decades for the first human genome to reach at a fraction of the cost and by the collaboration relatively few individuals (International Human Genome Sequencing Consortium et al., 2001). A recent example showcasing the power of hybrid *de novo* assembly was the sequencing and assembly of the domestic goat genome in 2017. Through the use of high-quality DNA extraction and combining recently stablished sequencing technologies the Bickhart et al., (2017) used long-read sequencing, short-read sequencing, and interaction mapping to produce a *de novo* assembly that was over two orders of magnitude more contiguous than the previously published goat. The paper also proposed it was the most continuous *de novo* mammalian assembly of its time. Since the publication of the *de novo* goat assembly, hybrid assembly publications are commonplace. As such, hybrid assembly is an accepted and reliable way to achieve a chromosome-scale high-quality reference genome (Bickhart et al., 2017; Rhie et al., 2021). Since 2017, over half of all vertebrate chromosome-level assemblies submitted to GenBank implemented a hybrid assembly approach to genome assembly (Hotaling et al., 2021).

In our assembly of a male and female delta smelt genome, we used PacBio HiFi long-reads, 10X Genomics linked-reads, and Phase Genomics chromatin confirmation capture. In addition to these three sequencing technologies, we sought to further increase the contiguity of the genome by incorporating information from a previously published linkage map (Lew et al., 2015). Finally, we independently validated the number of expected chromosomes in the final delta smelt assembly by karyotyping the species.

Prior to this work, only three highly fragmented genome assemblies from the Osmeridae(smelt) family were publicly available through GenBank and none of these was from endangered species. These assemblies came from three different genera (*Hypomesus, Thaleichthys*, and *Osmerus*) and the most closely related assembly publicly available was *Hypomesus nipponensis* (wakasagi smelt), a common species of smelt endemic to and used as a food commodity in Japan, introduced into the SFE and known to hybridize with delta smelt (Dill & Cordone, n.d.). While the wakasagi smelt genome, estimated to be 464 Mbp in size with 2n=26, was the most contiguous resource within the Smelt family, the assembly was not a chromosome-level assembly with an N50 of 0.46 Mbp and L50 of 477 (Table 2.1, Kitada et al., 1980; Xuan et al., 2021). Because of their listing status and the continuous genetic research activity of delta smelt, a highly contiguous reference genome was a necessary next step in investigating migration and genetic diversity in the wild population and domestication in the refuge colony.