De novo assembly and comparative genomics of teleosts

OLE KRISTIAN TØRRESEN

Dissertation presented for the degree of Philosophiae Doctor
(PhD)
2017

Centre for Ecological and Evolutionary Synthesis
Department of Biosciences
The Faculty of Mathematics and Natural Sciences
University of Oslo

© Ole Kristian Tørresen, 2017

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1878

ISSN 1501-7710

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

It's the questions we can't answer that teach us the most. They teach us how to think. If you give a man an answer, all he gains is a little fact. But give him a question and he'll look for his own answers.

– Patrick Rothfuss, *The Wise Man's Fear*

Getting an education was a bit like a communicable sexual disease. It made you unsuitable for a lot of jobs and then you had the urge to pass it on.

– Terry Pratchett, *Hogfather*

Acknowledgment

Almost by definition, an acknowledgment section will be incomplete. Inspiration, motivation and ideas come from a multitude of sources, and it is unrealistic to cover them all properly. I'll do my best.

My main supervisor has been Kjetill S. Jakobsen. Even though he sometimes misspells my last name, he is a well of knowledge in all kinds of areas and really supportive.

Another supervisor has been Lex Nederbragt, who introduced me to genome assembly and is always available to discuss all things in bioinformatics and life in sciences.

The third supervisor, Geir Kjetil Sandve, has helped me understand graph traversing and implementing it in Python.

Last supervisor, Sissel Jentoft, is always dragging me into new projects, and has apparently enough faith in me to think that I can handle it all.

I would like to thank my friends and co-workers for their support, discussions and fun: Martin, Monica, Ave, Even, Srinidhi, Tore, Kenneth, Kent, Cassie, Jo, Jostein, Kjetil, Mark, Øystein, William, Jonfinn, Anders, Elin, Helle, Rebekah, Marine, Bastiaan, Micha and Julia, and all others that I might have forgotten when writing this.

I don't have to forget to thank the high performing computing cluster Abel (previously Titan) and the people that built, maintain and support it, for all their help and computing power. I only have records for the last 2 years, but I used 7 million CPU hours in that time.

I would like to thank Mark Ravinet, William B. Reinar, Martin Malmstrøm, Ragnhild Røysland and Michael Matschiner for critical reading of this thesis. Unfortunately, I have to blame myself for any errors and not them.

I especially have to thank Anders (who has a boy born the same day, Christmas Eve, as my daughter) and Martin for extracurricular activities in swimming and beer drinking (not at the same time!), and others that have joined from time to time.

I want to thank all my co-authors, especially Brian Walenz and Jason Miller for their discussions about genome assembly and Rolf Edvardsen for good comments and guidance.

My parents have always been supportive, for which I thank them, even though they might not properly understand what I'm doing. I'm grateful to my brothers, Lars Even, Ingar and Lars Jørgen, for baby-sitting, drinking companionship and demolishing. My thanks also go to my extended family, for

participating in the same activities.

Most importantly, I thank my son, Nils Olav, my daughter, Kjersti, and my girlfriend and partner Ragnhild, for all their support and love. We have been through a lot these years. Moving twice, renovations, and two PhDs. No matter what, life will be simpler and less hectic after the PhD.

List of works

Work I

Ole K. Tørresen, Bastiaan Star, Sissel Jentoft, Kjetill S. Jakobsen, Alexander J. Nederbragt. **The new era of genome sequencing using high-throughput sequencing technology: generation of the first version of the Atlantic cod genome.** In *Genomics in Aquaculture*, edited by Simon MacKenzie and Sissel Jentoft. Cambridge, Massachusetts: Academic Press, 2016

Work II

Ole K. Tørresen, Bastiaan Star, Sissel Jentoft, William B. Reinar, Harald Grove, Jason R. Miller, Brian P. Walenz, James Knight, Jenny M. Ekholm, Paul Peluso, Rolf B. Edvardsen, Ave Tooming-Klunderud, Morten Skage, Sigbjørn Lien, Kjetill S. Jakobsen, Alexander J. Nederbragt. 2017. **An improved genome assembly uncovers prolific tandem repeats in Atlantic cod.** *BMC Genomics.* 18:95.

Work III

Martin Malmstrøm, Michael Matschiner, Ole K. Tørresen, Kjetill S. Jakobsen, Sissel Jentoft. 2017. **Whole genome sequencing data and *de novo* draft assemblies for 66 teleost species.** *Scientific Data.* 4:160132.

Work IV

Martin Malmstrøm, Michael Matschiner, Ole K. Tørresen, Bastiaan Star, Lars G. Snipen, Thomas F. Hansen, Helle T. Baalsrud, Alexander J. Nederbragt, Reinhold Hanel, Walter Salzburger, Nils C. Stenseth, Kjetill S. Jakobsen, Sissel Jentoft. 2016. **Evolution of the immune system influences speciation rates in teleost fishes.** *Nature Genetics.* 48, 1204–1210.

Work V

Ole K. Tørresen, Marine S. O. Brieuc, Monica H. Solbakken, Tomasz Furmanek, Elin Sørhus, Alexander J. Nederbragt, Kjetill S. Jakobsen, Sonnich Meier, Rolf B. Edvardsen, Sissel Jentoft. **Genomic architecture of codfishes featured by expansions of innate immune genes and short tandem repeats.** Manuscript.

Thesis summary

During the last 20 years, genome sequencing and assembly projects have changed from requiring large international collaborations to a task that a handful of people can plan and conduct. This has been driven by improvements in sequencing technology and computational methods. More and more sequencing and assembly projects are being conducted, with older assemblies being updated and improved, resulting in deeper understanding of the biology of a large and steadily growing number of species. The projects described in this thesis focus on genome assemblies created from species of the order Gadiformes, an order containing commercially and ecologically important fishes. Here, these assemblies are investigated in detail and compared to other teleost genome assemblies, with special attention to immune genes and short tandem repeats.

We have updated and substantially improved the Atlantic cod (*Gadus morhua*) genome assembly with the use of different sequencing technologies and computational approaches. A major finding was that the presence of short tandem repeats (STRs) is the main factor that led to the fragmentation of the previous assembly. STRs are hypermutating loci that occur at high frequency (loci/Mbp) and high density (bp/Mbp) in the cod genome, surpassing that of other published genome assemblies. The STRs likely contribute to substantial genetic variation in natural cod populations.

The Atlantic cod lacks genes involved in the major histocompatibility complex (MHC) II pathway, which is the pathway that normally detects and initiates a response against bacterial pathogens and thus is a crucial part of the adaptive immune system. To infer when in the ancestry of cod these genes were lost, we sequenced and assembled the genomes of 66 teleost species. We found that the loss is shared by all species in the order Gadiformes, and that there is an expanded repertoire of *MHCI* genes in the Gadiformes, which is likely connected with the large number of species in this order.

Since the 66 new teleost (including gadiform) genome assemblies are fragmented, the properties of STRs and multi-copy immune genes are not easily investigated. To further elucidate their role in Gadiformes, we sequenced and assembled the genome of haddock (*Melanogrammus aeglefinus*), a relative of cod. Our result shows that the high density and frequency of STRs is a feature likely shared by all codfishes (a family inside Gadiformes), and possibly all Gadiformes. Cod and haddock share a similar repertoire of the innate immune Toll-like receptor (*TLR*) genes, with both losses and expansions. The expan-

sions might be part of a compensatory mechanism for the absence of MHCII. Another class of genes, the NOD-like receptors (*NLRs*) has been reported in large numbers in species without an adaptive immune system. We find that cod and haddock as well as most other teleosts generally have a high number of NLRs, with a likely expansion at the root of this clade. Thus, a high number of NLRs in teleosts does not seem to be connected with the presence or absence of *MHCII*.

This thesis shows what kind of questions genome assemblies created for different purposes can answer. Ideally, genome assemblies for all kinds of species should be created, upgraded and updated based on the best available technologies. But this is costly. With the right planning and set-up, assemblies based on low-coverage sequencing can be very powerful with regards to topics such as the presence/absence of genes and for phylogeny. Also, even with moderate amounts of long-read PacBio sequencing, it is possible to create highly contiguous genome assemblies addressing issues that are impossible to elucidate with fragmented assemblies, such as the amount of multi-copy immune genes.

Contents

1 Introduction

Access to a genome assembly for a species of interest can contribute to knowledge in numerous biological aspects, including ecology (Ekblom and Wolf, 2014), evolutionary biology and speciation (Ellegren, 2014) and evolutionary developmental genomics (Braasch *et al.*, 2015). The first vertebrate genome sequencing and assembly project, i.e. human, was a huge and expensive endeavor, spanning more than a decade (International Human Genome Sequencing Consortium, 2004). Since then, with the advent of high-throughput sequencing (HTS), first with 454 (Margulies *et al.*, 2005) and Illumina (Bentley *et al.*, 2008), and later PacBio (Eid *et al.*, 2009), sequencing costs have decreased drastically (reviewed in Goodwin *et al.* (2016)), making genome sequencing and assembly feasible for even small research groups (Ekblom and Wolf, 2014).

1.1 Generating an annotated genome assembly

1.1.1 The underlying algorithms of genome assembly software

A genome assembly is a putative reconstruction of a genome based on information found in sequencing reads and possibly other sources of information, such as linkage maps (see below). There are two major approaches for genome assembly, the de Bruijn graph and overlap/layout/consensus (OLC) methods. The OLC approach was first implemented in Celera Assembler and used to assemble the Drosophila genome in 2000 (Myers *et al.*, 2000). This approach works by first detecting overlap between all sequencing reads, creating a graph based on the overlaps, simplifying and traversing the graph before outputting so-called unitigs (sequences that are either unique in the genome or are collapsed repeated sequence) based on a multiple sequence alignment from the overlaps (Miller *et al.*, 2010). Because the overlap step compares each read to all other reads, computational demand can be high, but it is reduced with fewer but longer reads because of fewer overlaps computed. The overlap step can also tolerate mismatches and indels (insertions and deletions between pairs of chromosomes in diploid organisms) between the reads, and therefore performs well with longer reads even if these are error-prone. The unitigs are further processed (categorized into unique and repeat unitigs), before they are scaffolded based on information from paired reads, outputting contigs (contiguous sequence based on consensus sequence from the reads) and their order and orientation into scaffolds (Figure 1.1). The String Graph Assembler (SGA) (Simpson and Durbin, 2011) is an alternative implementation of the overlap

graph, where a string graph is derived from the overlap graph. The main advantage of SGA compared to Celera Assembler is that it uses an efficient string indexing data structure, requiring less amount of memory when creating the string graph. The de Bruijn graph method, as implemented in assemblers such as ALLPATHS-LG (Gnerre *et al.*, 2011) and Velvet (Zerbino and Birney, 2008), sidesteps the computationally expensive explicit overlap step, and creates a graph where each node represents a fixed-length sequence (k-mer) found in the reads, and the edges connect to k-mers with k-1 sequence in common (which can be found in multiple reads) (Zerbino and Birney, 2008). This graph is then parsed, and contigs are output. These contigs can be scaffolded in much the same way as for the OLC method (Figure 1.1).

Reads generated for sequencing projects differ in their attributes. Illumina (Bentley *et al.*, 2008) reads are relatively short (100-250 bp), but contain few errors ($\geq$0.1 %) (Glenn, 2011). PacBio (Eid *et al.*, 2009) and Nanopore (Olasagasti *et al.*, 2010) reads are much longer (1-100 kbp), but contain higher amounts of errors (11-15 % for PacBio (Rhoads and Au, 2015), similar for Nanopore (Weirather *et al.*, 2017)). These characteristics often suggest an assembly strategy, but there are many assembly programs to choose from. For sequence alignment, a related area, a large number of different alignment programs have been developed (more than 70 in 2012, more than 90 in 2016 (`http://www.ebi.ac.uk/~nf/hts_mappers/`) (Fonseca *et al.*, 2012)). Questions about the usefulness of producing this many different programs that differ only slightly have been raised (`http://www.opiniomics.org/an-embargo-on-short-read-alignment-software/`). A substantial number of genome assembly programs have been released, but not as many as the alignment programs. Some are tailored to specific approach (e.g. ALLPATHS-LG requires Illumina libraries created according to a specific recipe (Gnerre *et al.*, 2011)), while others can handle a multitude of different technologies and methods (e.g. Celera Assembler and variants can assemble all the different sequencing technologies available (Miller *et al.*, 2008; Koren *et al.*, 2017)). Being able to run a specific program is often the first hurdle to pass, and the second is being able to assemble a specific dataset given a set of computing resources. De Bruijn graph based assemblers often require substantial amounts of memory on one computing node, and OLC based assemblers might require a large degree of parallel computing resources available (a large multiple CPU node or computing farm) to be able to assemble datasets based on sequences from a genome similar in size to the human genome.

Sometimes, a genome sequencing project might contain multiple forms of

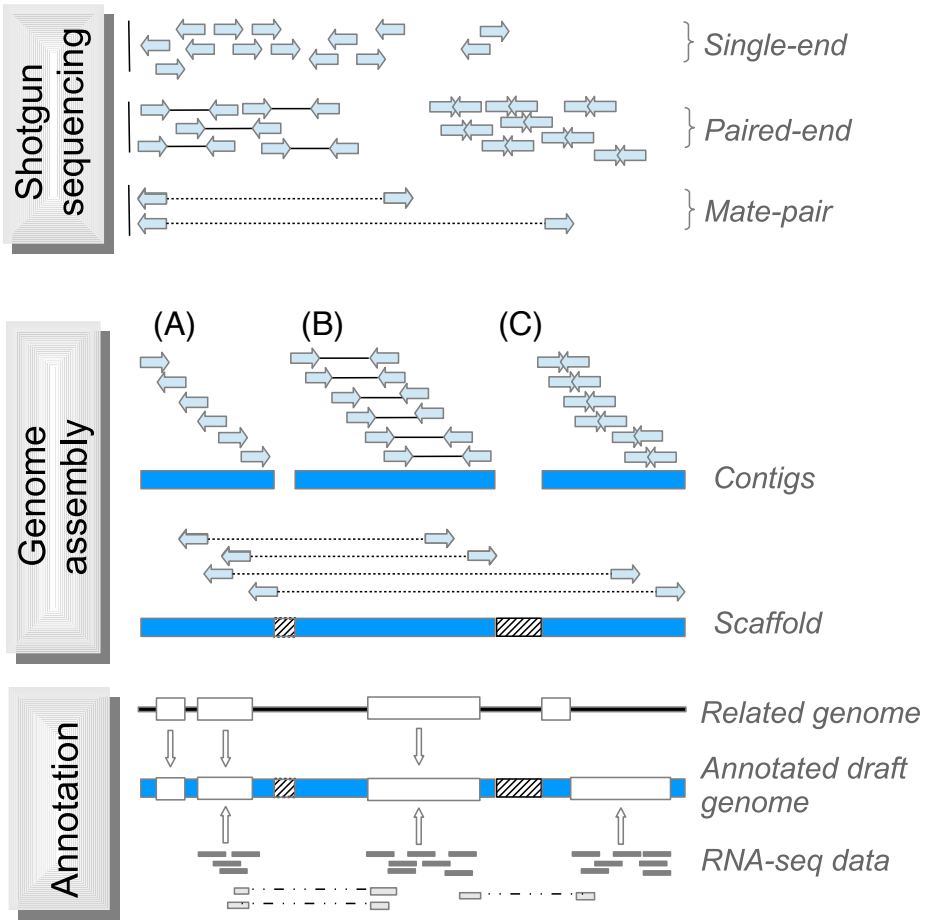

Figure 1.1: The sequencing, assembly and annotation process in brief. Single, paired-end or mate-pair reads are generated for a species. These are then assembled into contigs (blue boxes). Paired information (both paired-end and mate-pair) can be used to scaffold the contigs, with gaps representing unknown sequence between the contigs (hatched pattern). The assembly is then annotated, often by utilizing RNA-Seq data from the same species, where exons are shown in white boxes. If an annotation exists for related species, this can also be used. Used under CC BY 3.0. From Ekblom and Wolf (2014).

sequencing data (such as for the budgerigar in Assemblathon 2 (Bradnam *et al.*, 2013)), and there might not be an obvious choice for the final assembly strategy. Though, the use of different assemblers on the same data can lead to different assemblies (Bradnam *et al.*, 2013). Some assemblers, such as Celera Assembler, can use multiple types of sequencing data, but might not be the most efficient tools for a given task. For instance, a dataset consisting of Illumina reads generated for ALLPATHS-LG as well as PacBio reads, has multiple paths that lead to a final assembly. One path might be an assembly with Celera Assembler using all reads (Koren *et al.*, 2012). Another might be to create an assembly with the Illumina reads and ALLPATHS-LG, and then use PBJelly (English *et al.*, 2012) to close gaps in the ALLPATHS-LG assembly based on the longer (and presumably more repeat-spanning) PacBio reads. Both ALLPATHS-LG and Celera Assembler implements tools to correct the errors in the reads, creating more accurate reads before assembly, but standalone tools also exists (Alic *et al.*, 2016). A third path is to create several draft assemblies that could be merged or reconciled with Metassembler (Wences and Schatz, 2015) or a similar reconciliation tool (Alhakami *et al.*, 2017). There are also tools specifically created to use paired Illumina reads to close gaps, IMAGE (Tsai *et al.*, 2010) and GapFiller (Nadalin *et al.*, 2012). Finally, there are programs that recall the consensus sequence and correct errors, such as Pilon (Walker *et al.*, 2014) for Illumina reads and Quiver (Chin *et al.*, 2013) for PacBio reads. A second iteration of these programs might further improve the accuracy because more reads will map (because of the already increased accuracy of the consensus). In summary, for a relatively modest genome sequencing and assembly project, the choices and order of programs that can be run adds up to a multitude of possibilities.

1.1.2 Validation of genome assemblies

It is important to validate that the assembly generated is suitable for particular analyses. Simple statistics such as the N50 contig or scaffold lengths (the lengths at which half the assembly consists of sequences of those lengths or longer) are useful to get an impression of the contiguity and continuity, respectively, of an assembly, but they do not reflect its quality. Some tools such as FRC^{bam} (Vezzi *et al.*, 2012) and REAPR (Hunt *et al.*, 2013) compare the assembly to expectations based on the nature of the sequencing reads, i.e. insert size and orientation of the reads, and can give an impression of the best assembly when multiple have been generated. However, they are not as useful when only one assembly has been generated. The genic content of an assembly is often of particular interest, as sequencing and assembly of the genome of a particular

species is often conducted to investigate that species' repertoire of genes. The tools CEGMA (Parra *et al.*, 2007, 2009) and BUSCO (Simão *et al.*, 2015) both show the degree to which the genic content is captured by the assembly, and in particular BUSCO is recommended for all genome assembly projects (`http://www.acgt.me/blog/2015/5/18/goodbye-cegma-hello-busco`). To be scored as complete genes, the genes have to be contained on one sequence (contig, scaffold or chromosome), with most of its exons in correct order and orientation. Assemblies with most genes found complete by BUSCO and CEGMA can be assumed to be of good quality.

It is also possible to use existing genomic data such as finished bacterial artificial chromosomes (BACs), different forms of expressed sequence tags (ESTs) or complementary DNA (cDNA) sequences as validation of an assembly (Warren *et al.*, 2010). Transcriptomes generated from RNA-Seq or similar HTS technology can also be used (Ryan, 2013). While the assembly of a transcriptome and the validation of it might be more complicated than for a genome assembly (Smith-Unna *et al.*, 2016; Honaas *et al.*, 2016), mapping the transcriptome to a genome assembly gives a good indication of the quality of the latter. The more complete the assembly is, the more transcripts are expected to map at complete lengths.

For some species, such as human, a high-quality reference genome assembly exists. This is a good standard to compare against, for instance when testing a new sequencing technology or assembly approach. However, there are many differences between populations of humans, e.g. large inversions (Bansal *et al.*, 2007). These differences might be tagged as assembly errors if one is unaware of them.

1.1.3 Finalizing a genome assembly and annotation

For many sequenced species, their genome is represented by a set of scaffolds in a genome assembly, without information about which chromosome (or chromosomal region) the scaffolds originate. During evolution, chromosomes split and fuse, sometimes resulting in differences in chromosome number between closely related species. For example, chimpanzees and humans differ in chromosome number, with 24 and 23 pairs, respectively (Fan *et al.*, 2002). Arranging scaffolds into large scale reconstructions of chromosomes, like linkage groups (see below), is important to disentangle the evolution of genes as it allows the identification of syntenic regions (regions that share ancestry, often contain the same genes) between two species (Braasch *et al.*, 2015). Based on genotypes of family material, linkage maps traditionally made it possible to arrange data

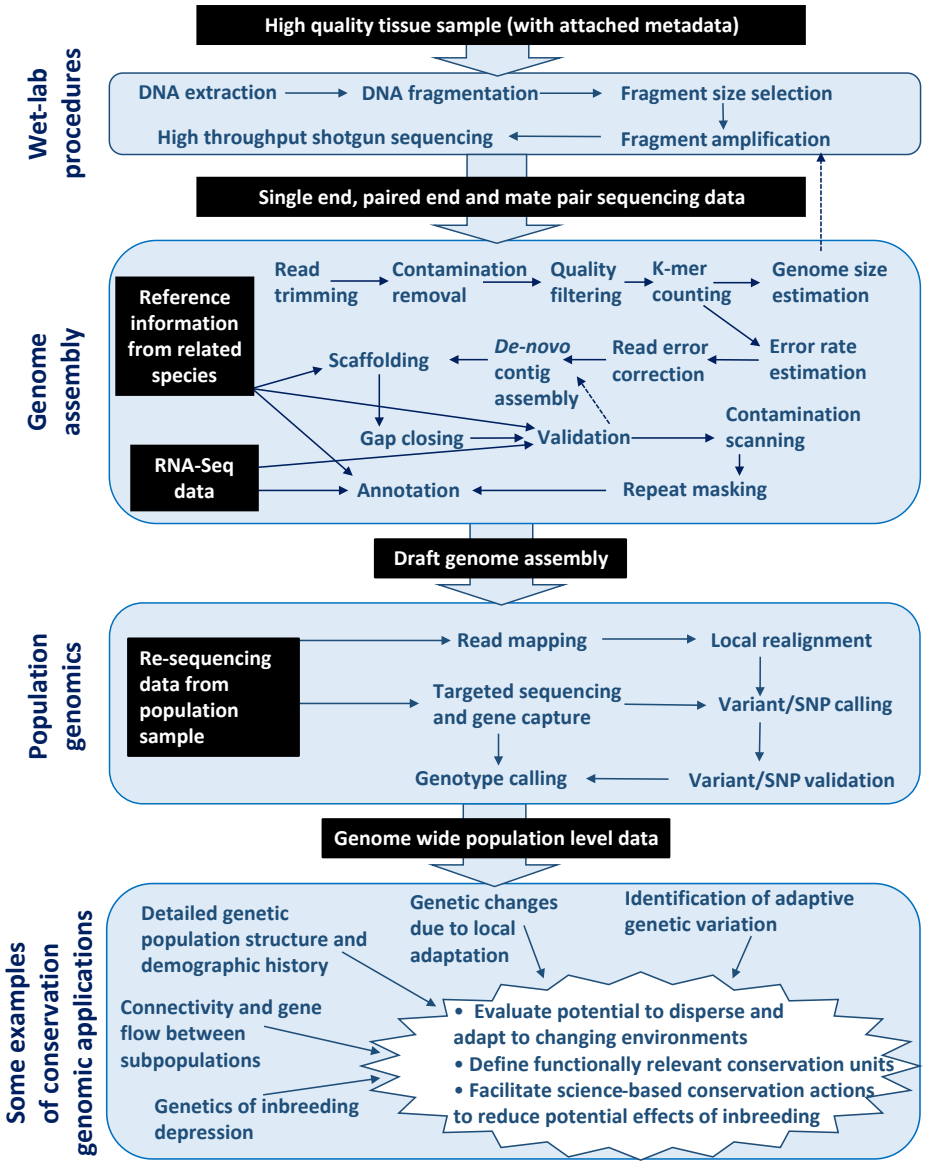

Figure 1.2: The workflow of a typical genome assembly and annotation project. The black boxes with white text represent resources that are generated during the project. The project starts with DNA extraction (preferably high-molecular weight DNA) and sequencing of the DNA. The reads then are processed (trimming, filtering, error correcting) before assembly, and the assembly validated before being annotated. As shown here, based on the annotated assembly, population re-sequencing data can be used to find genes under selection. The last box contains examples of uses of an annotated genome assembly in a conservation biology setting. Used under CC BY 3.0. From Ekblom and Wolf (2014).

into linkage groups by providing information on the order of sets of markers and the approximate distances between them (Rastas *et al.*, 2015). The scaffolds can be matched against these sets of markers, and ordered and oriented into linkage groups. This process is often costly and time-consuming, since it requires family material to track recombination events from the parents in the offspring. Two markers physically close to each other will recombine less often than two distant markers, which can in turn inform about the relative position of the markers on the genome. Presently, there are alternative sources of long-range information, enabling assemblies that have reconstructions (scaffolds) close to chromosome size, such as optical mapping (Howe and Wood, 2015; Seo *et al.*, 2016), chromosome conformation capture (Bickhart *et al.*, 2017) or linked reads (Yeo *et al.*, 2017; Weisenfeld *et al.*, 2017). All these create scaffolds with gaps of unknown sequence between contigs. For many species, sufficient coverage in PacBio reads can create chromosome arm reconstructions (Vij *et al.*, 2016; Koren *et al.*, 2017), which results in contiguous sequences without gaps.

A genome assembly is most useful when different features such as genes, transposable elements and other repeats are annotated, i.e. when they have a location on a scaffold/chromosome and an unique identifier (Figure 1.1 and 1.2). For instance, when investigating the difference in expression between two experimental set-ups with RNA-Seq (Conesa *et al.*, 2016), an annotated genome is a useful background. We often distinguish between structural annotation, finding all the genes with their intron and exon structure, and functional annotation, identifying the genes and their properties (active in which pathways etc.) (Yandell and Ence, 2012; Hoff and Stanke, 2015). Annotation is usually a multi-stage process. It starts with identifying as many repetitive elements as possible, possibly creating a custom-made repeat library using both homology-based and de novo tools (Bergman and Quesneville, 2007). Complete transposable elements contain genes (to facilitate transposition), so using a repeat library to mask them helps with identifying the genes of the species under investigation, and not genes found in transposable elements. After repeat masking, ab initio gene prediction programs such as AUGUSTUS (Stanke *et al.*, 2008), GeneMark (Lomsadze *et al.*, 2005), or SNAP (Korf, 2004) need to be trained, i.e., optimized for the specific species with regards to codon bias and splicing signals. RNA-Seq data can be used to train AUGUSTUS and GeneMark-ET (Hoff *et al.*, 2016), or a set of genes as annotated by CEGMA can be used to train SNAP (Campbell *et al.*, 2014a). A transcriptome assembled by Trinity (Grabherr *et al.*, 2011) or StringTie (Pertea *et al.*, 2016) (or by a combination of these tools with PASA (Haas *et al.*, 2003)) is often aligned to the genome

as evidence for expressed genes, and a non-redundant protein database such as UniProtKB/SwissProt (UniProt Consortium, 2015) can be included as a set of curated proteins, possibly in addition to proteins from well-annotated close relatives. All this information can be integrated by using a program such as MAKER (Campbell *et al.*, 2014b; Holt and Yandell, 2011) or EVM (Haas *et al.*, 2008). This approach provides a set of predicted transcripts and proteins, together with a GFF (General Feature Format) track with positions of all the annotated features, describing their properties. The predicted proteins can be used in InterProScan (Jones *et al.*, 2014) to classify which pathways and functions the different proteins have.

1.1.4 Impediments for optimal genome assembly

The most limiting aspect for obtaining a complete and contiguous genome assembly is the extent of the repetitive content of the genome (Treangen and Salzberg, 2012). Repeats are very similar or identical sequences that occur multiple times in a genome, and when they are longer than the read length (or the k-mer length for de Bruijn graph based assemblers), the assembler is unable to place reads originating from them uniquely. Repeats therefore fragment the assembly by limiting the length of contigs. However, if paired reads with an insert size longer than the longest repeat length are available, this limitation can partly be compensated by joining the contigs on each side of a repeat into scaffolds, with the repeat represented as a gap (Simpson and Pop, 2015) (Figure 1.1).

Repeats can be grouped into two categories: interspersed repeats (such as transposable elements occurring in multiple loci) and tandem repeats (a motif repeated in tandem). Transposable element (TE) content is highly correlated with genome size (Elliott and Gregory, 2015; Chalopin *et al.*, 2015), and is likely the largest factor contributing to fragmented assemblies (Sotero-Caio *et al.*, 2017). Large fractions of vertebrate genomes are filled with active and inactive fragments of TEs, with more than 40 % of the genome of zebrafish, and more than a third in the genomes of mammals consisting of TEs (Chalopin *et al.*, 2015). Evolutionary old TEs will accumulate mutations and will diverge from the original sequence, and therefore losing their repetitive nature over time. However, a high fraction of TEs in the genome might not be a large impediment if they are dissimilar. Many genomes with a high fraction of TEs have been assembled well; the human genome being a good example.

Tandem repeats, and in particular the short tandem repeats (STRs) also called microsatellites (unit size 1-10 bp), are also a limiting factor, especially

when these are longer than the read length. Additionally, tandem repeats mutate at a high rate, creating heterozygous loci in the genome, hampering assembly (see below). Variation in STRs has been a much-used genomic resource in many applications, ranging from forensics to elucidating population structure. They mutate by strand slippage or recombination (Ellegren, 2004), and have mutation rates 10 to 10 000-fold higher than other types of DNA (Verstrepen *et al.*, 2005). Most mutate at a rate of 10^{-6} to 10^{-3} per cellular generation (Verstrepen *et al.*, 2005), but mutation rates can be as high as $>10^{-2}$ (Ellegren, 2004). Tandem repeat content varies between 2,000 bp/Mbp to 55,000 bp/Mbp (0.2 to 5.5 %) in investigated eukaryotes (metazoans, green algae, plants and yeast) (Mayer *et al.*, 2010; Zhao *et al.*, 2014) when measured for unit sizes 1-50 bp. Contrary to TEs, STR content is not significantly correlated with genome size (Mayer *et al.*, 2010; Zhao *et al.*, 2014).

Another limiting aspect of genome assembly is the size of the genome being assembled. The amount of sequencing data optimal for assembly is directly based on the size of the genome under investigation. The larger the genome is, the more sequencing data must be generated to reach coverage suitable for assembly (depending on technology, this can be up to 100x). Insufficient coverage can lead to some parts of the genome not being sequenced, therefore creating gaps due to missing data. Smaller genomes require less computational power since smaller amounts of data need to be generated and assembled. Most of the genomes assembled to date have been less than 5 Gbp in size, and only a few very large plant genomes have so far been assembled, such as the Norway (*Picea abies*) and white spruce (*Picea glauca*) genomes (Nystedt *et al.*, 2013; Birol *et al.*, 2013) that are both around 20 Gbp in size, requiring extraordinary computational efforts. However, the resulting assemblies were quite fragmented, with N50 scaffolds at 5 kbp and 20 kbp, respectively.

Although a minority among all living species, diploid species (with paired chromosomes, one copy inherited from the mother and one copy from the father) make up a large fraction of the species investigated by scientists. However, diploid genomes are problematic for an accurate genome assembly, as most of the assemblers available today assume a haploid genome (a single set of unpaired chromosmes). This was an appropriate approximation for the first decade of genome sequencing and assembly, where most of the species assembled had little variation in their genomes, due to small effective population size (humans (Charlesworth, 2009)) or by being inbred lab strains (e.g. mouse (*Mus musculus*, Mouse Genome Sequencing Consortium *et al.* (2002)). With more and more non-model or wild-caught species being investigated, heterozygos-

ity (differences between the two homologous chromosomes from the mother and the father) is becoming a major issue. Since most assemblers are not designed to handle heterozygosity, they output one of the two alternatives, and the other is discarded. Alternatively, the heterozygous regions might induce gaps. This means that the output of an assembly of a heterozygous genome will not represent the actual genome it derives from. One solution is to output a graph representing the homologous regions, a feature that is implemented in, for instance, canu (Koren *et al.*, 2017), Supernova (Weisenfeld *et al.*, 2016) and Falcon-Unzip (Chin *et al.*, 2016). However, most downstream software is currently designed to use single sequences per individual rather than graphs and will also need to be updated. Ultimately, a single data structure representing the variation in a population or a species is a desired goal, and there are ongoing efforts to create this (Paten *et al.*, 2017) .

1.2 Comparative genomics

Comparing the genome assemblies of several species is a powerful method of discovering what is common and unique among the assemblies. The second vertebrate genome assembly published, that of fugu (*Fugu rubripes*, Aparicio *et al.* (2002), was compared in detail to the first vertebrate genome assembly, the human (Lander *et al.*, 2001; Venter *et al.*, 2001) and many similarities were found despite the species being separated by 450 million years of evolution. With HTS a multitude of genomes can be sequenced and assembled allowing investigation into such complex traits as the evolution of song in birds (Zhang *et al.*, 2014). While subject to much research, there are still unanswered questions in genome evolution, such as what exactly make up the non-coding part and how these sequences are gained and lost (Gregory, 2005). With multiple genome assemblies it is possible to investigate what exactly exists in different genomes, for instance, investigating the relationship between genome size and transposable elements (Canapa *et al.*, 2016). The roles of transposable elements in genome evolution are beginning to be determined, but the roles of another type of repeated element, the tandem repeats, are not.

1.2.1 STRs affecting function

Short tandem repeats (STRs) content varies across vertebrates, with frequencies from approximately 100 loci/Mbp to 1,000 loci/Mbp, and densities from 1,000 bp/Mbp to 50,000 bp/Mbp (Adams *et al.*, 2016) when measuring for unit sizes 2-6 bp. While using a different approach and different definition for a STR

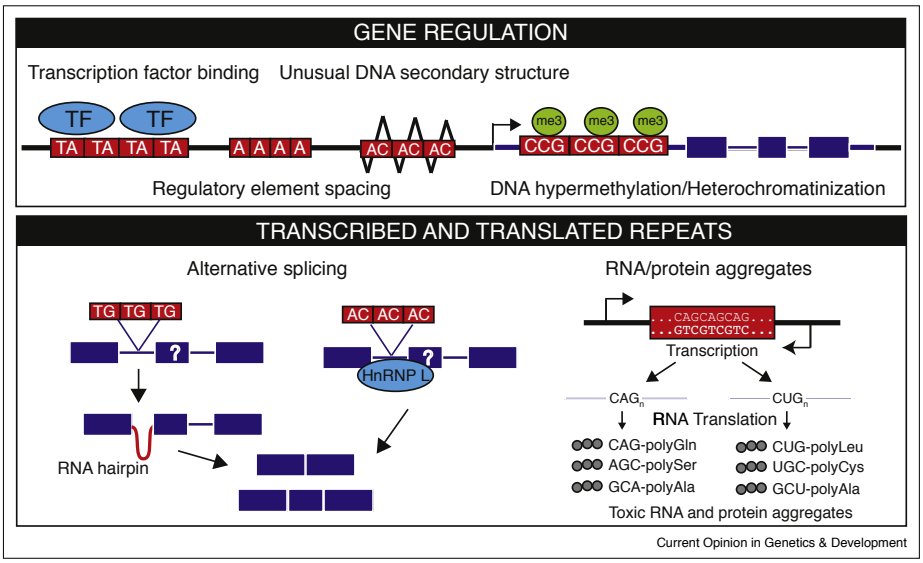

Figure 1.3: Different known and proposed mechanisms whereby STRs can influence function. Top, left to right: Transcription factors might bind to STRs, STRs can affect the distance between binding sites, STRs can induce different DNA secondary structures and can modulate DNA methylation and heterochromatinzation. Bottom, left to right: STRs can affect alternative splicing, RNA protein binding sites, and large expansions of STRs might create toxic RNA or protein aggregates. Purple boxes, exons; black lines, DNA; red boxes, STRs; blue circles, RNA/DNA binding proteins; grey circles, amino acids; green circles, DNA modifications. The main text discusses only the effect of STRs in exons. From Gymrek (2017). Used with permission.

(2-6 bp unit size in contrast to 1-50 bp), this is similar to the values found in eukaryotes in general above. Since genes only make up a fraction of vertebrate genomes, most of these STRs would occur outside genes, evolving mostly neutrally. Some STRs are located in protein coding regions (Figure 1.3), i.e. 4500 STRs occur in protein coding regions in humans (Willems *et al.*, 2014). Most of these are found in genes categorized in functional groups such as transcription, the regulation of transcription, and receptors (Mularoni *et al.*, 2010; Legendre *et al.*, 2007) in humans, and in similar groups in yeast (Albà *et al.*, 1999), fruit fly (Huntley and Clark, 2007) and plants and algae (Zhao *et al.*, 2014). An STR residing within a gene often encodes an amino acid repeat if the repeated motif is in the correct reading frame (Figure 1.3), and these can overlap with intrinsically unstructured regions (Simon and Hancock, 2009), which are abundant in proteins that interact with other proteins such as transcription factors and receptors (Huntley and Clark, 2007). STRs also occur in promotor regions (Figure 1.3), where they have been shown to affect the expression of genes in yeast (Vinces *et al.*, 2009) and humans (Gymrek *et al.*, 2016; Quilez *et al.*, 2016). Also, STRs in introns can affect the splicing of RNA (Hefferon *et al.*, 2004; Press *et al.*, 2017), and even contribute to speciation in primates by affecting gene expression divergence (Sonay *et al.*, 2015).

Following the completion of the human genome project, many researchers had hoped that we could start to untangle the genetic basis of traits and susceptibility to disease. Genome wide association studies have made much progress towards these goals. For instance, about 90 markers (single nucleotide polymorphisms, SNPs) have been found to explain 27.4 % of the heritability of height (Marouli *et al.*, 2017). Heritability is the concept of how much variation of a trait in a population is explained by genetic variation. For height, the expectation has been that 80 % of the variation is explained by genetic variation (Visscher, 2008). The genetic basis for many traits is more complicated and divided among more loci than many proponents expected, and gave rise to discussions about "missing heritability" (Eichler *et al.*, 2010), which would explain the difference between 27.4 % and 80 % in the case of genetic basis of height. Since studies only using SNPs cannot explain all the variation of a trait, variation in STR length has been suggested as one factor that could explain some of the "missing heritability" (Press *et al.*, 2014; Sawaya *et al.*, 2015). For instance, when investigating STRs in promoters that affect gene expression, it was found that variation in SNPs near these STRs does not necessary reflect the variation in the STRs, implying that they are not linked and evolve independently (Quilez *et al.*, 2016). The allele of a SNP does not inform about the

STR allele. This is in line with earlier theoretical studies, which showed that high mutation rates in an STR creates linkage disequilibrium with close SNPs (Sawaya *et al.*, 2015). This indicates that studies incorporating variation in both SNPs and STRs, and possibly other variation such as larger deletions and insertions, would have the potential to explain more of the heredity of different traits than current methods. Neither sequencing technology nor the tools used in used in such analyses are currently designed to do this.

1.2.2 The Gadiformes and their immune system

The Gadiformes, or cods and allies, consists of 613 extant species (Eschmeyer *et al.*, 2017), with its members some of the most important commercial fish species in the world (FAO, 2016). Many Gadiformes, and especially the codfishes such as cod and haddock, are commercially important fishes for many societies (Olsen *et al.*, 2010). Because of this importance, there has been a substantial research focus on Atlantic cod, with the release of its genome assembly in 2011 as a major milestone (Star *et al.*, 2011). This was the fourth teleost genome assembly published after fugu (*Takifugu rubripes*), tetraodon (*Tetraodon nigroviridis* and medaka (*Oryzias latipes*), and the first non-model, marine teleost. It was also the first vertebrate genome sequenced and assembled with a pure 454 sequencing approach. The main reported finding was a lack of the genes involved in the major histocompatibility complex (MHC) II (involved in defense against bacterial pathogens), accompanied by expansions in the *MHCI* (usually involved in defense against viral pathogens) and *TLR* (recognizing both bacterial and viral pathogens) genes (Star *et al.*, 2011). It is not clear how, or why, cod lost these genes (Star and Jentoft, 2012). Continuation of this work has identified that the *MHCI* genes have evolved into two distinct clades, one of which has a novel signaling peptide resembling those known from the MHCII pathway (Malmstrøm *et al.*, 2013). Within the innate defense the TLR genes have seen both expansions and losses in cod (Solbakken *et al.*, 2016b), and the expansions have been suggested to be associated with the loss of *MHCII* genes throughout Gadiformes (Solbakken *et al.*, 2017).

In addition to TLRs, other pattern recognition receptors (PRRs) such as RIG-I like receptors (RLRs), NOD-like receptors (NLRs), and C-type lectin receptors (CLRs) are responsible for the initial sensing of microorganisms (via pathogen-associated molecular patterns, PAMPs) and damage (damage-associated molecular patterns, DAMPs), and subsequently activating the inflammatory response (Takeuchi and Akira, 2010) (Figure 1.4). While the RLRs consist of three members (Zhu *et al.*, 2013), the NLRs have around 400 members

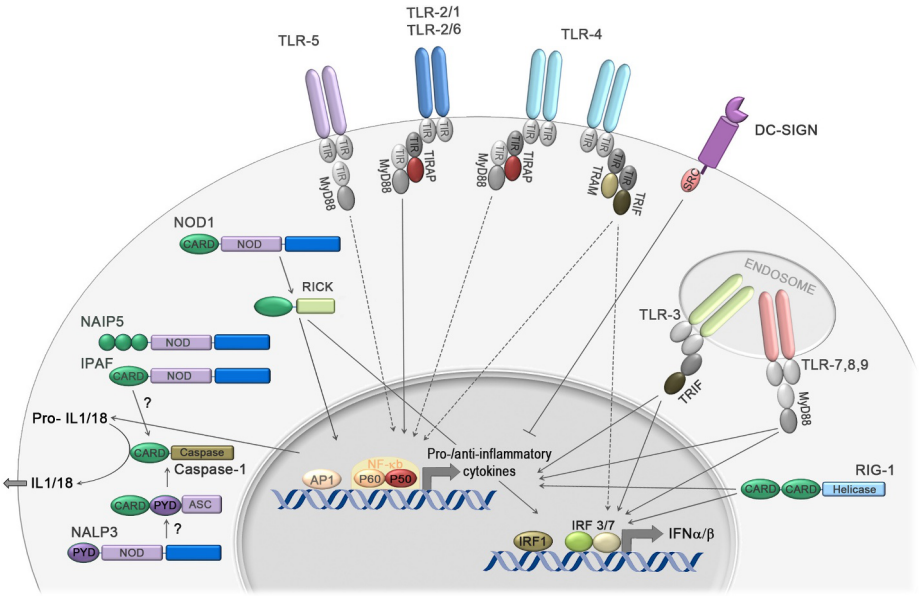

Figure 1.4: . An overview of the cellular location of innate immune receptors. The Toll-like receptors (TLRs) are located to the cell membrane and to endosomes together with the C-type lectin receptors (CLRs) such as DC-SIGN. NOD-like receptors (NLRs) are found in the cytosol, together with the RIG-like receptors (RLRs). These are all pattern-recognition receptors, recognizing molecular patterns either associated with damage or microbial pathogens. Used under CC BY 2.0. From Müller *et al.* (2011).

in zebrafish (Howe *et al.*, 2016), with varying numbers in other teleosts (Howe *et al.*, 2016; Stein *et al.*, 2007). These additional families of PRRs are just in their initial phase of characterization in teleosts, and the Gadiformes are especially interesting since they have a "non-traditional" immune system compared with other teleosts.

The first version of the cod genome (Star *et al.*, 2011)). was a rather fragmented assembly compared with Sanger sequenced or current HTS teleost assemblies. Most of teleost species' assemblies available at that time were traditionally Sanger sequenced, where reads were at least twice as long as those generated with the 454 sequencing technology used for cod. This shorter read length for cod could be expected to negatively influence the assembly contiguity. However, it was also found that 32 % of the contig edges contained an STR and 24 % of the gaps in scaffold were flanked by STRs (Supplementary Note 7 in Star *et al.* (2011)). Cod has further been identified as a species harboring a large amount of STRs compared to other marine species (Jiang *et al.*, 2014) and to a diverse set of vertebrates . Further, the Atlantic cod genome has extreme amounts of the dinucleotide tandem repeat AC compared to all vertebrate genomes within the Ensembl database (Star *et al.*, 2016a). These findings could therefore indicate that the fragmentation of the first version of the cod genome assembly may be connected to a large amount of STRs.

2 Aims

The main goal of this thesis has been to investigate and to evaluate how different aspects of the architecture of the genome (such as gene content, transposable elements, STRs, SNPs) and technical aspects (such as sequencing platforms, coverage, assembler and assembly algorithms) can affect the assembly of a genome, and in addition gain insight into what is 'good enough' for particular biological questions being asked (presence/absence of genes, synteny, usage in phylogeny, total gene content). A secondary goal has been to use these assemblies to investigate the evolution of immune genes in teleosts, in particular *MHCI*, *TLRs* and *NLRs* in Gadiformes and how their copy number estimations in light of their repetitive nature is affected by the genome assembly.

3 Summary of works

3.1 Work I

The new era of genome sequencing using high-throughput sequencing technology: generation of the first version of the Atlantic cod genome

In 2008 my colleagues received funding to sequence, assemble, and annotate the Atlantic cod (*Gadus morhua*) genome. This chapter published in the book *Genomics in Aquaculture* gives a summary of the rational for sequencing the Atlantic cod, the sequencing method, assembly methods, and annotation method used. It is the first chapter of the book, written as an introduction to sequencing and assembly using cod as a case study. Cod was one of the first vertebrates to be sequenced, assembled, and annotated based on HTS technology alone, and this was challenging compared to the older Sanger sequencing projects, partly because the 454 technology provided reads of shorter length. In this book chapter we discuss all the processes from sequencing to assembly, and the different assembly approaches (OLC and de Bruijn). Some challenges of genome assembly are described, including the size, repetitive content, heterozygosity and errors in sequencing reads. We discuss the relationship between repetitive sequences and paired reads (also called mate pair and paired-end reads). Two assemblies were originally generated for the Atlantic cod, one with Celera Assembler (long contigs) and one with Newbler (long scaffolds) and both were more fragmented than other publicly available teleost assemblies. We also describe the annotation process and the synteny between cod and other teleosts, before discussing the need for an improved genome assembly for cod and different methods for creating it.

3.2 Work II

An improved genome assembly uncovers prolific tandem repeats in Atlantic cod

While the Atlantic cod genome assembly published by Star *et al.* (2011) (gadMor1) was suitable for many purposes, it was more fragmented than other comparable assemblies, such as tetraodon, medaka and stickleback (*Gasterosteus aculeatus*). Short tandem repeats (STRs) were found at or near the end of many contig termini (32 % of them) and were suspected to be responsible for the fragmentation. We wanted to investigate the reason for the fragmentation and identify its biological underpinning.

We generated four different draft assemblies using different combinations

of sequencing technologies (Illumina, 454, PacBio, and Sanger), and assemblers (Celera Assembler, Newbler and ALLPATHS-LG), and validated these with a variety of tools. Each of them had its own strengths and weaknesses, therefore we created a reconciled assembly based on all four draft assemblies using a custom created program. This reconciled assembly (gadMor2) has a 50-fold increase in N50 contig length compared to the previous cod assembly, and reduced the amount of gap bases 15-fold. This new assembly was annotated using the automated MAKER pipeline.

We confirmed the high density (bp/Mbp) and frequency (loci/Mbp) of STRs in Atlantic cod, both inside and outside of genes. This amount of STRs is significantly higher than that in any other species found in the Ensembl database. By comparing the draft assemblies and gadMor1 to gadMor2, we were able to measure how many contigs from each assembly were overlapping with different features in gadMor2 (SNPs, indels, transposable elements, sequencing read coverage and STRs). While there is about 10 % STRs in gadMor2, up to half of the contig termini from gadMor1 and all draft assemblies except the one based on PacBio reads, overlap with STRs. This is a significant enrichment, and likely the reason why these assemblies are fragmented.

Because STRs are highly mutable, the high density and frequency of these STRs both inside and outside protein-coding regions suggests most of the STR loci are polymorphic at the population level, representing a substantial amount of standing genetic variation.

3.3 Work III

Whole genome sequencing data and *de novo* draft assemblies for 66 teleost species

To investigate when the loss of genes involved in the MHC class II pathway occurred in the lineage leading to cod, we needed assemblies sampled broadly enough to properly represent the teleost group, in addition to those publicly available. To produce these by the traditional approach with multiple sequencing libraries of different insert sizes would have been prohibitively expensive. We therefore explored different sequencing and assembly methods by utilizing publicly available sequencing data from the budgerigar (Bradnam *et al.*, 2013; Ganapathy *et al.*, 2014) to arrive at an optimal, cost-effective strategy for obtaining assemblies of sufficient quality for phylogenomic and gene presence/absence analyses.

We sequenced the genomes of 66 teleost species, representing all major lineages of teleosts, using a single sequencing library with insert size around 400

bp and read lengths of 150 bp, at 9-39x coverage. The qualities of the assemblies were assessed by gene-space completeness as measured by CEGMA and BUSCO. We found a significant correlation between N50 scaffold length and the number of genes found by CEGMA and BUSCO, but no correlation between number of genes found and coverage, and no correlation between coverage and N50 scaffold length.

We also report on a phylogeny based on mitochondrial sequences extracted from the sequenced genomes and from GenBank as a validation of the nuclear marker phylogeny reported in Work IV. We show that sequences taken from GenBank and the new mitochondrial genomes cluster according to expectation, indicating that there is little chance of contamination or taxonomic mis-assignment.

3.4 Work IV

Evolution of the immune system influences speciation rates in teleost fishes

We used the data reported in Work III together with 10 publicly available teleost genome assemblies to create a time-calibrated phylogeny using fossil constraints. By searching for the relevant genes in the assemblies of the 76 species, we were able to infer when the genes involved in the MHCII pathway were lost in the lineage leading to Atlantic cod. We found that this happened approximately 105 million years ago, in the common ancestor to all Gadiformes.

MHCI has expanded multiple times during the evolution of teleosts, in Gadiformes but also within Percomorphaceae. The expansion of *MHCI* is correlated with species diversity, suggesting that the copy number of *MHCI* can be a driving feature in speciation within Gadiformes and Percomorphaceae. With a large number of nearly identical *MHCI* gene copies in a genome, these may have the same effect as repetitive sequences for the assembly process, and might be excluded from the assembly. Therefore, we could not use the assemblies to count the number of *MHCI*, but had to compare read depth of assembled *MHCI* sequences, with the read-depth of assembled single-copy genes to estimate *MHCI* copy numbers.

3.5 Work V

Genomic architecture of codfishes featured by expansions of innate immune genes and short tandem repeats

According to the findings in Work II-IV, Gadiformes seem to have chosen

different strategies with regard to immune genes and STR content compared to other teleost species. Neither of these attributes is easily investigated with fragmented assemblies such as those used in Work III and IV. Because of it relatedness to cod (diverged from cod about 13 million years ago), and because it is an important commercial and ecological species, we performed PacBio and Illumina sequencing of haddock (*Melanogrammus aeglefinus*), generated a genome assembly, and annotated it.

We first compared selected immune related genes to the cod genome assembly (gadMor2), and found that about half of the estimated number of *MHCI* genes (from Work IV) can be found in the assemblies. There were only minor differences in the TLR repertoires of the two species. We also investigated the genes encoding NOD-like receptors (NLRs) and found that the teleosts likely have an expansion of these compared to other species. About twice as many are found in cod compared to haddock, but most are found on short contigs/scaffolds with high sequencing read coverage, indicating that multiple copies are collapsed into one. When investigating the unitigs from the assemblies that contributed to the final assemblies of cod and haddock, we find approximately the same number of putative *NLRs* in both species.

Both cod and haddock have a much higher density and frequency of STRs compared to other fish species of the Ensembl database. This is also reflected in the coding sequences, with about twice the number of genes containing an STR compared to the other species. For all fish species, there is a significant enrichment in STRs in genes involved in transcription, but for cod and haddock there is also enrichment in STRs in genes involved in signal transduction.

4 Discussion

4.1 The loss of MHCII and the background for a more contiguous assembly for cod

It has long been known that Atlantic cod does not respond to vaccination in a similar way that other teleosts do (Pilström *et al.*, 2005). However, it was not until the genome assembly of Atlantic cod and its annotation (Star *et al.*, 2011) became available that we properly understood the reasons why; i.e. the loss of genes involved in the MHCII pathway. While the first assembly of the Atlantic cod genome (gadMor1) was sufficiently complete to discover the loss of MHCII, the assembly was fragmented and the scaffolds/contigs did not have a chromosomal context; their relationship to each other was not clear. A low-density linkage map was available(Hubert *et al.*, 2010), but only half of the assembly could be anchored to it. The gadMor1 assembly was published in 2011, the same year that ALLPATHS-LG (Gnerre *et al.*, 2011) was published, which demonstrated remarkable assembly statistics using Illumina sequences alone. The PacBio sequencing platform was also beginning to gather momentum with the publication of tracing the origin of E. coli after an outbreak in Germany (Rasko *et al.*, 2011). The need for a more contiguous genome, and these developments in sequencing and assembly technologies, laid the foundation for further work on the cod genome assembly (as discussed in Work I (Tørresen *et al.*, 2016)).

4.2 The creation of highly and less contiguous assemblies of codfishes

For gadMor1, many genes are found on multiple contigs (e.g. Figure 2 in Work II). The mean gene lengths of the annotated teleosts on Ensembl range from 6 to 30 kbp depending on species. As gene length (exons plus introns) is correlated with genome size (Yandell and Ence, 2012), the gene length of Atlantic cod is expected to be similar to teleost species with similar genome sizes. Of the species included in the Ensembl database, platyfish, medaka, and Amazon molly have the most comparable genome sizes to cod, and gene lengths in these species range from 13 – 17 kbp. Thus, with a N50 contig size around 15 kbp, about half the genes of Atlantic cod should be fully contained in contigs. Unfortunately, none of the Newbler, ALLPATHS-LG, and Celera Assembler assemblies based on 454 and Illumina data had sufficient contig lengths for most of the genes to be fully contained (Work II (Tørresen *et al.*, 2017)). PacBio reads

are longer than Illumina and 454 reads, and therefore span more repeats. At the time, the usual approach was to correct the error-prone PacBio with Illumina/454 reads (Koren *et al.*, 2012). When attempting this correction, we ended up with fragmented datasets, likely since the limitations of the 454 and Illumina data were transferred to the PacBio data. This was due to either lack of coverage in certain regions or inability to span repeats in the PacBio reads. We used an unconventional approach by performing an assembly without correcting the PacBio reads first (also utilized in Work V). To our knowledge, the only other such published assembly was used for the salmon assembly (Lien *et al.*, 2016), but no actual sequence from the assembly created from uncorrected PacBio reads went into the final salmon assembly as it was merely used as quality control. However, there have been several recent publications advocating such approach to assembly with these kinds of data (Vaser *et al.*, 2017; Kamath *et al.*, 2017).

For both Work II and V, the amount of PacBio reads was moderate, with about 20-25x coverage. Similar amounts of PacBio data have dramatically improved the contiguity of assemblies compared to those without PacBio reads (Conte and Kocher, 2015; Koren *et al.*, 2012). Even higher numbers of PacBio reads can lead to contigs the lengths of chromosome arms (Vij *et al.*, 2016; Berlin *et al.*, 2015; Pendleton *et al.*, 2015). This has until recently been considerably costly. Notably, the moderate amounts of PacBio reads used in Work II and V have yielded assemblies where most genes are fully contained in contigs (Figure 2 in Work II, Figure 4.1). Thus, presuming the annotation is correct; multiple genes can be found on one contig (with around 700 Mbp genome and 20 000 genes; that is one gene present every 30 kbp), giving more confidence when looking at synteny between species. As an example of the improvements from gadMor1 to gadMor2, the region on LG11 associated with sex-determination (Star *et al.*, 2016b), contains substantial amounts of gaps in gadMor1, but the higher contiguity of gadMor2 provided more confidence in this region (also see Figure 4.1).

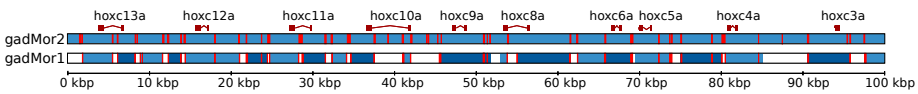

Figure 4.1: An example of the higher contiguity of gadMor2 compared to gadMor1. Seen here is the HoxC cluster in both assemblies. Blocks in shades of blue represents contig sequences, while white blocks are gaps and red lines tandem repeats. Gene models are drawn on top of the figure. In gadMor2 this region is a single contig, while it is 21 in gadMor1. Modified from Tørresen *et al.* (2017).

However, used in the right way, less contiguous assemblies can also be very powerful. We know that cod is missing the *MCHIIa* and *MHCIIb* genes (Star *et al.*, 2011). However, we did not know when this was lost from the evolutionary lineage leading to cod. By sequencing and assembling 66 teleost species (Work III (Malmstrøm *et al.*, 2017)), we were able to map the presence/absence of relevant genes on a fossil calibrated phylogeny (Work IV (Malmstrøm *et al.*, 2016)). While these kinds of assemblies cannot be used to reliably detect absence of one particular gene for a single species, it is a powerful tool when patterns emerge, for instance the absence of *MCHIIa* and *MHCIIb* in all Gadiformes despite their presence in all other species; Figure 2 in Work IV. These less contiguous assemblies seem to contain most genes (Figure 2 in Work III), and can therefore be used to investigate the evolution (for instance loss/expansion) of almost any gene/gene family, such as opsins (Cortesi *et al.*, 2015), *Mx* (Solbakken *et al.*, 2016a) and *TLRs* (Solbakken *et al.*, 2017).

4.3 The high frequency and density of short tandem repeats in codfishes

About 10 % of the cod genome assembly is found in short tandem repeats (STRs) (Work II). STRs are overrepresented at the edges of contigs from the previous assembly (gadMor1) and for three of the four draft assemblies (CA454ILM, ALPILM, NEWB454) that went into gadMor2 (Figure 6 in Work II). That is, around 30 – 50 % of contig edges overlap STRs. It is therefore likely that STRs fragment these assemblies and thus impede proper assembly. For CA454PB, the fourth draft assembly, we have utilized PacBio reads. Because of their length these PacBio reads were successful in spanning most STRs. Both cod and haddock (Work V) have a much higher amount of STRs in their genomes than other species, both in terms of percentage of sequence in STRs (density, bp/Mbp) and the frequency (loci/Mbp).

In Work V, we used the genome assemblies of two codfishes (haddock and cod) and compared these to other teleosts to find features shared among codfishes. The codfishes have about twice the number of genes with an STR (approximately 8000): 30 % of the annotated genes, with the other teleost at 17 % or less. All teleost assemblies are enriched for STRs in genes involved in transcription, similar to other eukaryotes (Mularoni *et al.*, 2010; Legendre *et al.*, 2007; Albà *et al.*, 1999; Huntley and Clark, 2007; Zhao *et al.*, 2014). The codfish assemblies are also enriched for STRs in genes involved in signal transduction. These genes encode proteins that regulate small GTPases, the GTPase-

activating proteins (GAPs) and guanine nucleotide-exchange factors (GEFs) by suppression (GAPs) or promotion (GEFs) of activity. The small GTPases are involved in a range of fundamental processes, from gene expression, cytoskeletal reorganization to intracellular vesicle trafficking and cytokinesis (Takai *et al.*, 2001; van Dam *et al.*, 2014). In mammals, some GTPases are involved in immune signalling as they are important for TLR signal transduction, especially for TLR2 and TLR4 (Bokoch, 2005). In codfishes, those particular *TLRs* are not found (Solbakken *et al.*, 2017), but other *TLRs* might have taken on their roles. Given the fundamental roles of GTPases, it is likely that some are important for signaling in the immune system of codfishes. The STRs in the genes encoding GAPs and GEFs could modulate their function, leading to differences in signal transduction of these pathways between populations of codfishes.

Cod and haddock seem unusual in their high density and frequency of STRs. It is not clear what aspect of their biology facilitates such an expansion compared with other species. However, the Atlantic herring also has a high amount of STRs (Supplementary File E in Barrio *et al.* (2016)). Cod, herring and haddock are all species with high fecundity and large population sizes (Barrio *et al.*, 2016) (Work V). Their genome sizes are moderate, around 700-900 Mbp. There have been indications that there is a negative correlation between genome size in teleosts and effective population (Yi and Streelman, 2005), but this disappears when correcting for phylogeny (Whitney and Garland Jr, 2010). There is a negative correlation between fecundity and egg size, with smaller eggs at higher fecundity (Sargent *et al.*, 1987). Further, there is a correlation between genome size and egg size (Hardie and Hebert, 2004). Thus, genome size and fecundity might influence the high density and frequency of STRs. There is no clear consensus of how STRs originate (*de novo*). They might be induced by transposon activity, but a high density of transposons does not necessarily lead to a high density of STRs (Oliveira *et al.*, 2006). However, it seems that as soon as there is a repeated sequence (two dinucleotides in tandem for instance), slippage in replication can then extend the STR. We are not aware of any studies showing a tendency for higher polymerase slippage in codfishes. Besides the origin of the high STR content, the maintenance of STRs also provides a puzzle: both cod, haddock and herring have large effective population sizes (Barrio *et al.*, 2016) (Work V) and should therefore be effective in removing deleterious or slightly deleterious alleles. Most likely, the large amount of STRs is either advantageous or neutral for these species.

STRs, like most other elements of the genome, can be advantageous, deleterious, or neutral. Some studies highlight the functional role of STRs in "cre-

ating and maintaining quantitative genetic variation" (Ohadi *et al.*, 2015), however, this might not accurately represent the all STRs. Likely for as long as we have been aware of them, investigators have discussed the role of STRs in genic regions and non-genic regions (King *et al.*, 1997; Sawyer *et al.*, 1997). Junk DNA has in the last decades been defined as the "fraction of the genome on which selection does not operate" (Graur, 2016) and it has been claimed that "a substantial percentage of the DNA in many eukaryotic genomes lacks an organism-level function" (Palazzo and Gregory, 2014). Most STRs in eukaryotic genomes do not play any role in "organism-level function". For instance, of 700,000 STRs in the human genome, only 4,500 are in protein coding regions (Willems *et al.*, 2014). In addition, 2,060 have been found to be significantly associated with expression of genes (Gymrek *et al.*, 2016). While this adds up to about 6,500 STRs in the human genome that can have direct effect on the sequence of a protein or the expression of the gene encoding the protein, this represents only about 1 % of the overall number of STRs. However, this is a substantial fraction of the total amount of genes in the human genome, about one third. Therefore, while STRs play a detrimental role in some human diseases (Orr and Zoghbi, 2007), or can be important for many genes, most STRs in the human genome are neutral. The genomes of codfishes are smaller than the human genome, and likely contains a larger fraction of DNA with "organism-level function". Given that there is a higher density and frequency of STRs in the codfish genomes, a higher fraction of them might be under selection than in the human genome. For instance, STRs are found in almost twice the amount of genes in codfishes compared to humans (8,000 vs 4,500 genes). However, many STRs would likely not be under selection, and therefore could be categorised as junk DNA.

4.4 The immune gene repertoire of species with an unusual immune system

We mapped the presence/absence of relevant genes on a phylogeny (Work IV) based on 76 species (the 66 new assemblies (Work III) plus 10 previously available assemblies) and found that the loss of the *MHCII* genes happened approximately 105 millions years ago in the lineage that became the Gadiformes of today. The absence of *MHCII* genes in Gadiformes has likely been compensated by several mechanisms. Firstly, many of the Gadiformes have a vastly expanded repertoire of *MHCI* genes (Work IV), with up to 100 gene copies in cod itself (Star *et al.* (2011) and Work IV). Within the pattern recognition recep-

tors (PRRs) family of Toll-like receptors genes (*TLRs*) some members are lost and others are expanded (Solbakken *et al.*, 2016b), with some expansions correlated to *MHCII* loss and species latitudinal distributions (Solbakken *et al.*, 2017). There is a high number of *TLRs* in species such as purple sea urchin (*Strongylocentrotus purpuratus*) and Florida lancet (*Branchiostoma floridae*) which could be connected to their lack of an adaptive immune system (Rast *et al.*, 2006; Hibino *et al.*, 2006; Huang *et al.*, 2008). Another PRR family, the NOD-like receptors (NLRs), has also been discussed in connection with large numbers of genes in species without an adaptive immune system such as cnidarians (Lange *et al.*, 2011) and the purple sea urchin (Rast *et al.*, 2006).

However, there is no clear connection between increases in number of *NLRs* and the loss of MHCII pathway genes in codfishes. While the *NLRs* are few in numbers in mammals (Stein *et al.*, 2007), there have been reports of 400 NLRs in zebrafish, 100 in Mexican tetra (cavefish or *Astyanax mexicanus*) and 50 in Northern pike (*Esox lucius*) (Howe *et al.*, 2016), 70 in fugu and 49 in tetraodon (Stein *et al.*, 2007), and 50 in the miiuy croaker (*Miichthys miiuy*) (Xu *et al.*, 2016). All of these species have genes necessary for the MHCII pathway. In Work V, we found that the *NLRs* are likely expanded in all teleosts, and possible lineage specific expansions in zebrafish, stickleback, tilapia (*Oreochromis niloticus*) and the codfishes. It is likely that the numbers of *NLR* genes are underestimated for all species. For instance, in cod about half the scaffolds with *NLRs* are unplaced in the linkage map, and have much higher average read coverage than the placed scaffolds (Work V), while in zebrafish only about 10 % of the scaffolds with putative *NLRs* are unplaced. Also, when investigating the unitig assemblies for cod (CA454PB) and haddock, the draft assemblies that contributed to the final assemblies, much higher numbers of putative *NLRs* are found in both species, around 600 copies. This is about three and nine times as many found in the final assemblies for cod and haddock, respectively. It is therefore difficult to confidently suggest a certain number of copies of *NLRs* in the assemblies of different fishes, aside from stating that the numbers are likely underestimated. Further improved assemblies for these species are needed to properly investigate the intriguing nature of these genes, not least for understanding how they propagate in the genome. This is continuing process, where new technologies and new assembly approaches extend how much of the genome is actually reconstructed in a genome assembly.

4.5 Genome assemblies are crucial for understanding biology

Biologi is the study of life, with the goal to understand different aspects of living creatures. One aspect of this is to understand the function and roles of the genes and the proteins they encode in different species (Pavey *et al.*, 2012). A genome assembly is useful as a starting point for this. By annotating the genes found on the assembly, we can get a catalogue of genes. However, the function of these are often unknown. Even for a species such as yeast, with more researchers than the number of genes (Peña-Castillo and Hughes, 2007), many genes are still uncharacterized, with unknown function (apparently 1,000 in 2007 (Peña-Castillo and Hughes, 2007), with many of these having a predicted function, but no experimental evidence (Eisenhaber, 2012)). For species less suited for experimental manipulation, and with a higher number of genes, such as humans, about half the genes/proteins lack functional characterization (Eisenhaber, 2012). This means that none these genes/proteins have a function that is known in humans, nor have their homologs known functions in other species. For humans and yeast we have good knowledge of the genomics sequence, and know exactly which predicted genes/proteins lack functional characterization.

How should we evaluate the situation for cod? While it is likely that most genes are structurally annotated and therefore have a known location, they might lack a proper functional characterization. Some of these might also be fragmented. Most teleost fishes annotated at Ensembl have around 20,000 genes, while gadMor2 was annotated to 23,000 (Work II, up from around 20,000 for gadMor1). It is possible that some of these are divided across multiple contigs/scaffolds, and this therefore increases the count (Denton *et al.*, 2014). Another issue is that of the 600 putative NLR genes, only 200 are found in the final assembly (Work V). It is not possible to assess the function of these multiple gene copies when they cannot even be found in the assembly. More complete genome assemblies are therefore needed.

While costly, highly contiguous assemblies that surpass even the earlier Sanger-based assemblies in quality can be created using sufficient coverage of PacBio reads (Bickhart *et al.*, 2017; Vij *et al.*, 2016; Warren *et al.*, 2016). A combination with longer range continuity information as generated by optical mapping (Howe and Wood, 2015; Seo *et al.*, 2016), chromosome conformation (Bickhart *et al.*, 2017) data, linked reads (Yeo *et al.*, 2017; Weisenfeld *et al.*, 2016), or combinations of these (Jiao *et al.*, 2017; Mostovoy *et al.*, 2016) would yield assemblies that are almost complete. A broad sampling of such assemblies across the teleosts would give much information about how multi-copy gene families

proliferate and spread in genomes, and facilitate investigations into how STRs originate and spread, thereby laying the groundwork for better understanding the biology of various fish species.

5 Concluding remarks and future perspectives

During their evolution, codfishes have changed fundamental aspects of their biology to something different from most other vertebrate species. They have reorganized their immune system with losses of genes previously thought to be ubiquitous among jawed vertebrates (*MHCII*, *CD4* and *Invariant chain*) with expansions of other gene families (*TLRs*, *NLRs*, *MHCIs*). They have an increased density and frequency of STRs compared to other species, inside and outside of protein coding sequence. It is striking that both of these features exists in the same group of species, and it is tempting to suggest that there might be some connections between them. For instance, there is enrichment of STRs in signal transduction genes in these species, and that might be one connection to the immune genes. In addition, codfishes have high fecundity and large effective population sizes, which would lead to even weak selection being effective, leading to local adaptations possibly based on the STR variation. The interconnections between these aspects of the codfishes require more investigation.

The present genome assemblies for cod and haddock reveal limitations when considering cases such as the NLRs; this is also the case all investigated teleost genome assemblies. These high copy-number genes collapse into fewer sequences during the assembly process, and impede proper investigation of their nature. Because of the collapse, surrounding sequence of the genes and therefore synteny between species is unavailable. This makes inference into the evolution of these genes difficult. The lack of properly updated genome assemblies also impedes this. For instance, we downloaded the assemblies we did not generate ourselves from a very convenient resource, the Ensembl database. However, many of the assemblies present there have not been updated for several years. For some, newer assemblies do exist, such as stickleback (Peichel *et al.*, 2016) and fugu. The new stickleback assembly is not found at Ensembl (N.B. neither is gadMor2 because funds have been prioritised to other purposes), but the consequence is that researchers will use an older assembly, possibly of lower quality. The newest fugu assembly, FUGU5, was produced in October 2011, and has been in Ensemble Pre! since then, i.e. not on the main website. Many genome projects often produce their own annotations, such as our own, and these are therefore not standardized. While this perspective deals with teleosts, these issues are likely affects other species across the tree of life as well. There is a need for an up-to-date, standardized resource, and hopefully initiatives such as Genome 10K (Bernardi *et al.*, 2012; Koepfli *et al.*, 2015) might facilitate this. It is crucial for comparative analy-

ses, as well. For a few selected species (mouse, human, chicken and zebrafish; https://www.ncbi.nlm.nih.gov/grc) there is an ongoing effort to improving their genome assemblies with new sequencing technologies and assembly approaches. However, as argued in this thesis, improved and updated genome assemblies for other non-model species will reveal fascinating aspects of their biology and evolution.

6 References

Adams, R. H., Blackmon, H., Reyes-Velasco, J., Schield, D. R., Card, D. C., Andrew, A. L., Waynewood, N., and Castoe, T. A., 2016. Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. *Genome*, **59**(5), 295–310.

Albà, M. M., Santibáñez-Koref, M. F., and Hancock, J. M., 1999. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process. *Journal of Molecular Evolution*, **49**(6), 789–797.

Alhakami, H., Mirebrahim, H., and Lonardi, S., 2017. A comparative evaluation of genome assembly reconciliation tools. *Genome Biology*, **18**(1), 93.

Alic, A. S., Ruzafa, D., Dopazo, J., and Blanquer, I., 2016. Objective review of *de novo* stand-alone error correction methods for NGS data. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **6**(2), 111–146.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M. D. S., Roach, J., Oh, T., Ho, I. Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S. F., Clark, M. S., Edwards, Y. J. K., Doggett, N., Zharkikh, A., Tavtigian, S. V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y. H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., and Brenner, S., 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297**(5585), 1301–1310.

Bansal, V., Bashir, A., and Bafna, V., 2007. Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Research*, **17**(2), 219–230.

Barrio, A. M., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman, D., Höppner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma, K., Zhan, X., Feng, C., Gustafson, U., Rubin, C.-J., Almén, M. S., Blass, M., Casini, M., Folkvord, A., Laikre, L., Ryman, N., Lee, S. M.-Y., Xu, X., Andersson, L., and Nordborg, M., 2016. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife*, **5**, e12081.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell,

J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**(7218), 53–59.

Bergman, C. M. and Quesneville, H., 2007. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, **8**(6), 382–392.

Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., and Phillippy, A. M., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, **33**(6), 623–630.

Bernardi, G., Wiley, E. O., Mansour, H., Miller, M. R., Orti, G., Haussler, D., O'Brien, S. J., Ryder, O. A., and Venkatesh, B., 2012. The fishes of Genome 10K. *Marine Genomics*, **7**, 3–6.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., Lee, J., Lam, E. T., Liachko, I., Sullivan, S. T., Burton, J. N., Huson, H. J., Nystrom, J. C., Kelley, C. M., Hutchison, J. L., Zhou, Y., Sun, J., Crisà, A., de León, F. A. P., Schwartz, J. C., Hammond, J. A., Waldbieser, G. C., Schroeder, S. G., Liu, G. E., Dunham, M. J., Shendure, J., Sonstegard, T. S., Phillippy, A. M., Van Tassell, C. P., and Smith, T. P. L., 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, **431**, 931.

Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., Yuen, M. M. S., Keeling, C. I., Brand, D., Vandervalk, B. P., Kirk, H., Pandoh, P., Moore, R. A., Zhao, Y., Mungall, A. J., Jaquish, B., Yanchuk, A., Ritland, C., Boyle, B., Bousquet, J., Ritland, K., MacKay, J., Bohlmann, J., and Jones, S. J. M., 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics*, **29**(12), 1492–1497.

Bokoch, G. M., 2005. Regulation of innate immunity by Rho GTPases. *Trends in Cell Biology*, **15**(3), 163–171.

Braasch, I., Peterson, S. M., Desvignes, T., McCluskey, B. M., Batzel, P., and Postlethwait, J. H., 2015. A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, **324**(4), 316–341.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.-C., Corbeil, J., Del Fabbro, C., Docking, T., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T.-W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., MacManes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto,

T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D., Ruby, J., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S.-M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., and Korf, I. F., 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, **2**(1), 10.

Campbell, M. S., Holt, C., Moore, B., and Yandell, M., 2014a. Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics*, **48**, 4.11.1–4.11.39.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S.-H., Childs, K. L., Sun, Y., Jiang, N., and Yandell, M., 2014b. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology*, **164**(2), 513–524.

Canapa, A., Barucca, M., Biscotti, M. A., Forconi, M., and Olmo, E., 2016. Transposons, genome size, and evolutionary insights in animals. *Cytogenetic and Genome Research*, **0**(0).

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.-N., 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, **7**(2), 567–580.

Charlesworth, B., 2009. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**(3), 195–205.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., and Korlach, J., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, **10**(6), 563–569.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., and Schatz, M. C., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, **13**(12), 1050–1054.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and

Mortazavi, A., 2016. A survey of best practices for RNA-seq data analysis. *Genome Biology*, **17**(1), 13.

Conte, M. A. and Kocher, T. D., 2015. An improved genome reference for the African cichlid, Metriaclima zebra. *BMC Genomics*, **16**(1), 1.

Cortesi, F., Musilová, Z., Stieb, S. M., Hart, N. S., Siebeck, U. E., Malmstrøm, M., Tørresen, O. K., Jentoft, S., Cheney, K. L., Marshall, N. J., Carleton, K. L., and Salzburger, W., 2015. Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes. *Proceedings of the National Academy of Sciences*, **112**(5), 1493–1498.

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., and Hahn, M. W., 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Computational Biology*, **10**(12), e1003998.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H., 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**(6), 446–450.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. *Science*, **323**(5910), 133–138.

Eisenhaber, F., 2012. A decade after the first full human genome sequencing: when will we understand our own genome? *Journal of Bioinformatics and Computational Biology*, **10**(5), 1271001.

Ekblom, R. and Wolf, J. B. W., 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, **7**(9), 1026–1042.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**(6), 435–445.

Ellegren, H., 2014. Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**(1), 51–63.

Elliott, T. A. and Gregory, T. R., 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society of London Series B: Biological sciences*, **370**(1678), 20140331.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., and Gibbs, R. A., 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, **7**(11), e47768.

Eschmeyer, W. N., Fricke, R., and van der Laan, R., 2017. Catalog of fishes: genera, species, references.

Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., and Trask, B. J., 2002. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13-2q14.1 and paralogous regions on other human chromosomes. *Genome Research*, **12**(11), 1651–1662.

FAO, 2016. The State of World Fisheries and Aquaculture 2016, 1–204.

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C., 2012. Tools for mapping high-throughput sequencing data. *Bioinformatics*, **28**(24), 3169–3177.

Ganapathy, G., Howard, J. T., Ward, J. M., Li, J., Li, B., Li, Y., Xiong, Y., Zhang, Y., Zhou, S., Schwartz, D. C., Schatz, M., Aboukhalil, R., Fedrigo, O., Bukovnik, L., Wang, T., Wray, G., Rasolonjatovo, I., Winer, R., Knight, J. R., Koren, S., Warren, W. C., Zhang, G., Phillippy, A. M., and Jarvis, E. D., 2014. High-coverage sequencing and annotated assemblies of the budgerigar genome. *GigaScience*, **3**(1), 11.

Glenn, T. C., 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**(5), 759–769.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., Berlin, A. M., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B., 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(4), 1513–1518.

Goodwin, S., McPherson, J. D., and McCombie, W. R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**(6), 333–351.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**(7), 644–652.

Graur, D., 2016. Rubbish DNA: The functionless fraction of the human genome. *arXiv.org*.

Gregory, T. R., 2005. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics*, **6**(9), 699–708.

Gymrek, M., 2017. A genomic view of short tandem repeats. *Current Opinion in Genetics & Development*, **44**, 9–16.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M. J., Price, A. L., Pritchard, J. K., Sharp, A. J., and Erlich, Y., 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*, **48**(1), 22–29.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., and White, O., 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**(19), 5654–5666.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R., 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, **9**(1), R7.

Hardie, D. C. and Hebert, P. D., 2004. Genome-size evolution in fishes. *Canadian Journal of Fisheries and Aquatic Sciences*, **61**(9), 1636–1646.

Hefferon, T. W., Groman, J. D., Yurk, C. E., and Cutting, G. R., 2004. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(10), 3504–3509.

Hibino, T., Loza-Coll, M., Messier, C., Majeske, A. J., Cohen, A. H., Terwilliger, D. P., Buckley, K. M., Brockton, V., Nair, S. V., Berney, K., Fugmann, S. D., Anderson, M. K., Pancer, Z., Cameron, R. A., Smith, L. C., and Rast, J. P.,

2006. The immune gene repertoire encoded in the purple sea urchin genome. *Developmental Biology*, **300**(1), 349–365.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M., 2016. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**(5), 767–769.

Hoff, K. J. and Stanke, M., 2015. Current methods for automated annotation of protein-coding genes. *Current Opinion in Insect Science*, **7**, 8–14.

Holt, C. and Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**(1), 491.

Honaas, L. A., Wafula, E. K., Wickett, N. J., Der, J. P., Zhang, Y., Edger, P. P., Altman, N. S., Pires, J. C., Leebens-Mack, J. H., and dePamphilis, C. W., 2016. Selecting superior *de novo* transcriptome assemblies: lessons learned by leveraging the best plant genome. *PLoS ONE*, **11**(1), e0146062.

Howe, K., Schiffer, P. H., Zielinski, J., Wiehe, T., Laird, G. K., Marioni, J. C., Soylemez, O., Kondrashov, F., and Leptin, M., 2016. Structure and evolutionary history of a large family of NLR proteins in the zebrafish. *Open Biology*, **6**(4), 160009–224.

Howe, K. and Wood, J. M., 2015. Using optical mapping data for the improvement of vertebrate genome assemblies. *GigaScience*, **4**(1), 10.

Huang, S., Yuan, S., Guo, L., Yu, Y., Li, J., Wu, T., Liu, T., Yang, M., Wu, K., Liu, H., Ge, J., Yu, Y., Huang, H., Dong, M., Yu, C., Chen, S., and Xu, A., 2008. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Research*, **18**(7), 1112–1126.

Hubert, S., Higgins, B., Borza, T., and Bowman, S., 2010. Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics*, **11**(1), 191.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D., 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biology*, **14**(5), R47.

Huntley, M. A. and Clark, A. G., 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 Drosophila species. *Molecular Biology and Evolution*, **24**(12), 2598–2609.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011), 931–945.

Jiang, Q., Li, Q., Yu, H., and Kong, L., 2014. Genome-wide analysis of simple sequence repeats in marine animals—a comparative approach. *Marine Biotechnology*, **16**(5), 604–619.

Jiao, W.-B., Accinelli, G. G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E.-M., Piednoel, M., Woetzel, S., Madrid-Herrero, E., Huettel, B., Hümann, U., Reinhard, R., Koch, M. A., Swan, D., Clavijo, B., Coupland, G., and Schneeberger, K., 2017. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Research*, **27**(5), 778–786.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S., 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**(9), 1236–1240.

Kamath, G. M., Shomorony, I., Xia, F., Courtade, T. A., and Tse, D. N., 2017. HINGE: long-read assembly achieves optimal repeat resolution. *Genome Research*, **27**(5), 747–756.

King, D. G., Soller, M., and Kashi, Y., 1997. Evolutionary tuning knobs. *Endeavour*, **21**(1), 36–40.

Koepfli, K.-P., Paten, B., Genome 10K Community of Scientists, and O'Brien, S. J., 2015. The Genome 10K Project: a way forward. *Annual Review of Animal Biosciences*, **3**(1), 57–111.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Phillippy, A. M., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**(7), 693–700.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, gr.215087.116.

Korf, I. F., 2004. Gene finding in novel genomes. *BMC Bioinformatics*, **5**(1), 59.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T.,

Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Lange, C., Hemmrich, G., Klostermeier, U. C., López-Quintero, J. A., Miller, D. J., Rahn, T., Weiss, Y., Bosch, T. C. G., and Rosenstiel, P., 2011. Defining the origins of the NOD-like receptor system at the base of animal evolution. *Molecular Biology and Evolution*, **28**(5), 1687–1702.

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K. J., 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research*, **17**(12), 1787–1796.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., Hvidsten, T. R., Leong, J. S., Minkley, D. R., Zimin, A., Grammes, F., Grove, H., Gjuvsland, A., Walenz, B. P., Hermansen, R. A., von Schalburg, K., Rondeau, E. B., Di Genova, A., Samy, J. K. A., Olav Vik, J., Vigeland, M. D., Caler, L., Grimholt, U., Jentoft, S., Våge, D. I., de Jong, P., Moen, T., Baranski, M., Palti, Y., Smith, D. R., Yorke, J. A., Nederbragt, A. J., Tooming-Klunderud, A., Jakobsen, K. S., Jiang, X., Fan, D., Hu, Y., Liberles, D. A., Vidal, R., Iturra, P., Jones, S. J. M., Jonassen, I., Maass, A., Omholt, S. W., and Davidson, W. S., 2016. The Atlantic salmon genome provides insights into rediploidization. *Nature*, **533**(7602), 200–205.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M., 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, **33**(20), 6494–6506.

Malmstrøm, M., Jentoft, S., Gregers, T. F., and Jakobsen, K. S., 2013. Unraveling the evolution of the Atlantic cod's (*Gadus morhua* L.) alternative immune strategy. *PLoS ONE*, **8**(9), e74004.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Jakobsen, K. S., and Jentoft, S., 2017. Whole genome sequencing data and de novo draft assemblies for 66 teleost species. *Scientific Data*, **4**, 160132.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., Baalsrud, H. T., Nederbragt, A. J., Hanel, R., Salzburger, W., Stenseth, N. C., Jakobsen, K. S., and Jentoft, S., 2016. Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, **48**(10), 1204–1210.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**(7057), 376–380.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., Rüeger, S., Thorleifsson, G., Justice, A. E., Lamparter, D., Stirrups, K. E., Turcot, V., Young, K. L., Winkler, T. W., Esko, T., Karaderi, T., Locke, A. E., Masca, N. G. D., Ng, M. C. Y., Mudgal, P., Rivas, M. A., Vedantam, S., Mahajan, A., Guo, X., Abecasis, G., Aben, K. K., Adair, L. S., Alam, D. S., Albrecht, E., Allin, K. H., Allison, M., Amouyel, P., Appel, E. V., Arveiler, D., Asselbergs, F. W., Auer, P. L., Balkau, B., Banas, B., Bang, L. E., Benn, M., Bergmann, S., Bielak, L. F., Blüher, M., Boeing, H., Boerwinkle, E., Böger, C. A., Bonnycastle, L. L., Bork-Jensen, J., Bots, M. L., Bottinger, E. P., Bowden, D. W., Brandslund, I., Breen, G., Brilliant, M. H., Broer, L., Burt, A. A., Butterworth, A. S., Carey, D. J., Caulfield, M. J., Chambers, J. C., Chasman, D. I., Chen, Y.-D. I., Chowdhury, R., Christensen, C., Chu, A. Y., Cocca, M., Collins, F. S., Cook, J. P., Corley, J., Galbany, J. C., Cox, A. J., Cuellar-Partida, G., Danesh, J., Davies, G., de Bakker, P. I. W., de Borst, G. J., de Denus, S., de Groot, M. C. H., de Mutsert, R., Deary, I. J., Dedoussis, G., Demerath, E. W., den Hollander, A. I., Dennis, J. G., Di Angelantonio, E., Drenos, F., Du, M., Dunning, A. M., Easton, D. F., Ebeling, T., Edwards, T. L., Ellinor, P. T., Elliott, P., Evangelou, E., Farmaki, A.-E., Faul, J. D., Feitosa, M. F., Feng, S., Ferrannini, E., Ferrario, M. M., Ferrieres, J., Florez, J. C., Ford, I., Fornage, M., Franks, P. W., Frikke-Schmidt, R., Galesloot, T. E., Gan, W., Gandin, I., Gasparini, P., Giedraitis, V., Giri, A.,

Girotto, G., Gordon, S. D., Gordon-Larsen, P., Gorski, M., Grarup, N., Grove, M. L., Gudnason, V., Gustafsson, S., Hansen, T., Harris, K. M., Harris, T. B., Hattersley, A. T., Hayward, C., He, L., Heid, I. M., Heikkilä, K., Helgeland, Ø., Hernesniemi, J., Hewitt, A. W., Hocking, L. J., Hollensted, M., Holmen, O. L., Hovingh, G. K., Howson, J. M. M., Hoyng, C. B., Huang, P. L., Hveem, K., Ikram, M. A., Ingelsson, E., Jackson, A. U., Jansson, J.-H., Jarvik, G. P., Jensen, G. B., Jhun, M. A., Jia, Y., Jiang, X., Johansson, S., Jørgensen, M. E., Jørgensen, T., Jousilahti, P., Jukema, J. W., Kahali, B., Kahn, R. S., Kähönen, M., Kamstrup, P. R., Kanoni, S., Kaprio, J., Karaleftheri, M., Kardia, S. L. R., Karpe, F., Kee, F., Keeman, R., Kiemeney, L. A., Kitajima, H., Kluivers, K. B., Kocher, T., Komulainen, P., Kontto, J., Kooner, J. S., Kooperberg, C., Kovacs, P., Kriebel, J., Kuivaniemi, H., Küry, S., Kuusisto, J., La Bianca, M., Laakso, M., Lakka, T. A., Lange, E. M., Lange, L. A., Langefeld, C. D., Langenberg, C., Larson, E. B., Lee, I.-T., Lehtimäki, T., Lewis, C. E., Li, H., Li, J., Li-Gao, R., Lin, H., Lin, L.-A., Lin, X., Lind, L., Lindström, J., Linneberg, A., Liu, Y., Liu, Y., Lophatananon, A., Luan, J., Lubitz, S. A., Lyytikäinen, L.-P., Mackey, D. A., Madden, P. A. F., Manning, A. K., Männistö, S., Marenne, G., Marten, J., Martin, N. G., Mazul, A. L., Meidtner, K., Metspalu, A., Mitchell, P., Mohlke, K. L., Mook-Kanamori, D. O., Morgan, A., Morris, A. D., Morris, A. P., Müller-Nurasyid, M., Munroe, P. B., Nalls, M. A., and Nauck, M., 2017. Rare and low-frequency coding variants alter human adult height. *Nature*, **542**(7640), 186–190.

Mayer, C., Leese, F., and Tollrian, R., 2010. Genome-wide analysis of tandem repeats in *Daphnia pulex* - a comparative approach. *BMC Genomics*, **11**, 277.

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. G., 2008. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**(24), 2818–2824.

Miller, J. R., Koren, S., and Sutton, G. G., 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, **95**(6), 315–327.

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., Lee, J., Chu, C., Lin, C., Džakula, Ž., Cao, H., Schlebusch, S. A., Giorda, K., Schnall-Levin, M., Wall, J. D., and Kwok, P.-Y., 2016. A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, **13**(7), 587–590.

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R.,

Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I. F., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P. A., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E. M., Zody, M. C., and Lander, E. S., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), 520–562.

Mularoni, L., Ledda, A., Toll-Riera, M., and Albà, M. M., 2010. Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Research*, **20**(6), 745–754.

Müller, A., Oertli, M., and Arnold, I. C., 2011. H. pylori exploits and manipulates innate and adaptive immune cell signaling pathways to establish persistent infection. *Cell Communication and Signaling*, **9**(1), 25.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I., Fasulo, D., Flanigan, M., Kravitz, S., Mobarry, C., Reinert, K. H., Remington, K., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C., 2000. A whole-genome assembly of Drosophila. *Science*, **287**(5461), 2196–2204.

Nadalin, F., Vezzi, F., and Policriti, A., 2012. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics*, **13**(Suppl 14), S8.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, Å., Rilakovic, N., Ritland, C., Rosselló, J. A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T. R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Lee Thompson, S., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P. K., Lundeberg, J., and Jansson, S., 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**(7451), 579–584.

Ohadi, M., Valipour, E., Ghadimi Haddadan, S., Namdar Aligoodarzi, P., Bagheri, A., Kowsari, A., Rezazadeh, M., Darvish, H., and Kazeminasab, S., 2015. Core promoter short tandem repeats as evolutionary switch codes for primate speciation. *American Journal of Primatology*, **77**(1), 34–43.

Olasagasti, F., Lieberman, K. R., Benner, S., Cherf, G. M., Dahl, J. M., Deamer, D. W., and Akeson, M., 2010. Replication of individual DNA molecules under electronic control using a protein nanopore. *Nature Nanotechnology*, **5**(11), 798–806.

Oliveira, E. J., Pádua, J. G., Zucchi, M. I., Vencovsky, R., Vieira, M. L. C., and Universidade de São Paulo, Brasil, 2006. Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, **29**(2), 294–307.

Olsen, E., Aanes, S., Mehl, S., Holst, J. C., Aglen, A., and Gjosaeter, H., 2010. Cod, haddock, saithe, herring, and capelin in the Barents Sea and adjacent waters: a review of the biological value of the area. *ICES Journal of Marine Science*, **67**(1), 87–101.

Orr, H. T. and Zoghbi, H. Y., 2007. Trinucleotide repeat disorders. *Annual Review of Neuroscience*, **30**(1), 575–621.

Palazzo, A. F. and Gregory, T. R., 2014. The case for junk DNA. *PLoS Genetics*, **10**(5), e1004351.

Parra, G., Bradnam, K. R., and Korf, I. F., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**(9), 1061–1067.

Parra, G., Bradnam, K. R., Ning, Z., Keane, T., and Korf, I. F., 2009. Assessing the gene space in draft genomes. *Nucleic Acids Research*, **37**(1), 289–297.

Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E., 2017. Genome graphs and the evolution of genome inference. *Genome Research*, **27**(5), 665–676.

Pavey, S. A., Bernatchez, L., Aubin-Horth, N., and Landry, C. R., 2012. What is needed for next-generation ecological and evolutionary genomics? *Trends in Ecology & Evolution*, **27**(12), 673–678.

Peichel, C. L., Sullivan, S. T., Liachko, I., and White, M. A., 2016. Improvement of the threespine stickleback (*Gasterosteus aculeatus*) genome using a Hi-C-based Proximity-Guided Assembly method. *bioRxiv*, 068528.

Peña-Castillo, L. and Hughes, T. R., 2007. Why are there still over 1000 uncharacterized yeast genes? *Genetics*, **176**(1), 7–14.

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., Guo, Y., Paxinos, E. E., Korbel, J. O., Darnell, R. B., McCombie, W. R., Kwok, P.-Y., Mason, C. E., Schadt, E. E., and Bashir, A., 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, **12**(8), 780–786.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., and Salzberg, S. L., 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc*, **11**(9), 1650–1667.

Pilström, L., Warr, G. W., and Strömberg, S., 2005. Why is the antibody response of Atlantic cod so poor? The search for a genetic explanation. *Fisheries Science*, **71**(5), 961–971.

Press, M. O., Carlson, K. D., and Queitsch, C., 2014. The overdue promise of short tandem repeat variation for heritability. *Trends in Genetics*, **30**(11), 504–512.

Press, M. O., McCoy, R. C., Hall, A. N., Akey, J. M., and Queitsch, C., 2017. Short tandem repeats with massive variation and functional consequences across strains of Arabidopsis thaliana. *bioRxiv*, 145128.

Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R. S., Mittelman, D., and Sharp, A. J., 2016. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Research*, **44**(8), 3750–3762.

Rasko, D. A., Webster, D. R., Sahl, J. W., Bashir, A., Boisen, N., Scheutz, F., Paxinos, E. E., Sebra, R., Chin, C.-S., Iliopoulos, D., Klammer, A., Peluso, P., Lee, L., Kislyuk, A. O., Bullard, J., Kasarskis, A., Wang, S., Eid, J., Rank, D., Redman, J. C., Steyert, S. R., Frimodt-Møller, J., Struve, C., Petersen, A. M., Krogfelt, K. A., Nataro, J. P., Schadt, E. E., and Waldor, M. K., 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *The New England Journal of Medicine*, **365**(8), 709–717.

Rast, J. P., Smith, L. C., Loza-Coll, M., Hibino, T., and Litman, G. W., 2006. Genomic insights into the immune system of the sea urchin. *Science*, **314**(5801), 952–956.

Rastas, P., Calboli, F. C. F., Guo, B., Shikano, T., and Merilä, J., 2015. Construction of ultra-dense linkage maps with Lep-MAP2: stickleback F2 recombinant crosses as an example. *Genome Biology and Evolution*, **8**(1), evv250–93.

Rhoads, A. and Au, K. F., 2015. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, **13**(5), 278–289.

Ryan, J. F., 2013. Baa. pl: A tool to evaluate de novo genome assemblies with RNA transcripts. *arXiv.org*.

Sargent, R. C., Taylor, P. D., and Gross, M. R., 1987. Parental care and the evolution of egg size in fishes. *American Naturalist*.

Sawaya, S., Jones, M., and Keller, M., 2015. Linkage disequilibrium between single nucleotide polymorphisms and hypermutable loci. *bioRxiv*, 020909.

Sawyer, L. A., Hennessy, J. M., Peixoto, A. A., Rosato, E., Parkinson, H., Costa, R., and Kyriacou, C. P., 1997. Natural variation in a Drosophila clock gene and temperature compensation. *Science*, **278**(5346), 2117–2120.

Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, J., Kuk, J., Park, G. H., Kim, J., Ryu, H., Kim, J., Roh, M., Baek, J., Hunkapiller, M. W., Korlach, J., Shin, J.-Y., and Kim, C., 2016. De novo assembly and phasing of a Korean human genome. *Nature*, **538**(7624), 243–247.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**(19), 3210–3212.

Simon, M. and Hancock, J. M., 2009. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology*, **10**(6), R59.

Simpson, J. T. and Durbin, R., 2011. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, **22**(3), gr.126953.111–556.

Simpson, J. T. and Pop, M., 2015. The theory and practice of genome sequence assembly. *Annual Review of Genomics and Human Genetics*, **16**(1), 153–172.

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J., and Kelly, S., 2016. TransRate: reference free quality assessment of de novo transcriptome assemblies. *Genome Research*, **26**(8), gr.196469.115–1144.

Solbakken, M. H., Rise, M. L., Jakobsen, K. S., and Jentoft, S., 2016a. Successive losses of central immune genes characterize the Gadiformes' alternate immunity. *Genome Biology and Evolution*, **8**(11), 3508–3515.

Solbakken, M. H., Tørresen, O. K., Nederbragt, A. J., Seppola, M., Gregers, T. F., Jakobsen, K. S., and Jentoft, S., 2016b. Evolutionary redesign of the Atlantic cod (*Gadus morhua* L.) Toll-like receptor repertoire by gene losses and expansions. *Scientific Reports*, **6**(1), 25211.

Solbakken, M. H., Voje, K. L., Jakobsen, K. S., and Jentoft, S., 2017. Linking species habitat and past palaeoclimatic events to evolution of the teleost innate immune system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **284**(1853), 20162810.

Sonay, T. B., Carvalho, T., Robinson, M., Greminger, M., Krutzen, M., Comas, D., Highnam, G., Mittelman, D. A., Sharp, A. J., Marques-Bonet, T., and Wagner, A., 2015. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Research*, **25**(11), gr.190868.115–1599.

Sotero-Caio, C. G., Platt, R. N., Suh, A., and Ray, D. A., 2017. Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biology and Evolution*, **9**(1), 161–177.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**(5), 637–644.

Star, B., Hansen, M. H., Skage, M., Bradbury, I. R., Godiksen, J. A., Kjesbu, O. S., and Jentoft, S., 2016a. Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples. *STAR: Science & Technology of Archaeological Research*, **2**(1), 36–45.

Star, B. and Jentoft, S., 2012. Why does the immune system of Atlantic cod lack MHC II? *BioEssays*, **34**(8), 648–651.

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lanzén, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R. B., Tina, K. G., Espelund, M., Nepal, C., Previti, C., Karlsen, B. O., Moum, T., Skage, M., Berg, P. R., Gjøen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S. D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S. W., Stenseth, N. C., and Jakobsen, K. S., 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, **477**(7363), 207–210.

Star, B., Tørresen, O. K., Nederbragt, A. J., Jakobsen, K. S., Pampoulie, C., and Jentoft, S., 2016b. Genomic characterization of the Atlantic cod sex-locus. *Scientific Reports*, **6**(1), 31235.

Stein, C., Caccamo, M., Laird, G., and Leptin, M., 2007. Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biology*, **8**(11), R251.

Takai, Y., Sasaki, T., and Matozaki, T., 2001. Small GTP-Binding Proteins. *Physiological Reviews*, **81**(1), 153–208.

Takeuchi, O. and Akira, S., 2010. Pattern recognition receptors and inflammation. *Cell*, **140**(6), 805–820.

Tørresen, O. K., Star, B., Jentoft, S., Jakobsen, K. S., and Nederbragt, A. J., 2016. The new era of genome sequencing using high-throughput sequencing technology: generation of the first version of the Atlantic cod genome. In S. MacKenzie and S. Jentoft, editors, *Genomics in Aquaculture*. Academic Press, pages 1–20.

Tørresen, O. K., Star, B., Jentoft, S., Reinar, W. B., Grove, H., Miller, J. R., Walenz, B. P., Knight, J., Ekholm, J. M., Peluso, P., Edvardsen, R. B., Tooming-Klunderud, A., Skage, M., Lien, S., Jakobsen, K. S., and Nederbragt, A. J., 2017. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics*, **18**(1), 95.

Treangen, T. J. and Salzberg, S. L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, **13**(1), 36–46.

Tsai, I. J., Otto, T. D., and Berriman, M., 2010. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, **11**(4), R41.

UniProt Consortium, 2015. UniProt: a hub for protein information. *Nucleic Acids Research*, **43**(Database issue), D204–12.

van Dam, T. J. P., Bos, J., and Snel, B., 2014. Evolution of the Ras-like small GTPases and their regulators. *Small GTPases*, **2**(1), 4–16.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M., 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, **27**(5), 737–746.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira,

C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S.,

Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X., 2001. The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Verstrepen, K. J., Jansen, A., Lewitter, F., and Fink, G. R., 2005. Intragenic tandem repeats generate functional variability. *Nature Genetics*, **37**(9), 986–990.

Vezzi, F., Narzisi, G., and Mishra, B., 2012. Reevaluating assembly evaluations with feature response curves: GAGE and Assemblathons. *PLoS ONE*, **7**(12), e52210.

Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., van Heusden, P., Singh, S., Thevasagayam, N. M., Prakki, S. R. S., Purushothaman, K., Saju, J. M., Jiang, J., Mbandi, S. K., Jonas, M., Tong, A. H. Y., Mwangi, S., Lau, D., Ngoh, S. Y., Liew, W. C., Shen, X., Hon, L. S., Drake, J. P., Boitano, M., Hall, R., Chin, C.-S., Lachumanan, R., Korlach, J., Trifonov, V., Kabilov, M., Tupikin, A., Green, D., Moxon, S., Garvin, T., Sedlazeck, F. J., Vurture, G. W., Gopalapillai, G., Katneni, V. K., Noble, T. H., Scaria, V., Sivasubbu, S., Jerry, D. R., O'Brien, S. J., Schatz, M. C., Dalmay, T., Turner, S. W., Lok, S., Christoffels, A., and Orban, L., 2016. Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS Genetics*, **12**(4), e1005954.

Vinces, M. D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K. J., 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, **324**(5931), 1213–1216.

Visscher, P. M., 2008. Sizing up human height variation. *Nature Genetics*, **40**(5), 489–490.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M., 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly Improvement. *PLoS ONE*, **9**(11), e112963.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A. J., Fairley, S., Heger, A., Kong, L., Ponting, C. P., Jarvis, E. D., Mello, C. V., Minx, P., Lovell, P., Velho, T. A. F., Ferris, M., Balakrishnan, C. N., Sinha, S., Blatti, C., London, S. E., Li, Y., Lin, Y.-C., George, J., Sweedler, J., Southey, B., Gunaratne, P., Watson, M., Nam,

K., Backström, N., Smeds, L., Nabholz, B., Itoh, Y., Whitney, O., Pfenning, A. R., Howard, J., Völker, M., Skinner, B. M., Griffin, D. K., Ye, L., McLaren, W. M., Flicek, P., Quesada, V., Velasco, G., López-Otín, C., Puente, X. S., Olender, T., Lancet, D., Smit, A. F. A., Hubley, R., Konkel, M. K., Walker, J. A., Batzer, M. A., Gu, W., Pollock, D. D., Chen, L., Cheng, Z., Eichler, E. E., Stapley, J., Slate, J., Ekblom, R., Birkhead, T., Burke, T., Burt, D., Scharff, C., Adam, I., Richard, H., Sultan, M., Soldatov, A., Lehrach, H., Edwards, S. V., Yang, S.-P., Li, X., Graves, T., Fulton, L., Nelson, J., Chinwalla, A., Hou, S., Mardis, E. R., and Wilson, R. K., 2010. The genome of a songbird. *Nature*, **464**(7289), 757–762.

Warren, W. C., Hillier, L. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., Markovic, C., Bouk, N., Pruitt, K. D., Thibaud-Nissen, F., Schneider, V., Mansour, T. A., Brown, C. T., Zimin, A., Hawken, R., Abrahamsen, M., Pyrkosz, A. B., Morisson, M., Fillon, V., Vignal, A., Chow, W., Howe, K., Fulton, J. E., Miller, M. M., Lovell, P., Mello, C. V., Wirthlin, M., Mason, A. S., Kuo, R., Burt, D. W., Dodgson, J. B., and Cheng, H. H., 2016. A new chicken genome assembly provides insight into avian genome structure. *G3: Genes | Genomes | Genetics*, g3.116.035923.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K. F., 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, **6**, 100.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D., and Jaffe, D. B., 2016. Direct determination of diploid genome sequences. *bioRxiv*, 070425.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., and Jaffe, D. B., 2017. Direct determination of diploid genome sequences. *Genome Research*, **27**(5), 757–767.

Wences, A. H. and Schatz, M. C., 2015. Metassembler: merging and optimizing de novo genome assemblies. *Genome Biology*, **16**(1), 1.

Whitney, K. D. and Garland Jr, T., 2010. Did genetic drift drive increases in genome complexity? *PLoS Genetics*, **6**(8), e1001080.

Willems, T., Gymrek, M., Highnam, G., 1000 Genomes Project Consortium, Mittelman, D., and Erlich, Y., 2014. The landscape of human STR variation. *Genome Research*, **24**(11), 1894–1904.

Xu, T., Xu, G., Che, R., Wang, R., Wang, Y., Li, J., Wang, S., Shu, C., Sun, Y., Liu, T., Liu, J., Wang, A., Han, J., Chu, Q., and Yang, Q., 2016. The genome of the miiuy croaker reveals well-developed innate immune and sensory systems. *Scientific Reports*, **6**, 21902.

Yandell, M. and Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, **13**(5), 329–342.

Yeo, S., Coombe, L., Chu, J., Warren, R. L., and Birol, I., 2017. ARCS: Assembly Roundup by Chromium Scaffolding. *bioRxiv*, 100750.

Yi, S. and Streelman, J. T., 2005. Genome size is negatively correlated with effective population size in ray-finned fish. *Trends in Genetics*, **21**(12), 643–646.

Zerbino, D. R. and Birney, E., 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**(5), 821–829.

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., Yang, W., Hu, J., Xiao, J., Yang, Z., Liu, Y., Xie, Q., Yu, H., Lian, J., Wen, P., Zhang, F., Li, H., Zeng, Y., Xiong, Z., Liu, S., Zhou, L., Huang, Z., An, N., Wang, J., Zheng, Q., Xiong, Y., Wang, G., Wang, B., Wang, J., Fan, Y., da Fonseca, R. R., Alfaro-Núñez, A., Schubert, M., Orlando, L., Mourier, T., Howard, J. T., Ganapathy, G., Pfenning, A., Whitney, O., Rivas, M. V., Hara, E., Smith, J., Farré, M., Narayan, J., Slavov, G., Romanov, M. N., Borges, R., Machado, J. P., Khan, I., Springer, M. S., Gatesy, J., Hoffmann, F. G., Opazo, J. C., Håstad, O., Sawyer, R. H., Kim, H., Kim, K.-W., Kim, H. J., Cho, S., Li, N., Huang, Y., Bruford, M. W., Zhan, X., Dixon, A., Bertelsen, M. F., Derryberry, E., Warren, W., Wilson, R. K., Li, S., Ray, D. A., Green, R. E., O'Brien, S. J., Griffin, D., Johnson, W. E., Haussler, D., Ryder, O. A., Willerslev, E., Graves, G. R., Alström, P., Fjeldsa, J., Mindell, D. P., Edwards, S. V., Braun, E. L., Rahbek, C., Burt, D. W., Houde, P., Zhang, Y., Yang, H., Wang, J., Consortium, A. G., Jarvis, E. D., Gilbert, M. T. P., Wang, J., Ye, C., Liang, S., Yan, Z., Zepeda, M. L., Campos, P. F., Velazquez, A. M. V., Samaniego, J. A., Avila-Arcos, M., Martin, M. D., Barnett, R., Ribeiro, A. M., Mello, C. V., Lovell, P. V., Almeida, D., Maldonado, E., Pereira, J., Sunagar, K., Philip, S., Dominguez-Bello, M. G., Bunce, M., Lambert, D., Brumfield, R. T., Sheldon, F. H., Holmes, E. C., Gardner, P. P., Steeves, T. E., Stadler, P. F., Burge, S. W., Lyons, E., Smith, J., McCarthy, F., Pitel, F., Rhoads, D.,

and Froman, D. P., 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**(6215), 1311–1320.

Zhao, Z., Guo, C., Sutharzan, S., Li, P., Echt, C. S., Zhang, J., and Liang, C., 2014. Genome-wide analysis of tandem repeats in plants and green algae. *G3: Genes | Genomes | Genetics*, **4**(1), 67–78.

Zhu, L.-y., Nie, L., Zhu, G., Xiang, L.-x., and Shao, J.-z., 2013. Advances in research of fish immune-relevant genes: A comparative overview of innate and adaptive immunity in teleosts. *Developmental and Comparative Immunology*, **39**(1), 39–62.