

DR. PHILINE G. D. FEULNER (Orcid ID : 0000-0002-8078-1788)

Article type : Resource Article

Title: A *de novo* chromosome-level genome assembly of *Coregonus* sp. “Balchen”: one representative of the Swiss Alpine whitefish radiation

Running title: A chromosome-level whitefish genome assembly

Authors: Rishi De-Kayne ^{1,2}, Stefan Zoller ³ and Philine G. D. Feulner ^{1,2}

Affiliations:

1. Department of Fish Ecology and Evolution, Centre of Ecology, Evolution and Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Switzerland
2. Division of Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University of Bern, Switzerland
3. Genetic Diversity Centre (GDC), ETH Zürich, Zürich, Switzerland

Keywords:

Alpine whitefish, *Coregonus*, Whitefish, Salmonidae, genome assembly

Corresponding author:

Philine Feulner

EAWAG Swiss Federal Institute of Aquatic Science and Technology
Seestrasse 79, 6047 Kastanienbaum, Switzerland

Phone: +41 (0)58 765 21 06

Fax: +41 (0)58 765 21 68

philine.feulner@eawag.ch

Abstract:

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the [Version of Record](#). Please cite this article as [doi: 10.1111/1755-0998.13187](https://doi.org/10.1111/1755-0998.13187)

This article is protected by copyright. All rights reserved

Salmonids are of particular interest to evolutionary biologists due to their incredible diversity of life-history strategies and the speed at which many salmonid species have diversified. In Switzerland alone, over 30 species of Alpine whitefish from the subfamily Coregoninae have evolved since the last glacial maximum, with species exhibiting a diverse range of morphological and behavioural phenotypes. This, combined with the whole genome duplication which occurred in the ancestor of all salmonids, makes the Alpine whitefish radiation a particularly interesting system in which to study the genetic basis of adaptation and speciation and the impacts of ploidy changes and subsequent rediploidization on genome evolution. Although well curated genome assemblies exist for many species within Salmonidae, genomic resources for the subfamily Coregoninae are lacking. To assemble a whitefish reference genome, we carried out PacBio sequencing from one wild-caught *Coregonus* sp. “Balchen” from Lake Thun to ~90x coverage. PacBio reads were assembled independently using three different assemblers, Falcon, Canu and wtdbg2 and subsequently scaffolded with additional Hi-C data. All three assemblies were highly contiguous, had strong synteny to a previously published *Coregonus* linkage map, and when mapping additional short-read data to each of the assemblies, coverage was fairly even across most chromosome-scale scaffolds. Here, we present the first *de novo* genome assembly for the Salmonid subfamily Coregoninae. The final 2.2 Gb wtdbg2 assembly included 40 scaffolds, an N50 of 51.9 Mb, and was 93.3% complete for BUSCOs. The assembly consisted of ~52% TEs and contained 44,525 genes.

Introduction

Members of the *Coregonus* genus, known as lake whitefish, are distributed throughout freshwater systems across Europe and North America (Bernatchez & Dodson, 1990; Douglas, Brunner, & Bernatchez, 1999; Hudson, Vonlanthen, & Seehausen, 2011; Kottelat & Freyhof, 2007; Østbye, Bernatchez, Naesje, Himberg, & Hindar, 2005). In many lakes across their range, multiple whitefish species have evolved in the last 12,000 years following the melting of glaciers after the last glacial maximum (Hudson et al., 2011; Kottelat & Freyhof, 2007; Lu & Bernatchez, 1999). Today a particularly speciose clade of whitefish is found throughout pre-alpine lakes across Switzerland, known as the Alpine whitefish radiation (Vonlanthen et al., 2012). Over 30 species are thought to make up this radiation which was previously described as the *C. lavaretus* spp. complex, and new studies continue to identify additional cryptic diversity within the

radiation using genetic methods (Doenz, Bittner, Vonlanthen, Wagner, & Seehausen, 2018; Hudson, Lundsgaard-Hansen, Lucek, Vonlanthen, & Seehausen, 2017; Hudson et al., 2011; Østbye et al., 2005). Within Switzerland, independent, monophyletic, radiations of up to six species have evolved rapidly following the last glacial maximum (Doenz et al., 2018; Hudson et al., 2011). Sympatric whitefish species in these lakes are differentiated in many phenotypic traits including body size, gill-raker number (linked to their feeding ecology) as well as spawning depth and season (Doenz et al., 2018; Hudson et al., 2017; Kottelat & Freyhof, 2007). Repeated phenotypic differentiation has evolved independently across different lake systems resulting in allopatric species exhibiting analogous life history strategies e.g. large, shallow spawning, benthic macro-invertebrate eaters *C. sp. "Bodenbalchen" sp. nov.*, *C. sp. "Balchen"* and *C. duplex* are present in lakes Luzern (Reuss system), Thun/Brienzen (Aare system) and Walen/Zürich (Limmat system), respectively. Likewise, in the same lakes *C. zugensis*, *C. albellus*, and *C. heglungus*, small bodied pelagic zooplanktivores with high numbers of gill rakers have also evolved, alongside up to four other sympatric species. This rapid and repeated evolution of multiple whitefish phenotypes and life history strategies has made the Alpine whitefish a particularly interesting system in which to study the genomic basis of adaptation and speciation. The recent use of genomic data gained from reduced representation libraries has demonstrated the power of genomic approaches for species designation amongst closely related sympatric species (Feulner & Seehausen, 2019). Further, it was demonstrated that genetic differentiation across the genome is widespread when comparing sympatric species from contrasting habitats (Feulner & Seehausen, 2019). However, the low density and uncertainty of positioning of markers along the genome currently still limits a true genome-wide view of adaptation and speciation within these radiations.

The Salmonidae are a focal family in which to study genome evolution, specifically the rediploidization process following whole genome duplication. As part of the Salmonidae family, Coregonids share a common ancestor with the Salmoninae and Thymallinae. Before these sub-families split from one another, the whole lineage experienced a whole genome duplication 80-100 Mya (Lien et al., 2016; Macqueen & Johnston, 2014; Near et al., 2012). Recent studies have determined that different Salmonid genomes are uniquely shaped by rediploidization following this whole genome duplication, referred to as the salmonid-specific fourth vertebrate whole-genome duplication, Ss4R (Robertson et al., 2017). It has been shown that whilst many regions of Salmonid genomes rediploidized prior to the diversification of the three sub-families, and thus are shared across the family, each lineage also has unique patterns of rediploidization for some genomic regions leading to substantial variation in genome structure between lineages (Robertson et al., 2017). To fully understand the impact of whole genome duplication and subsequent rediploidization on genome structure in the Salmonidae, high quality genome assemblies for all major lineages are needed.

Although many salmonid species now have suitably well assembled and curated reference genomes, including Atlantic Salmon (*Salmo salar*; Lien et al., 2016), Rainbow Trout (*Oncorhynchus mykiss*; Berthelot et al., 2014; Pearse et al., 2019), Chinook Salmon (*Oncorhynchus tshawytscha*; Christensen, Leong, et al., 2018), Coho Salmon (*Oncorhynchus kisutch*; NCBI BioProject: PRJNA352719), Arctic Charr (*Salvelinus alpinus*; Christensen, Rondeau, et al., 2018), and European grayling (*Thymallus thymallus*; Sävilammi et al., 2019; Varadharajan et al., 2018), genomic resources for the Coregoninae subfamily are largely limited. To date, the best curated genomic resources for the Coregoninae are next-generation sequencing linkage maps, one for the North American whitefish *Coregonus clupeaformis* (Gagnaire, Normandeau, Pavey, & Bernatchez, 2013), one for the cisco *Coregonus artedii* (Blumstein et al., 2020) and one for the Alpine whitefish radiation (*Coregonus* sp. “*Albock*”; De-Kayne & Feulner, 2018). Here we add to the genomic resources available for Coregonids by producing a chromosome-scale genome assembly for one species of Swiss Alpine whitefish, *Coregonus* sp. “*Balchen*”. To produce the best assembly, we tested three of the best and widely used assemblers, Falcon, Canu and wtdbg2 with ~90x PacBio data, validated each of the resulting assemblies, and selected the best assembly for annotation. The final assembly was produced using wtdbg2 and 94% of its total length was assembled into 40 scaffolds, in addition to 7815 unassigned contigs. This assembly was shown to be made up of ~52% transposable elements (TEs) and contained 93.3% of complete BUSCOs and a total of 44,525 genes.

Materials and Methods

Sample preparation and sequencing

DNA was extracted in multiple batches from heart and somatic muscle tissue of one wild caught (outbred) female *C. sp.* “*Balchen*” from Lake Thun (in December 2016) using the MagAttract HMW DNA Kit (Qiagen). From this high molecular weight DNA, 45µg was used to prepare nine libraries for sequencing on the single-molecule real-time sequencing (SMRT) platform from Pacific Biosciences (Sequel with 2.0 chemistry) using 60 SMRT cells to generate 240 gigabases of sequence data (Next Generation Sequencing Platform, University of Bern). In addition, one Illumina TruSeq library was sequenced (paired end reads of 150 bp; average fragment size for Illumina library preparation 582 bp) on the Illumina HiSeq 3000 platform (Next Generation Sequencing Platform, University of Bern) which generated 87 Gb of data. These Illumina reads were evaluated for quality using FastQC (Andrews 2010) before being used for assembly polishing.

Estimation of genome size

To estimate genome size for the focal species *C. sp.* “*Balchen*”, we used Jellyfish v.1.1.11 (Marçais & Kingsford, 2011) to produce frequency distributions of 17-, 21-, 25 and 30-mers for all Hi-Seq reads. GenomeScope v.1 (Vurture et al., 2017) was then used to estimate genome size from these histograms.

Genome assembly and polishing

Raw PacBio data was assembled independently using three different assemblers (Figure 1), Falcon/Falcon Unzip v.1.9.1 (Chin et al., 2016), Canu v.1.6 (Koren et al., 2017), and wtdbg2 v.2.2 (Ruan & Li, 2019), which have each demonstrated their ability to produce highly contiguous assemblies. In all three cases assembly was carried out using only PacBio data. Read polishing was carried out using both the original raw PacBio reads (Arrow; SMRT link v.5.0.1; <https://github.com/PacificBiosciences/GenomicConsensus>) and low-error rate, short-read, Illumina data (Pilon v.1.22; Walker et al., 2014). After each round of polishing BUSCO v.3.0.2 (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) was run against the core gene set from ray-finned fishes (actinopterygii_odb9) to evaluate the improvement of the assembly. If the number of complete BUSCOs did not improve after running the Arrow algorithm, then only Pilon was used. A second round of Pilon polishing was also used for the wtdbg2 assembly where an additional BUSCO improvement was observed (whereas no improvement was observed after a second round of Pilon polishing of the Canu assembly). All assembly parameters can be found at <https://github.com/RishiDeKayne/WhitefishReferenceGenome>.

Falcon/Falcon Unzip

Genome assembly was carried out by DNAnexus (Mountain View, CA) utilizing the DAMASKER suite (<https://github.com/thegenemyers/DAMASKER>) and the Falcon (v1.9.1) pipeline (Pacific Biosciences; Chin et al., 2016). First the REPmask and TANmask modules of the DAMASKER suite were used on the raw PacBio reads and the resulting output was used as input for the Falcon 1.9.1 pipeline. For the first two steps of the Falcon pipeline, error-correction and read overlap identification, a length cut-off of 5000 bp was used. This assembly was then phased and polished using Falcon-Unzip (Chin et al., 2016) and Pacific Biosciences' Arrow algorithm (<https://github.com/PacificBiosciences/GenomicConsensus>). The final step involved polishing these contigs using ~33x Illumina reads in the Pilon program v.1.22 (Walker et al., 2014). This resulted in primary contigs, thought to represent the haploid whitefish genome, and haplotig contigs, thought to represent alternative alleles at heterozygous sites in the subject fish. For downstream processing of the Falcon assembly this primary contigs assembly was combined with the reads identified as haplotigs by Falcon Unzip. This allowed us to find misidentified primary contigs (which may rather represent haplotigs or mitochondrial DNA) as well as misidentified haplotigs (which could in fact be low coverage contigs or repetitive/duplicated regions).

Canu

Assembly following the Canu (v.1.6) pipeline (Koren et al., 2017) was carried out using the same raw PacBio data. Canu assembly includes three main steps, error correction followed by read trimming and

finally, assembly. Canu read correction was carried out using the default settings regarding minimum read length (1000 bp) and minimum read overlap (500 bp) whilst specifying a genome size of 4 Gb (aimed at minimizing the potential collapsing of sequence similar homeologous regions of the genome). The same parameters were used for the trimming step however, for the assembly step minimum read length was extended to 1200 bp and minimum read overlap to 600 bp. Similar to the Falcon/Falcon Unzip assembly, the final step involved polishing contigs using ~33x Illumina reads in the Pilon program v.1.22 (one round; (Walker et al., 2014)).

wtdbg2

Finally, an assembly was carried out with the least computationally intensive of the three programs - wtdbg2 (v.2.2; Ruan & Li, 2019). wtdbg2 involves two steps, the first of which assembles long reads and the second derives a consensus sequence. For long read assembly kmer psize was set to 21 (-p 21), and 1/3 kmers were subsampled (using -S 3), the maximum length variation of two aligned fragments was set to 0.05 (-s 0.05) and the minimum length of alignment was 5000 bp (-L 5000). After the consensus was derived one round of polishing was carried out using Arrow (v.5.0.1) followed by two rounds of polishing with Pilon v.1.22 (Walker et al., 2014).

Haplotig purging

Following each assembly, we used purge_haplotigs v.1 (Roach, Schmidt, & Borneman, 2018) to identify contigs that were more likely to represent alternative alleles (from heterozygous regions of the genome) or mitochondrial DNA rather than the haploid nuclear genome. In each case, all raw PacBio data was mapped against the assembly and a read-depth histogram was produced. A low, mid, and high value of coverage was then selected from this histogram to flag suspect haplotigs and regions with exceptionally high coverage which should minimize the likelihood of removing sequence similar homeologous regions (all thresholds and histograms can be found in Table S1). Suspect haplotigs were then mapped to the rest of the assembly to identify their allelic partner before the contigs with good matches were reassigned as haplotigs. To assess the gene-level completeness of each assembly after running purge_haplotigs each assembly was again compared against the core gene set from ray-finned fishes (actinopterygii_odb9) using BUSCO v.3.0.2 (Simão et al., 2015).

Genome scaffolding

Hi-C scaffolding of the purged assemblies, including tissue processing, library preparation, and sequencing, was carried out by Phase Genomics (Seattle, WA). Chromatin conformation capture data was generated using a Phase Genomics (Seattle, WA) Proximo Hi-C Animal Kit, which is a commercially available version of the Hi-C protocol (Lieberman-Aiden et al., 2009). Following the manufacturer's instructions for

the kit, intact cells from the same whitefish female were crosslinked using a formaldehyde solution, digested using the Sau3AI restriction enzyme, and proximity ligated with biotinylated nucleotides to create chimeric molecules composed of fragments from different regions of the genome that were physically proximal *in vivo*, but not necessarily genomically proximal. Continuing with the manufacturer's protocol, molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Sequencing was performed on an Illumina HiSeq 4000, generating a total of 249,544,461 100 bp read pairs.

Reads were aligned independently to each of the three draft assemblies (Canu, Falcon and wtdbg2). Briefly, reads were aligned using BWA-MEM v.0.7.17 (Li & Durbin, 2010) with the -5SP and -t 8 options specified, and all other options default. SAMBLASTER (Faust & Hall, 2014) was used to flag PCR duplicates, which were later excluded from analysis. Alignments were then filtered with samtools v.1.9 (Li et al., 2009) using the -F 2304 filtering flag to remove non-primary and secondary alignments and matlock (<https://github.com/phasegenomics/matlock>) using default options. Putative misjoined contigs were broken using Juicebox (Durand et al., 2016) based on the Hi-C alignments; 11 breaks were introduced into the Canu assembly, 42 breaks into the Falcon assembly, and 11 breaks into the wtdbg2 assembly.

Phase Genomics' Proximo Hi-C genome scaffolding platform was used to create scaffolds from each draft assembly in a method similar to that described by Bickhart et al. (2017). As in the LACHESIS method (Burton et al., 2013), this process computes a contact frequency matrix from the aligned Hi-C read pairs, normalized by the number of Sau3AI restriction sites (GATC) on each contig, and constructs scaffolds in such a way as to optimize expected contact frequency. For each of the four assemblies, approximately 100,000 separate Proximo runs were performed to optimize the number of scaffolds and the scaffold construction in order to make the scaffolds as concordant with the observed Hi-C data as possible. Finally, each set of scaffolds was polished an additional time using Juicebox (Durand et al., 2016).

The differential log-likelihood of each set of scaffolds was calculated and examined in the same manner demonstrated by LACHESIS. A threshold of 100 was used to identify contigs scaffolded in a position and orientation in which the log-likelihood (base e) of the chosen orientation was more than 100 times greater than the alternative, a method shown by Burton et al. (2013) to be effective for identifying contigs that are well ordered and oriented in their region of a scaffold. Following scaffolding each of the assemblies, BUSCO v.3.0.2 (Simão et al., 2015) was run again on each complete assembly as well as the 40 scaffolds (denoted by WFSs, CFs, and FSs for the wtdbg2, Canu, and Falcon assemblies, respectively).

Validation of whitefish assemblies

Illumina short read mapping

To assess the qualitative differences between the three scaffolded assemblies we used two independent data sets, the Illumina short reads and a previously published *Coregonus sp.* “Albock” linkage map (see next section). Mapping the Illumina data helped to assess the composition of each of the scaffolds. The Illumina data, collected from the same individual from which the genome was sequenced, was mapped back to each of the reference assemblies. In this way we assessed the consistency of coverage across the assembly and identify potentially duplicated regions of the whitefish genome which may have been collapsed into one sequence during the assembly process. Illumina reads were mapped to each assembly using BWA-MEM v.0.7.17 (Li & Durbin, 2010; with default parameters). A summary of this mapping was produced using samtools (Li et al., 2009; samtools flagstat). Coverage was then calculated in 30 Kb windows using bedtools v.2.27.1 (Quinlan & Hall, 2010) and a custom perl script cov.per.window.pl. Coverage statistics were then calculated in R (R Core Team, 2017).

Linkage map synteny

In addition, we were able to assess the reliability of scaffolding by investigating the synteny between the 40 scaffolds from each assembly and the *C. sp.* “Albock” linkage groups (De-Kayne & Feulner, 2018). RAD loci (90 bp containing a marker) with a known position in the linkage map were mapped to the 40 scaffolds constituting each of the three assemblies using BWA-MEM v.0.7.17 (Li & Durbin, 2010; with default parameters). Synteny plots were then visualized using the *circlize* package in R (Gu et al., 2014; R Core Team, 2017).

Repeat masking and genome annotation

To characterize the repeat landscape of the whitefish genome we first produced a repeat library using RepeatModeler v.1.0.11 (Arian F. A. Smit & Hubley, 2008) for each of the haplotig-purged assemblies. These libraries were then combined with a Salmonidae repeat library (from RepeatMasker repeat database; queryRepeatDatabase.pl -species Salmonidae) to produce one concatenated reference library. Each of the scaffolded assemblies was then repeat masked using this concatenated library with RepeatMasker v.4.0.7 (A. F. A. Smit, Hubley, & Green, 2015). An interspersed repeat landscape was then produced for the best assembly, from wtdbg2, using the RepeatMasker scripts calcDivergenceFromAlign.pl and createRepeatLandscape.pl.

Annotation of the wtdbg2 assembled genome was carried out using a three-pass iterative approach with MAKER2 v.2.31 (Holt & Yandell, 2011). First, an initial gene model was made using our repeat library (described above), protein evidence from *Salmo salar* (UPID: UP000087266) and *Oncorhynchus mykiss* (UPID: UP000193380) proteomes from Uniprot and the Swissprot vertebrate database

(uniprot_sprot_vertebrates), a recently published whitefish transcriptome (Carruthers et al., 2018) and alternative transcriptome evidence from a *Danio rerio* transcriptome (TSA: GDQQ01000001:GDQQ01083602). Next, this gene model was used to produce HMMs with SNAP (Korf, 2004) and Augustus v.3.2.1 (Stanke, Diekhans, Baertsch, & Haussler, 2008). A second pass of MAKER2 was then carried out using these ab initio gene prediction models and the models were optimized before a final third MAKER2 pass was carried out. These final MAKER2 gene models were filtered to remove spurious genes with AED scores < 0.6 (in accordance to Campbell, Holt, Moore & Yandell 2014). Finally, functional annotation of this gene set was carried out using Pannzer2 (Törönen, Medlar, & Holm, 2018) and the accuracy of the annotation was determined using BUSCO v.3.0.2 (Simão et al., 2015; with the core gene set from ray-finned fishes actinopterygii_odb9) on the final gene set. To further evaluate our gene set we used Orthofinder v2.3.11 (Emms & Kelly 2015, 2019) to construct orthologous gene sets. This analysis included protein sequences from whitefish (42,695 genes annotated with an AED < 0.6 and positioned on the 40 wtdbg2 whitefish scaffolds) and three other salmonids (*Hucho hucho* (ASM331708v1; GCA_003317085.1; submitted by University Of Aberdeen on July 2018), *Salmo salar* (ICSASG_v2; GCA_000233375.4; submitted by International Cooperation to Sequence the Atlantic Salmon Genome on June 2015), *Salmo trutta* (fSalTru1.1; GCA_901001165.1; submitted by SC on June 2019) and the outgroup (which did not go through the Ss4R whole genome duplication) *Esox lucius* (Eluc_V3; GCA_000721915.3; submitted by Ben F Koop, Jong S Leong on January 2017). All protein files were downloaded from ENSEMBL (<https://www.ensembl.org/index.html>; 21.Feb 2020).

Identification of homeologous regions in the whitefish genome

Following the whole genome duplication that occurred in an ancestral salmonid it is possible to determine which whitefish scaffolds (WFSs; many of which correspond to chromosomes) are homeologs of one another by identifying pairs of scaffolds that are sequence similar. After hard-masking the WFSs resulting from the wtdbg2 assembly and scaffolding we aligned each WFS to all other WFSs in SyMAP v.5.0 (Soderlund, Bomhoff & Nelson 2011; Soderlund, Nelson, Shoemaker & Paterson 2006), using default parameters. For each of the 55 links between homeologous WFS blocks identified in SyMAP, Lastz v.1.02 (Harris 2007) was run in both direction (using the parameters: --gfextend --nochain --nogapped --matchcount=100; similarly to Lien et al., 2016) to align these regions to one another and subsequently determine sequence similarity between the two. Following Lastz alignment, matches were filtered to remove those with sequence similarity < 75% (in keeping with Lien et al., 2016) and/or smaller than 1000bp, and sequence similarity was then averaged across alignments within each block.

Ancestral chromosome identification

We also aimed to determine the single (non-duplicated) ancestral chromosome that each homeologous pair of scaffolds corresponds to, and subsequently determine the level to which chromosomal rearrangements may have taken place in whitefish. Each wtdbg2 WFS was mapped to the Northern Pike (*Esox lucius*) genome (GCF_004634155.1) using SyMAP v.5.0 (Soderlund, Bomhoff & Nelson 2011; Soderlund, Nelson, Shoemaker & Paterson 2006) following the ancestral chromosome identification convention used by Sutherland et al. (2016) and Blumstein et al. (2020). The identified syntenic links between WFSs and pike chromosomes then allowed us to determine which ancestral pike chromosome (PK; pre-whole genome duplication) corresponds to each homeologous pair of WFSs (identified above; as the result of whole genome duplication) or to a single WFS, evidential of one copy of a pair having been lost.

Results

Estimation of genome size

Reports of *Coregonus* genome sizes vary widely with estimates ranging from 2.4 Gb (Hardie & Hebert, 2003) to 3.5 Gb (Lockwood, Seavey, Dillinger, & Bickham, 1991). Using the best model (k=25), as determined using the 'Model Fit' output from GenomeScope, we estimated *C. sp. "Balchen"* to have a genome size of 2.63 Gb (Figure 2). Based on this estimate of genome size the PacBio sequencing used equates to ~91x coverage (and the Illumina HiSeq to ~33x).

Genome assembly and polishing

After the first step of the Falcon pipeline, prior to phasing and polishing with Falcon Unzip the assembly contained 52,448 primary contigs covering 2.78 Gb with an N50 contig length of 204 Kb. After the Falcon Unzip step the primary contigs assembly was made up of 19,553 contigs covering 2.41 Gb and with an N50 of 280 Kb (Table 1). For downstream analysis this primary contigs assembly was merged with the haplotigs assembly, as described above. This merged assembly was 4.11 Gb long, contained 60,605 contigs, had an N50 of 136 Kb and included 89.5% of complete BUSCOs (Table 1; Table S1). The Canu assembly was substantially larger than the primary reads from the Falcon/Falcon Unzip assembly and covered 3.28 Gb across 52,023 contigs, with an N50 of 131 Kb and including 88.7% BUSCOs (Table 1; Table S1). The wtdbg2 assembly was the shortest of the three assemblies with a total length of 2.38 Gb and also had the fewest contigs (28,224; Table 1). However, it had the highest N50 of 424 Kb and contained the highest percentage of complete BUSCOs, 93.4% (Table 1; Table S1).

Haplotig purging

After haplotig purging the differences between the three assemblies was reduced dramatically, with the range of contigs now from 16,440 to 22,627 (for wtdbg2 and Canu, respectively; Table S1). The N50 of all three assemblies also increased, particularly in the Falcon and Canu assemblies, from 136 Kb and 131 Kb

to 281 Kb and 258 Kb each. The N50 of the wtdbg2 assembly also increased, although less significantly, from 424 Kb to 491 Kb (Table S1). The number of complete BUSCOs went up in both the Falcon and Canu assemblies after haplotig purging (by 1.3% and 4.4%), but dropped slightly (by 0.3%) in the wtdbg2 assembly (Table S1). The high completeness percentage of BUSCOs for each of the assemblies prior to scaffolding suggests that we have succeeded in capturing a large proportion of the whitefish genome sequence during the assembly process.

Genome scaffolding

Hi-C scaffolding of contigs into scaffolds resulted in a set of 40 scaffolds, many of which were chromosome-scale, for each of the three assemblies, containing 2.38 Gb (96% of all sequence; Falcon), 2.41 Gb (98% of all sequence; Canu) and 2.07 Gb (94% of all sequence; wtdbg2). The differential log-likelihood calculation, which was used to identify the length of confidently ordered and orientated scaffolds, resulted in 1.08 Gb (45.41%) for the Falcon scaffolds (FSs), 1.21 Gb (50.1%) for the Canu scaffolds (CSs), and 1.74 Gb (84.1%) for the wtdbg2 scaffolds (WFSs), meeting this criterion. These results are in agreement with the patterns observable in the final scaffold heatmaps for each assembly, in which the patterns observable for the wtdbg2 scaffolds are more in alignment with a priori expectations about Hi-C linkage density patterns (Figure 3; contact plots for the Falcon and Canu assemblies are displayed in Figures S1 and S4; van Berkum et al., 2010), yielding qualitative confirmation of the quantitative scaffold quality assessment. The percentage of complete BUSCOs went up for each assembly following Hi-C scaffolding, with the Falcon, Canu, and wtdbg2 assemblies now having 91.8%, 93.7%, and 93.3% (Table 1). When considering only scaffolds, the Canu assembly retained the highest complete percentage of BUSCOs at 93.7% with the Falcon and wtdbg2 assemblies dropping only slightly to 91.5% and 93.0% each. Based on having the highest length of confidently scaffolded contigs and the high number of complete BUSCOs, the Hi-C scaffolded wtdbg2 assembly was selected as the best of the three assemblies and was uploaded to the European Nucleotide Archive (Accession: GCA_902810595.1). Falcon and Canu assemblies are available on Dryad (<https://doi.org/10.5061/dryad.xd2547ddf>).

Validation of whitefish assemblies

Illumina short read mapping

Summaries of the mapping of Illumina reads to each of the assemblies can be seen in Table S2. The wtdbg2 assembly had the highest number of mappings over MAPQ 30 (71.6%) with the Canu and Falcon assemblies having slightly lower proportions of high-quality mappings (69.5% and 60.7% each). When considering the proportion of read mates mapped to a different scaffold, however, the Canu assembly looks the best of the three with only 1.4% of mates mapped to a different scaffold with a MAPQ > 5, compared to 2% for Falcon and 2.4% for wtdbg2. As a result of our coverage analysis the highest mean coverage (in 30

Kb windows) was 17.3x observed in the wtdbg2 assembly. The Falcon and Canu assemblies had lower mean coverages of 15.0x and 15.9x, respectively. Plots of coverage across the 40 wtdbg2 WFSs are shown in Figure 4 and the equivalent plots for the Falcon and Canu assemblies in Figures S2 and S5. Based on these coverage plots we identified regions which are likely to represent collapsed duplicated regions, spread across each genome assembly (Table S3 and Table S4). As expected in these regions, the coverage was approximately double that of the rest of the assembly. In other salmonid genome assemblies, which have successfully resolved each copy of a duplicated region, these duplicated regions typically span whole chromosome arms or even chromosomes. Similarly, in the whitefish assemblies we identify collapsed blocks which encompass whole scaffolds or parts of scaffolds. In the wtdbg2 assembly some WFSs (WFS4, 7, 14 and 37) likely represent collapsed regions which span chromosome arms, and in other WFSs (WFS22, 28, 32, 36, and 38) the whole chromosome appears collapsed (a BED file containing the locations of these wtdbg2 collapsed duplicates is included in Table S3). We estimate that in total 309 Mb of the wtdbg2 assembly (representing 14% of the assembly) is collapsed compared to 413 Mb and 517 Mb in the Canu and Falcon assemblies, respectively (representing 17% and 21% of each assembly; Table S4).

Linkage map synteny

Out of the 5395 markers from the *C. sp.* “Albock” linkage map which were mapped to each assembly, only high-quality mappings (MAPQ > 30) were retained, resulting in a mapping position for 3648, 4494, and 4744 markers in the sequence of the Falcon, Canu, and wtdbg2 assemblies respectively (Table S2). For all three assemblies, concordance between sequence and recombination position across the majority of markers was very high, suggesting a high synteny between linkage groups and scaffolds. In the wtdbg2 assembly 95% of markers (4489/4744) showed strong synteny between one linkage group and one WFS (38 out of the 40 linkage groups; Figure 5; equivalent to figures S3 and S6 for the Falcon and Canu assemblies, respectively). Only two scaffolds, (WFS38 and 40) could not be matched to any linkage group. We also identified a series of substantial deviations from the broader pattern, where a number of markers from a linkage group also mapped to a second, alternative scaffold. This was the case for markers from Calb01 – WFS37, Calb02 – WFS32, Calb08 – WFS38, Calb13 – WFS35, Calb16 – WFS06, Calb20 – WFS28, Calb34 – WFS07, and Calb36 – WFS04. Strikingly the mapping location of seven of these deviations (WFS04, 07, 28, 32, 35, 37, and 38) also represent seven of the nine scaffolds we identified as collapsed duplicates showing inflated coverage (shown in grey in Figures 4, 5 and 7). Although part of WFS17 looked like a collapsed duplicate based on coverage, no significant deviations of markers mapped to this scaffold. Additionally, despite having an unusual mapping pattern with markers from Calb16 in addition to those from Calb33, WFS06 showed a consistent coverage of Illumina reads mapping. Out of all deviating markers, 83% (212/255) mapped to regions identified as collapsed duplicates. No markers from the linkage map were successfully mapped to WFS40, the smallest of the scaffolds, at only 1.1 Mb long.

Additionally, in two cases markers from two linkage groups predominantly mapped to one WFS. Markers from both Calb35 and Calb40 mapped to WFS31 and from Calb38 and Calb39 to WFS22. The few deviations from the patterns of synteny between linkage groups and scaffolds could be caused by a number of interacting factors. These include the potential collapse of some scaffolds or parts of scaffolds, small mistakes in either the linkage map or the sequence assembly, and the distribution of repetitive sequence similar regions which all reduce the accuracy of mapping the short (100 bp) RAD loci from the linkage map. However, collapsed regions likely play the most significant role in driving deviations in the observed patterns of synteny as evidenced by the fact that the majority of deviating RAD loci map to regions that are thought to be collapsed.

Repeat masking and genome annotation

Around 52% of each assembly was masked with the most abundant repetitive elements being DNA elements followed by LINEs and then unclassified repeats (Table S1). DNA elements alone made up nearly a quarter of each assembly (24.65% of Falcon, 23.79% of Canu, and 24.41% of the wtdbg2 assembly). The resulting landscape (Figure 6) identified the Class II TE superfamily Tc1-*mariner* as the most abundant in the *Coregonus sp. "Balchen"* genome, making up 18% of the interspersed repeats. The most abundant Class I TEs were LINE-2 elements, however, these only made up 4.2% of the interspersed repeats. The three-pass MAKER2 annotation resulted in the identification of 44,525 protein-coding genes (42,695 on scaffolds and 1,830 on unscaffolded contigs) and included 357,479 identified exons (Table 2), with the final set of genes 81.8% complete for BUSCOs (C:81.8%[S:54.8%,D:27.0%],F:9.3%,M:8.9%,n:4584). Functional annotation with Pannzer2 allowed the assignment of gene ontology terms to 29,046 genes. Across a total of 415,276 genes in the five species, we identified 41,042 orthogroups. This includes 7,673 species-specific orthogroups, 725 single-copy orthologs, and 16,599 orthogroups with all five species present. Out of the 42,695 whitefish genes (on scaffolds) 38,219 could be assigned to 22,311 orthogroups. The number of whitefish genes annotated on scaffolds (42,695) is similar to the number of genes annotated in the diploid outgroup (*Esox lucius* 43,143 genes), but substantially lower than the ENSMBL annotation of any of the other three salmonids (*Hucho hucho* 91,817 genes, *Salmo salar* 121,064 genes, *Salmo trutta* 116,557 genes). This suggests that more transcriptomic work on various tissues and developmental stages is warranted to further improve the annotation of the whitefish genome. We found 4,504 orthogroups that were duplicated in whitefish (present with more than one copy) that were present in a single copy in pike (*Esox lucius*). A total of 2,746 orthogroups show duplication patterns relative to the pike consistent across all four salmonids. However, we also identify 3,459 orthogroups where only the other three salmonids show duplications and 474 orthogroups where only whitefish appear duplicated. This could reflect true biological differences in duplication loss in the different salmonid lineages, however technical artefacts

during the assembly (collapsing of highly identical regions) or gene annotation differences likely contribute as well.

Identification of homeologous regions in the whitefish genome

Using SyMAP we identified 55 syntenic links between 34 of the 40 WFSs (Figure 7; Table S5). Sequence similarity calculations for each link (110 mappings, one in each direction, for each of the 55 homeologous blocks) showed that the majority of identified syntenic blocks had sequence similarity $\geq 90\%$ (shown in orange and red in Figure 7). Slightly lower sequence similarity was observed for syntenic links between WFS02 and WFS03 and multiple smaller links between WFS01 and WFS31, WFS07 and WFS34, and WFS08 and WFS17. The WFSs for which no syntenic links were identified were WFS22, WFS28, WFS32, WFS38, WFS39, and WFS40. Of those, WFS22, WFS28, WFS32, and WFS38 have been identified as fully collapsed based on their unusual coverage patterns, hence likely represent residual tetraploid regions or a collapse of two very sequence similar homeologs. It is likely that the shortest scaffold, WFS40, constitutes part of a chromosome, explaining why no homeologous scaffold was found.

Ancestral chromosome identification

As above, SyMAP was used to identify the corresponding non-duplicated ancestral chromosome for each syntenic WFS pair. This comparison of the WFSs and the Northern Pike genome identified that the majority of pike chromosomes (18 out of 25) had two equivalent WFSs, the remaining seven pike chromosomes match to only one WFS (Table 3; Figure S7). These seven pike chromosomes included PK2, PK9, PK11, PK20, and PK23 which corresponded to WFS32, WFS38, WFS35, WFS28, and WFS4. Additionally, PK22 and PK25 are both in single copy in whitefish but appear to have been rearranged into a single WFS, WFS22.

Discussion

To enable future studies to investigate both the genetic basis of adaptation and speciation within Coregoninae and genome evolution following whole genome duplication across the Salmonidae family we have assembled the first whitefish reference genome. Assembling $> 90\times$ PacBio data from one female *Coregonus sp.* “Balchen” with three of the most commonly used assemblers resulted in three high quality assemblies, each with $>90\%$ complete BUSCOs and 40 scaffolds, many of which are chromosome-scale. Out of these three assemblies we judge the assembly produced by wtdbg2 as the best. This new draft whitefish genome is 2.2 Gb and comprises 40 scaffolds (containing 94% of nucleotides) and 7,815 unscaffolded contigs, has an N50 of 51.9 Mb and contains 93.3% complete BUSCOs. Annotation of the assembly identified 44,525 genes in total and showed that transposable elements make up 52% of the *C. sp.* “Balchen” genome.

To assemble the first reference genome from Coregoninae, we made use of three different but widely used genome assemblers, Falcon/Falcon Unzip, Canu, and wtdbg2. Although studies describing new assembly software often compare genome assemblies produced with the same input data and multiple different assemblers, the high cost and prohibitive computational time usually restricts these comparisons to genomes < 150 Mb including *Arabidopsis thaliana* and *Drosophila melanogaster* (Chin et al., 2016) and those of model systems including human cell line CHM1 (Koren et al., 2017; Ruan & Li, 2019). Few studies have reported such performance comparisons with non-model species, despite an increasing number of *de novo* reference genome assembly projects for organisms with large, complex genomes. All three assemblies were subsequently polished using Arrow (Falcon and wtdbg2) and/or Pilon (Falcon, Canu and wtdbg2) and scaffolded using Hi-C technology into 40 scaffolds. At the contig stage the Falcon assembly was the longest at 4.11 Gb (containing 60,605 contigs), and the wtdbg2 the shortest at 2.38 Gb (containing 28,224 contigs). Although the structure of each assembly at the contig stage varied, each of the three assemblies had high complete BUSCO percentages (Table S1). This shows that all three assemblers performed well with the input data, producing contigs which incorporate around 90% of genes known to be present in all ray-finned fishes. After haplotig purging, the three assemblies became more similar in size and N50 value (Table S1), which suggests that the assemblers differed largely in their ability to resolve alleles. Next, Hi-C scaffolding was used, resulting in 40 scaffolds for each assembly. For the Falcon, Canu and wtdbg2 assemblies, 2.38/2.47 Gb (96%), 2.41/2.46 (98%), and 2.07/2.2 Gb (94%), respectively were assigned to the 40 scaffolds. Out of the three assemblies more of the wtdbg2 assembly could be confidently scaffolded (84.1% meeting criterion compared to 45.4% and 50.1% in the Falcon and Canu assemblies). Also, during the assembly validation process the wtdbg2 assembly appeared to be the best of the three assemblies, having the highest mean coverage across the genome and lowest proportion of the genome in potentially collapsed regions (Table S4) as well as the largest proportion of confidently mapped linkage map markers. Although the complete BUSCO scores were slightly lower than that of the Canu assembly (93% in the wtdbg2 assembly compared to 93.7% in the Canu assembly), the increased confidence of scaffolding in the wtdbg2 assembly, the superior mapping metrics, the lowest proportion of the genome being collapsed, and highest synteny with the linkage map, all led to us selecting this as the best assembly.

The scaffold N50 of the wtdbg2 *C. sp. "Balchen"* assembly was 51.9 Mb, which is higher than a number of recently published salmonid genomes including Chinook salmon (1.138 Mb; Christensen, Leong, et al., 2018), Arctic charr (1.02 Mb; Christensen, Rondeau, et al., 2018) and grayling (33 Mb; Sävilammi et al., 2019). The characterization of the repeat landscape of the whitefish genome also highlighted the broad similarity in the proportion of many families of transposable elements between salmonid species. We identified that around 52% of the whitefish genome is repetitive, a similar proportion to that of Chinook

salmon (56%; Christensen, Leong, et al., 2018), Arctic charr (56%; Christensen, Rondeau, et al., 2018), and European grayling (47%; Sävilammi et al., 2019). The relative abundances of different types of repetitive element are similar to those reported in other salmonid assemblies including that of Atlantic Salmon (Lien et al., 2016) and Chinook Salmon (Christensen, Leong, et al., 2018). The relatively high abundance of Class II TE superfamily Tc1-*mariner* and LINE-2 elements amongst the youngest elements suggest that these families are still expanding and potentially diversifying in the whitefish genome. Conversely, the lack of new LTR elements suggests that their abundance and diversity peaked in the past and they are no longer diversifying in the genome. Annotation of the wtdbg2 assembly identified 44,525 genes, similar to those reported in the publications associated with the rainbow trout genome (46,585 by Berthelot et al., 2014; 53,383 by Pearse et al., 2019) but higher than the 37,206 genes identified by Lien et al. (2016) in the Atlantic salmon genome and the 36,216 identified by Christensen, Leong et al. (2018) in the Chinook salmon genome.

Although whole genome duplications have punctuated the tree of life, few have occurred recently enough to allow investigations into the subsequent rediploidization process at the genomic level. Salmonids are therefore an ideal family in which to study rediploidization since the genomic signals of whole genome duplication and genomic rearrangements which followed have not yet been confounded by other genomic processes such as mutations (including small point mutations and large structural changes such as inversions and deletions; Macqueen & Johnston, 2014). One recent investigation into rediploidization within Salmonidae already identified substantial genomic differences between 16 salmonid species, which evolved independently as rediploidization proceeded (Robertson et al., 2017). However, with high-quality genomes available for an increasing number of salmonid species, the resolution with which we can identify differences which have occurred in genome structure and composition following whole genome duplication is vastly increasing. Therefore, highly contiguous reference genomes, particularly for underrepresented groups such as the salmonid subfamily Coregoninae, are invaluable to fill gaps in the genomic resources currently available. Here, we have been able to determine pairs of whitefish scaffolds that represent homeologs and their corresponding ancestral chromosome (using Northern Pike chromosome numbering). This will facilitate future comparisons across the salmonid family (similar to those by Blumstein et al., 2020) to assess the independent rediploidization process in different salmonid lineages. We also identified a number of whitefish scaffolds for which no homeolog was present. By combining synteny data with our coverage-based validation of the new *C. sp.* “*Balchen*” assembly we showed that some of these regions were due to the collapsing of highly sequence similar regions (e.g. WFS22 and WFS28). In other instances, we rather identified potential genomic rearrangements as the driver of this pattern, for example, pike PK24 showed homology with WFS39 but also a part (potentially a chromosome arm) of WFS08 (Table 3; Figure S7). More complex patterns were also identified such as the merging of two different PK chromosomes (22

and 25) into one whitefish scaffold WFS22 that showed inflated coverage estimates across the scaffold suggesting that the both homeologs of the fused PK chromosomes were collapsed in our assembly due to their high sequence similarity. Although their high sequence similarity makes duplicated regions difficult to assemble and subsequently sometimes causes their collapse during assembly, we have assembled a highly continuous reference genome. This is despite a high sequence similarity between large parts of the genome, with at least 80% of the genome being $\geq 90\%$ sequence similar. Our estimates of sequence similarity are comparable to observations in Atlantic salmon where it has been shown that 94% of chromosome sequence is duplicated, with 26% of the genome having a duplicate region with sequence similarity $>90\%$ (Lien et al., 2016). Future work should aim to investigate the partially collapsed scaffolds WFS04, WFS07, WFS17 and WFS37 and collapsed regions which span the length of scaffolds WFS22, WFS28, WFS32, WFS35, and WFS38, which we identified (Figure 4; Figure 7; Table S3), to determine the evolutionary history of these duplicates and the process by which the non-duplicated regions of the whitefish genome may have rediploidized independently, or not, compared to other salmonid species. Specifically, it should be determined whether the lack of identified homeologs for WFS22, WFS28, WFS32, WFS35 and WFS38 is an artefact of our assembly (as indicated by their increased coverage; Figure 4; for example for WFS28, WFS32, WFS38), the result of genomic rearrangements resulting in coregonid-specific arrangement (such as WFS39), or a combination of the two (such as WFS22).

In addition to facilitating the investigation of salmonid genome evolution, the highly contiguous whitefish assembly presented here will also support future genomic studies within the subfamily Coregoninae. Coregoninae are distributed across the Northern hemisphere (North America and Eurasia), widely fished and of economic importance, and exhibit an extraordinary ecological diversity. Studying whitefish diversification is of fundamental scientific interest to understand the processes driving and facilitating such diversification and to assist in the conservation of this diverse group. We anticipate that the whitefish genome assembly presented here will aid future investigations into the ecology and evolution of all whitefish. Specifically, it will facilitate investigations into the genetic basis of adaptation across the Alpine whitefish radiation including determining the level of parallelism across multiple pre-alpine lake systems at a genome-wide resolution.

Acknowledgements:

Thanks to Benjamin Gugger and team from Lake Thun whitefish hatchery for providing us with the *C. sp.* “*Balchen*” used for genome assembly. Also, thanks to Daniel Jeffries for many helpful discussions regarding the genome assembly pipeline, Oliver Selz for his taxonomic expertise and Cameron Hudson and David Frei for their suggestions to improve the clarity of the manuscript.

Thanks to the Triple A Winter School 2017 (ETH Zurich) for providing invaluable core skills for the project. Data produced and analysed in this paper were generated in collaboration with the Next Generation Sequencing Platform, University of Bern and the Genetic Diversity Centre (GDC), ETH Zurich. We also thank DNAnexus for the Falcon assembly and Phase Genomics for the scaffolding of all three assemblies. This project is funded by the Swiss National Science Foundation (SNSF project 31003A_163446/1 awarded to PGDF).

References:

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

Bernatchez, L., & Dodson, J. J. (1990). Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial-DNA restriction analysis. *Evolution; International Journal of Organic Evolution*, 44, 1263–1271.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., ... Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, 5, 3657.

Bickhart, D. M., Rosen, B. D., Koren, S., Sayre, B. L., Hastie, A. R., Chan, S., ... Smith, T. P. L. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nature Genetics*, 49, 643–650.

Blumstein, D.M., Campbell, M.A., Hale, M.C., Sutherland, B.J.G., McKinney, G.J., Stott, W., & Larson, W.A. (2020). Comparative genomic analyses and a novel linkage map for cisco (*Coregonus artedii*) provides insight into chromosomal evolution and rediploidization across salmonids. *bioRxiv*, 834937.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31, 1119–1125.

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. *Current protocols in bioinformatics*, 48, 4-11.

Carruthers, M., Yurchenko, A. A., Augley, J. J., Adams, C. E., Herzyk, P., & Elmer, K. R. (2018). De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. *BMC Genomics*, 19, 32.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., ... Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, 13, 1050–1054.

Christensen, K. A., Leong, J. S., Sakhrani, D., Biagi, C. A., Minkley, D. R., Withler, R. E., ... Devlin, R. H. (2018). Chinook salmon (*Oncorhynchus tshawytscha*) genome and transcriptome. *PloS One*, 13, e0195461.

Christensen, K. A., Rondeau, E. B., Minkley, D. R., Leong, J. S., Nugent, C. M., Danzmann, R. G., ... Koop, B. F. (2018). The Arctic charr (*Salvelinus alpinus*) genome and transcriptome assembly. *PloS One*, 13, e0204076.

De-Kayne, R., & Feulner, P. G. D. (2018). A European whitefish linkage map and its implications for understanding genome-wide synteny between salmonids following whole genome duplication. *G3*, 8, 3745–3755.

Doenz, C. J., Bittner, D., Vonlanthen, P., Wagner, C. E., & Seehausen, O. (2018). Rapid buildup of sympatric species diversity in Alpine whitefish. *Ecology and Evolution*, 8, 9398–9412.

Douglas, M. R., Brunner, P. C., & Bernatchez, L. (1999). Do assemblages of *Coregonus* (Teleostei: Salmoniformes) in the Central Alpine region of Europe represent species flocks? *Molecular Ecology*, 8, 589–603.

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, 3, 99–101.

Emms, D.M., & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* 20:238

Emms, D.M., & Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157

Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, 30, 2503–2505.

Feulner, P. G. D., & Seehausen, O. (2019). Genomic insights into the vulnerability of sympatric whitefish species flocks. *Molecular Ecology*, 28, 615–629.

Gagnaire, P.-A., Normandeau, E., Pavey, S. A., & Bernatchez, L. (2013). Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, 22, 3036–3048.

Gu Z., Gu L., Eils R., Schlesner M., Brors B. (2014) circlize implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812.

Hardie, D. C., & Hebert, P. D. (2003). The nucleotypic effects of cellular DNA content in cartilaginous and ray-finned fishes. *Genome*, 46, 683-706.

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

Hudson, A. G., Lundsgaard-Hansen, B., Lucek, K., Vonlanthen, P., & Seehausen, O. (2017). Managing cryptic biodiversity: Fine-scale intralacustrine speciation along a benthic gradient in Alpine whitefish (*Coregonus* spp.). *Evolutionary Applications*, 10, 251–266.

Hudson, A. G., Vonlanthen, P., & Seehausen, O. (2010). Rapid parallel adaptive radiations from a single hybridogenic ancestral population. *Proceedings of the Royal Society B: Biological Sciences*, 278, 58-66.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27, 722–736.

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

Kottelat, M., & Freyhof, J. (2007). *Handbook of European freshwater fishes*. Cornol, Switzerland: Publications Kottelat

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., ... Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326, 289–293.

Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., ... Davidson, W. S. (2016). The Atlantic salmon genome provides insights into rediploidization. *Nature*, 533, 200–205.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Lockwood, S. F., Seavey, B. T., Dillinger, R. E., Jr., & Bickham, J. W. (1991). Variation in DNA content among age classes of broad whitefish (*Coregonus nasus*) from the Sagavanirktok River delta. *Canadian Journal of Zoology*, 69, 1335–1338.

Lu, G., & Bernatchez, L. (1999). Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): Support for the ecological speciation hypothesis. *Evolution*, 53, 1491–1505.

Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, 281, 20132881.

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27, 764–770.

Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., ... & Smith, W. L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences*, 109, 13698–13703.

Østbye, K., Bernatchez, L., Naesje, T. F., Himberg, K.-J. M., & Hindar, K. (2005). Evolutionary history of the European whitefish *Coregonus lavaretus* (L.) species complex as inferred from mtDNA phylogeography and gill-raker numbers. *Molecular Ecology*, 14, 4371–4387.

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., ... & Moen, T. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature ecology & evolution*, 3, 1731–1742.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19, 460.

Robertson, F. M., Gundappa, M. K., Grammes, F., Hvidsten, T. R., Redmond, A. K., Lien, S., ... Macqueen, D. J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biology*, 18, 111.

Ruan, J., & Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *BioRxiv*, 530972.

Sävilampi, T., Primmer, C. R., Varadharajan, S., Guyomard, R., Guiguen, Y., Sandve, S. R., ... & Lien, S. (2019). The chromosome-level genome assembly of European grayling reveals aspects of a unique genome evolution process within salmonids. *G3*, 9, 1283-1294.

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

Smit, A. F. A., & Hubley, R. (2008). RepeatModeler Open-1.0. Available from [Http://www.repeatmasker.org](http://www.repeatmasker.org)

Smit, A. F. A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. Available from [Http://www.repeatmasker.org](http://www.repeatmasker.org)

SMRT Link v.5.0.1 – Pacific Biosciences SMRT Tools Reference Guide.
<https://www.pacb.com/support/software-downloads/>

Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic acids research*, 39, e68-e68.

Soderlund, C., Nelson, W., Shoemaker, A., & Paterson, A. (2006). SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome research*, 16, 1159-1168.

Sutherland, B. J., Gosselin, T., Normandeau, E., Lamothe, M., Isabel, N., Audet, C., & Bernatchez, L. (2016). Salmonid chromosome evolution as revealed by a novel method for comparing RADseq linkage maps. *Genome Biology and Evolution*, 8, 3600-3617

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

Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: a rapid functional annotation web server. *Nucleic Acids Research*, 46, W84–W88.

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., ... & Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE (Journal of Visualized Experiments)*, (39), e1869.

Varadharajan, S., Sandve, S. R., Gillard, G. B., Tørresen, O. K., Mulugeta, T. D., Hvidsten, T. R., ... & Jakobsen, K. S. (2018). The grayling genome reveals selection on gene expression regulation after whole-genome duplication. *Genome biology and evolution*, 10, 2785-2800.

Vonlanthen, P., Bittner, D., Hudson, A. G., Young, K. A., Müller, R., Lundsgaard-Hansen, B., ... Seehausen, O. (2012). Eutrophication causes speciation reversal in whitefish adaptive radiations. *Nature*, 482, 357–362.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33, 2202–2204.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, 9, e112963.

Data accessibility

Fastq files for PacBio and Illumina raw reads are deposited at the European Nucleotide Archive (ENA study accession: PRJEB33097, sample accession: ERS359599). The wtdbg2 reference assembly available

online at the European Nucleotide Archive (accession: GCA_902810595.1). Falcon and Canu assemblies are available on Dryad: <https://doi.org/10.5061/dryad.xd2547ddf>. All parameters and scripts for genome analysis are accessible at <https://github.com/RishiDeKayne/WhitefishReferenceGenome>

Author contributions

P.G.D.F. conceived and designed the study. Genome assembly and annotation was carried out by R.D.K. and S.Z., and genome validation by R.D.K. The manuscript was written by R.D.K. and P.G.D.F. Funding was awarded to P.G.D.F. All authors read and revised the manuscript and approved the final version for submission.

Figures:

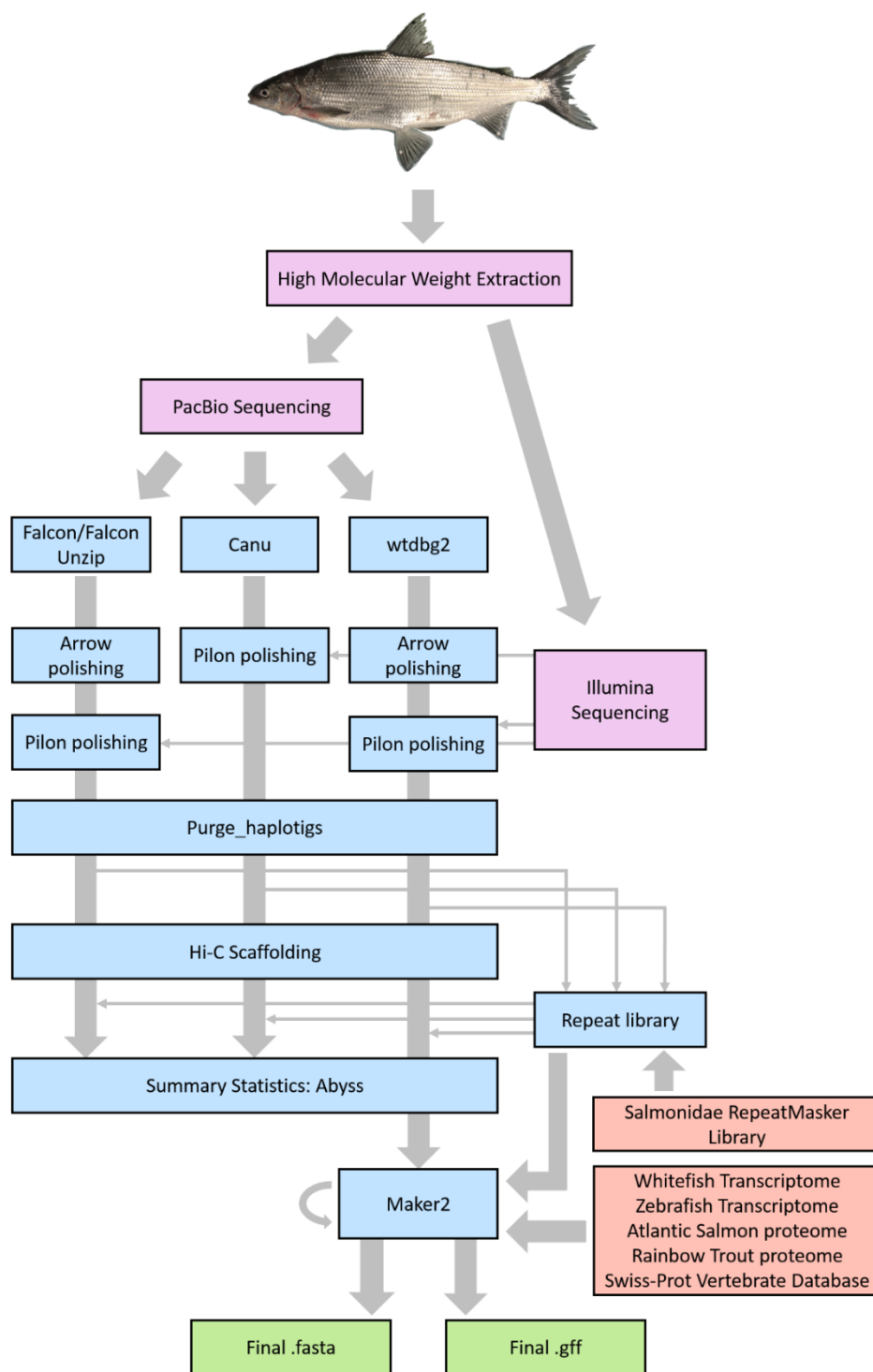


Figure 1. Workflow outlining the different steps and tools used to assemble the whitefish genome (coloured in blue). New input produced for this study is coloured in purple and previously published resources used for repeat masking and annotation in orange. Final outputs are shown in green.

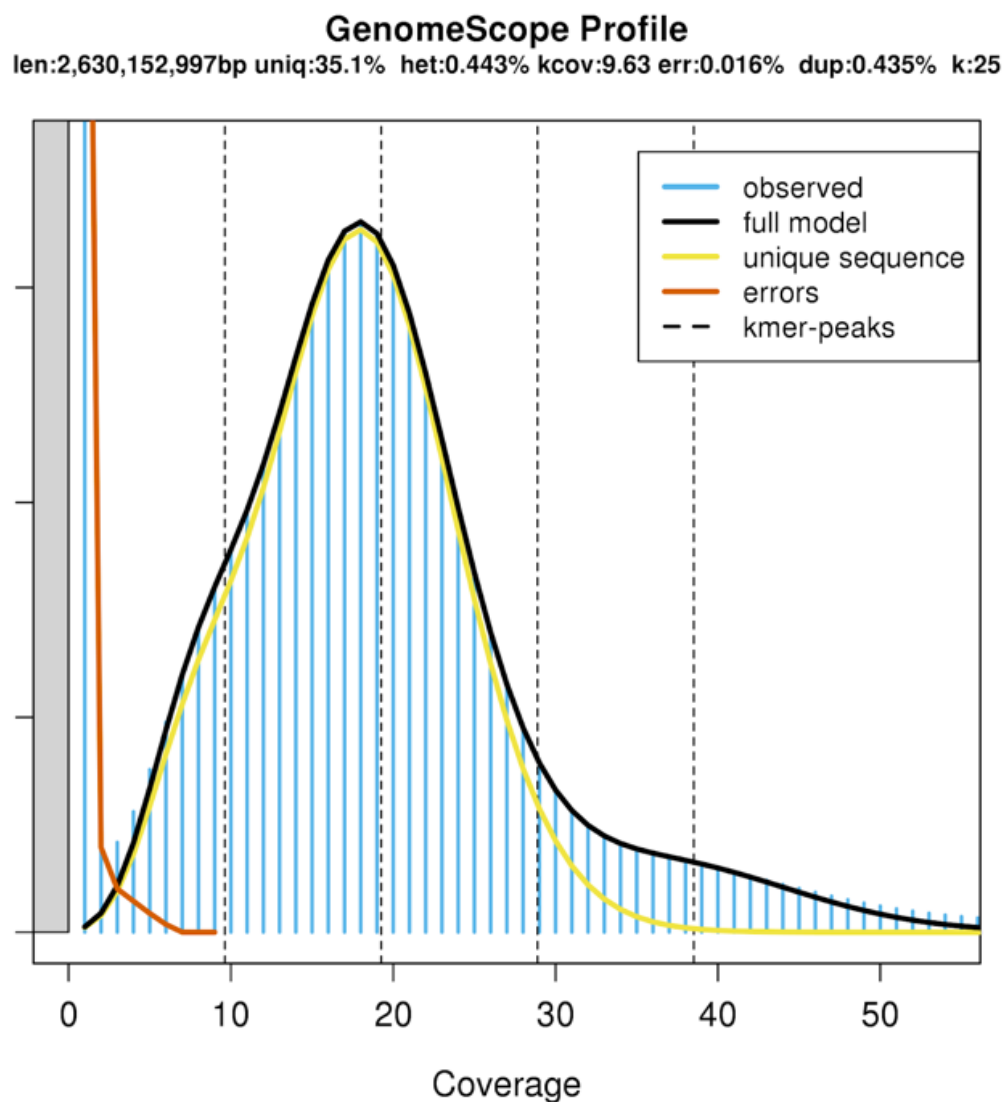


Figure 2. GenomeScope profile established based on short read data, which estimates the genome size of *Coregonus sp. "Balchen"* to be 2.6 Mb.

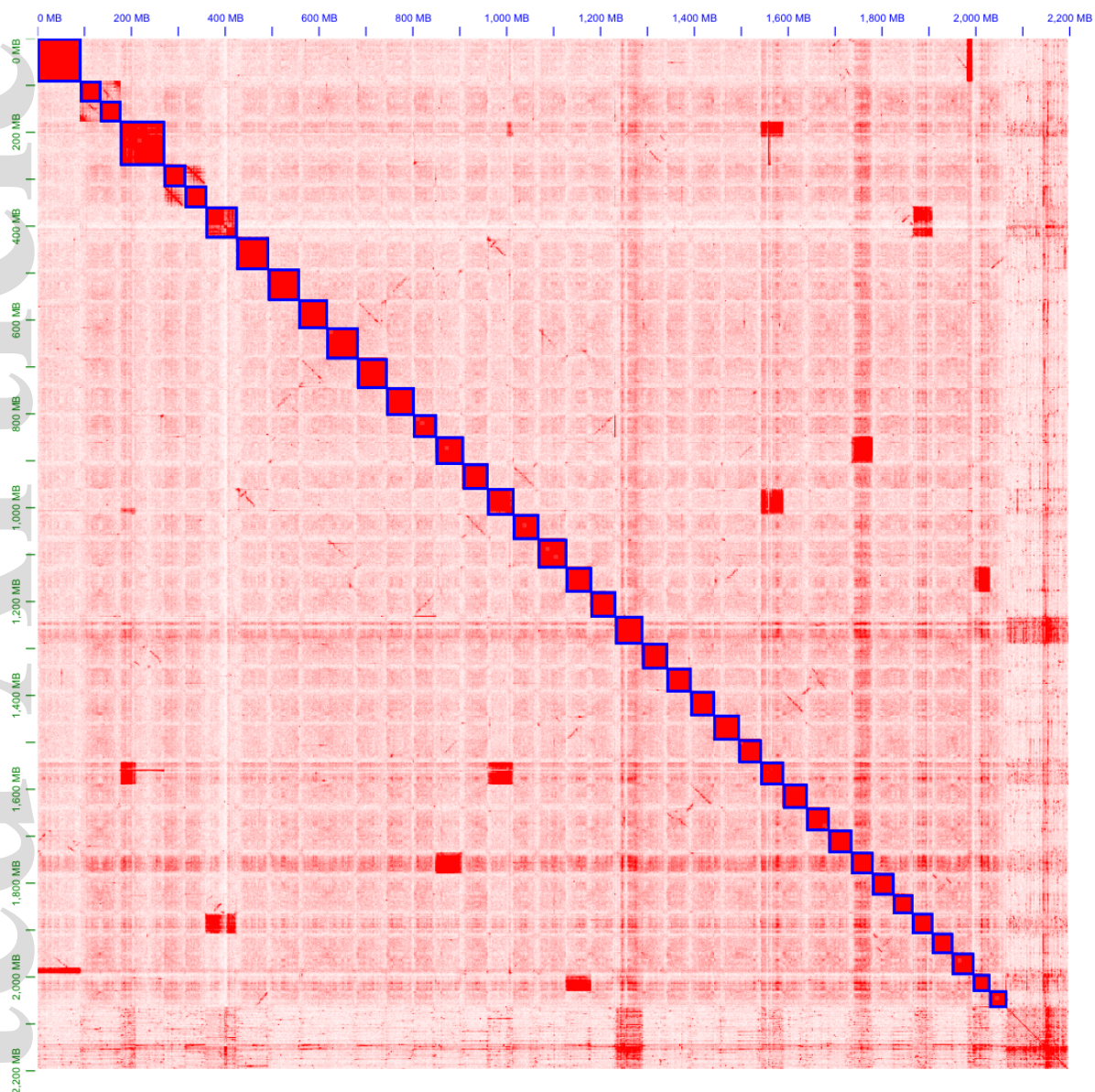


Figure 3. *Coregonus* sp. “Balchen” contig contact map from Hi-C scaffolding of the wtdbg2 assembly. The intensity of red represents the relative contact density between contigs. The highest contact density is found within whitefish scaffolds (WFSs), which are outlined in blue.

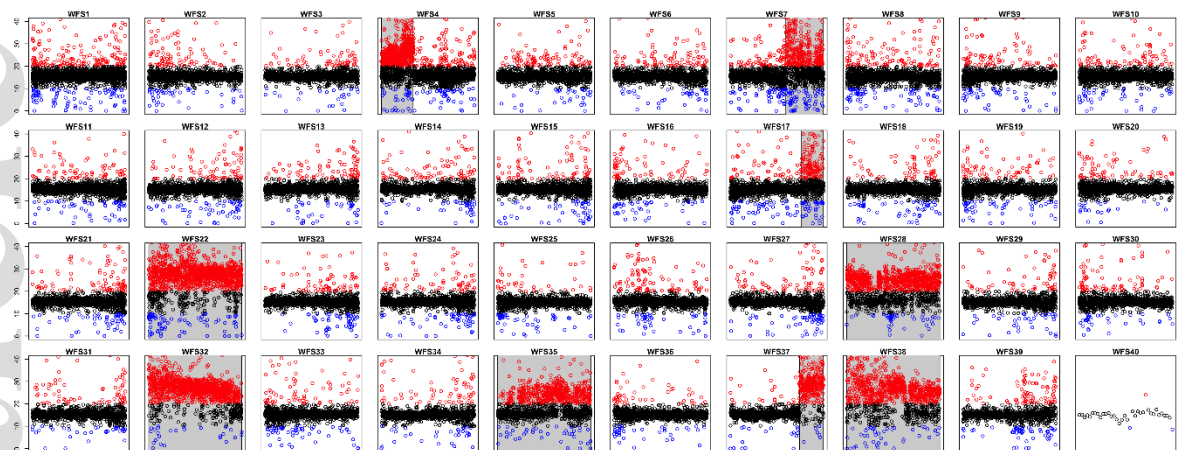


Figure 4. For each of the whitefish scaffolds (WFSs) coverage of Illumina data mapped to the wtdbg2 assembly is plotted in 30 Kb windows. Most windows show an average coverage of around 17x (black points). Windows with coverage >20, and <10 are coloured in red and blue respectively. Putative collapsed duplicate regions are highlighted in grey.

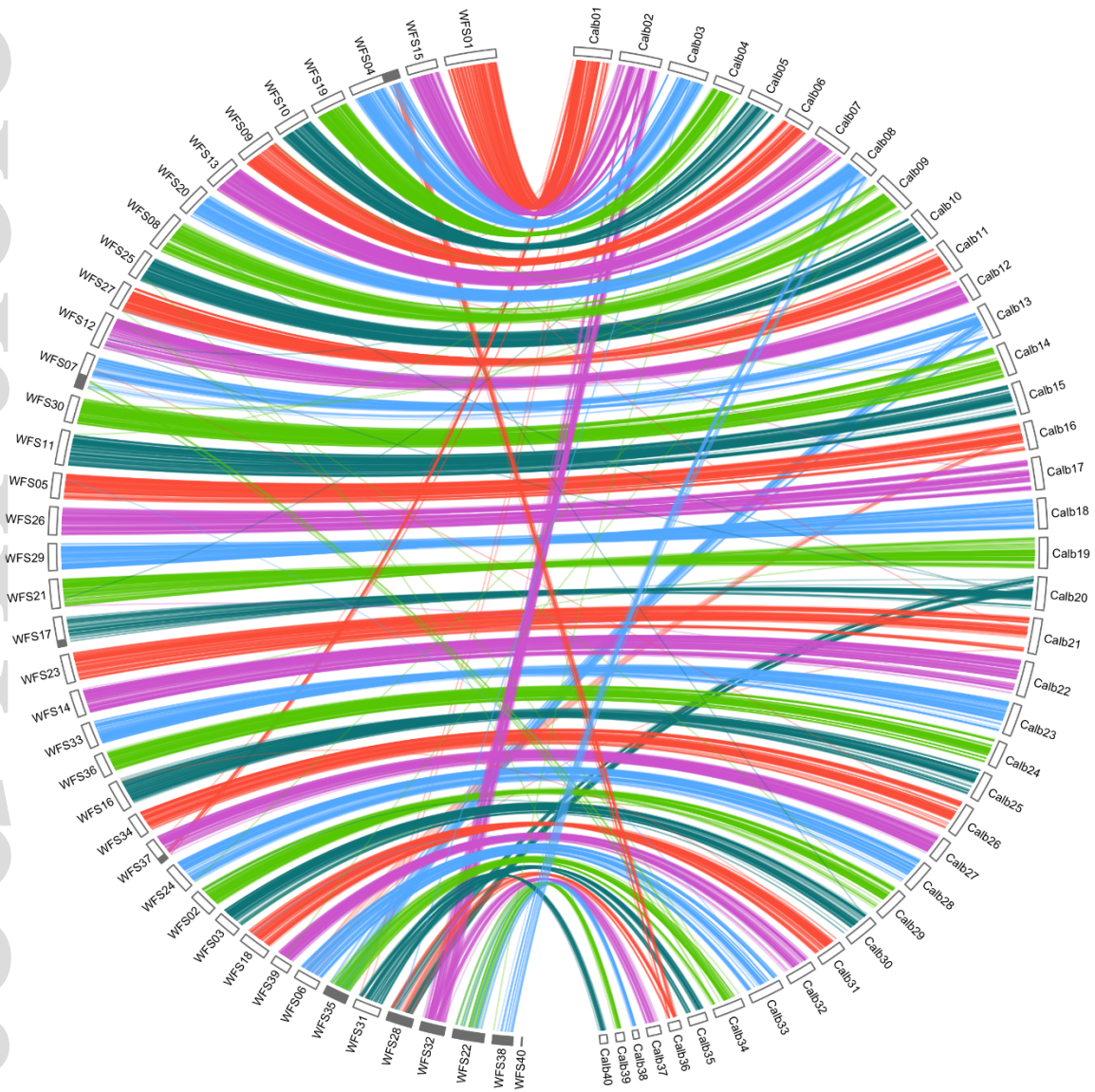


Figure 5. Circos plot comparing the structure of the *C. sp.* “*Albock*” linkage map (right; De-Kayne & Feulner, 2018) and the 40 whitefish scaffolds (WFSs) of the wtdbg2 *C. sp.* “*Balchen*” assembly (left). Lines indicate mapping locations of RAD loci from the linkage map in the genome assembly. Most mappings suggest a good match between linkage map and genome assembly (high synteny between linkage groups and WFSs) and only few lines map discordantly. Genome assembly regions which represent collapsed duplicate regions are identified in grey around the left perimeter.

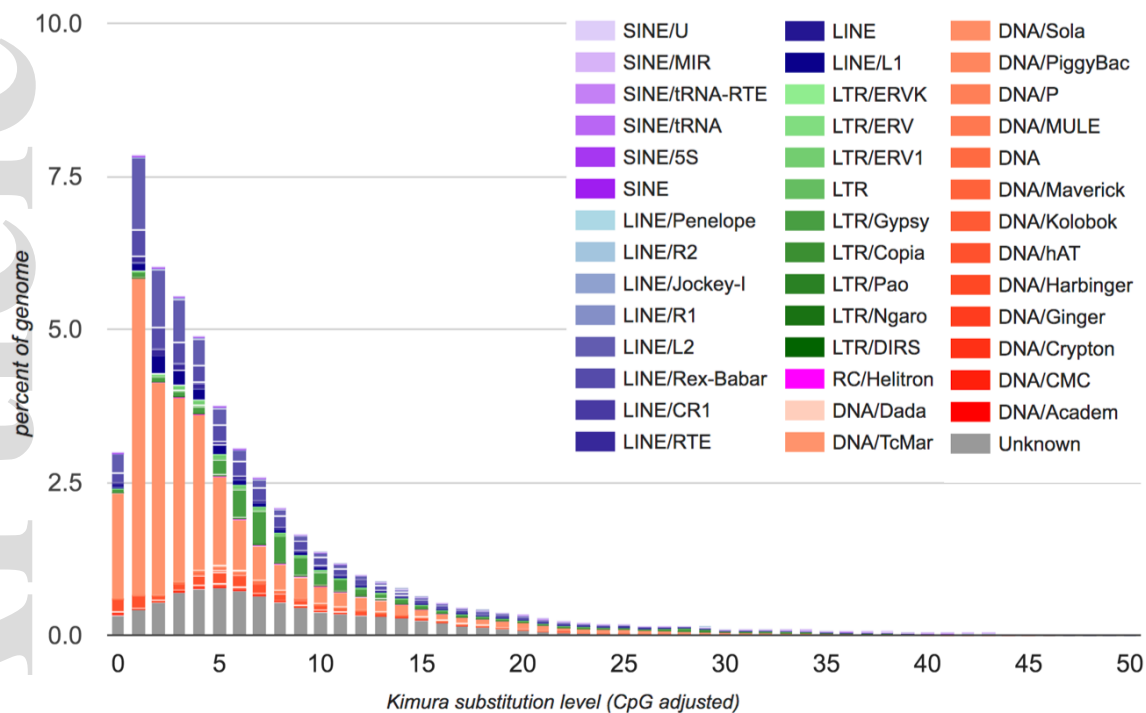


Figure 6. *Coregonus* sp. "Balchen" transposable element divergence landscape. Transposable elements within the whitefish genome have been characterised (different classes represented by distinct colours). The plot shows the relative abundance of each class and their relative age (molecular clock estimate). Note the ongoing DNA element diversification within the whitefish genome, particularly in DNA elements and LINEs.

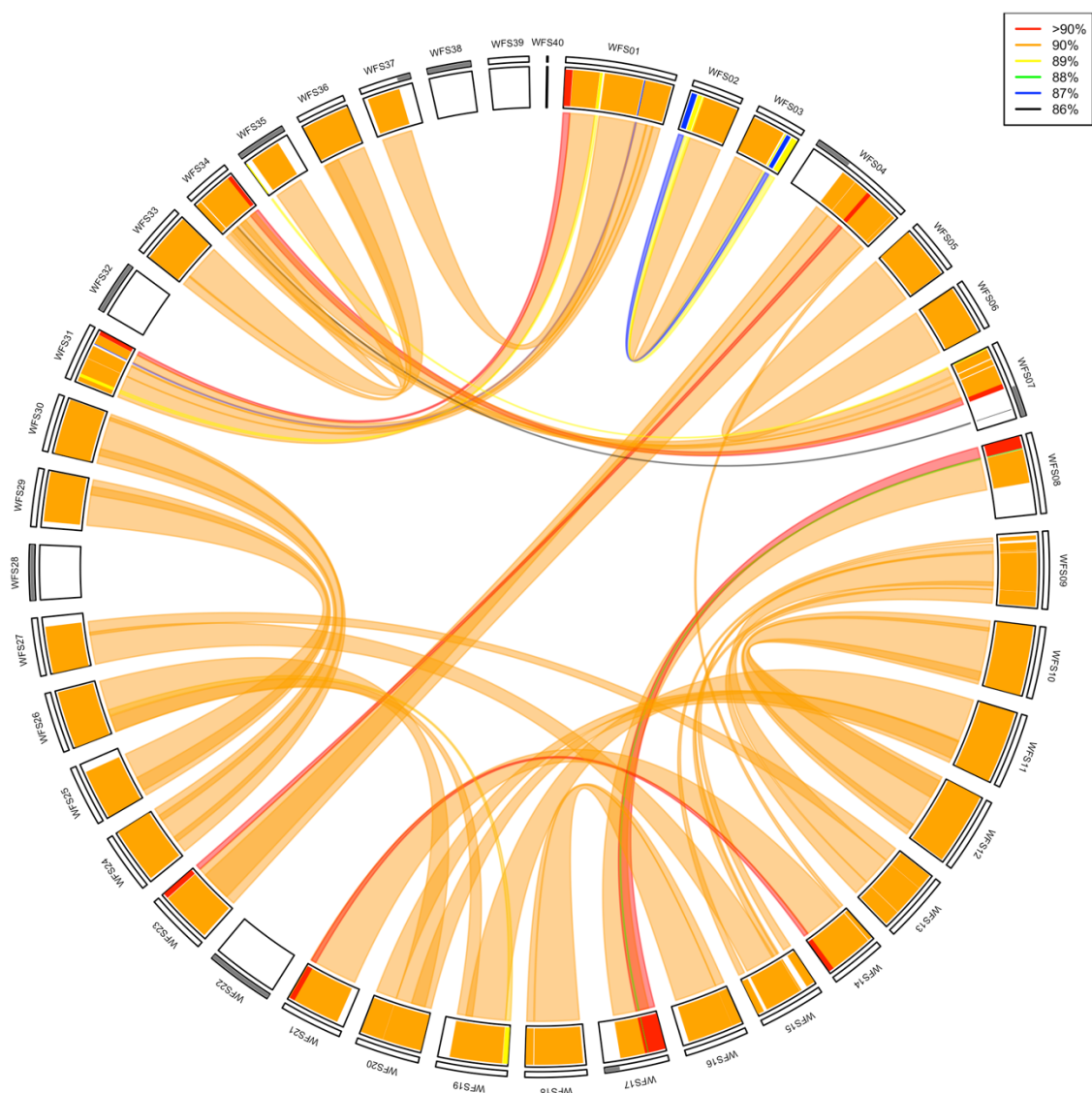


Figure 7. Homeologous whitefish scaffolds (WFSs) within the whitefish genome identified using SyMAP. Links between homeologs are coloured according to their mean sequence similarity based on Lastz alignments for each syntenic link in both directions. Genome assembly regions which are thought to be collapsed are identified in grey on the outer most track.

Table 1. Summary statistics at the contig (pre- haplotig purging) and scaffold stage (each scaffolded assembly contains 40 scaffolds and a number of unscaffolded contigs) for the Falcon, Canu and wtdbg2 assemblies.

| | | Falcon (primary contigs) | Canu | wtdbg2 |
|--------------------------------|--------------------------------|---------------------------------|-------------|---------------|
| Contig statistics | | | | |
| | Number of contigs | 60605 (19553) | 52023 | 28224 |
| | Contig N50 (bp) | 136418 (279657) | 130955 | 424474 |
| | Longest contig (bp) | 6516619 (6516619) | 5278180 | 5201837 |
| | Total contig length (Gb) | 4.11 (2.41) | 3.28 | 2.38 |
| Scaffolded assembly statistics | | | | |
| | Number of scaffolds | 40 | 40 | 40 |
| | Number of unscaffolded contigs | 3705 | 3513 | 7815 |
| | Combined N50 (bp) | 62840000 | 59340000 | 51930000 |
| | Longest scaffold (bp) | 111300000 | 104000000 | 93420000 |
| | Total combined length (Gb) | 2.47 | 2.46 | 2.20 |
| Scaffolded assembly BUSCOs/ 40 | | | | |

| scaffolds BUSCOs | | | | |
|---------------------|------------|------------------------------|------------------------------|------------------------------|
| | Complete | 4209 (91.8%)/4195 (91.5%) | 4299 (93.7%)/4297 (93.7%) | 4274 (93.3%)/4263 (93%) |
| | Single | 2713 (59.2%)/2732 (59.6%) | 2578 (56.2%)/2583 (56.3%) | 2551 (55.7%)/2551 (55.7%) |
| | Duplicated | 1496 (32.6%)/1463 (31.9%) | 1721 (37.5%)/1714 (37.4%) | 1723 (37.6%)/1712 (37.3%) |
| | Fragmented | 89 (1.9%)/77 (1.7%) | 78 (1.7%)/78 (1.7%) | 95 (2.1%)/83 (1.8%) |
| | Missing | 286 (6.3%)/312 (6.8%) | 207 (4.6%)/209 (4.6%) | 215 (4.6%)/238 (5.2%) |

Table 2. Genome annotation summary statistics for final wtdbg2 assembly following three-pass MAKER2 annotation.

| | | |
|-------|------------------------------|-----------|
| Genes | Number | 44525 |
| | Mean length (bp) | 11473.3 |
| | Median length (bp) | 4850 |
| | Min/max (bp) | 77/181605 |
| | Gene frequency (genes/Mb) | 20.24 |

| | | |
|-------|--------------------|---------|
| Exons | Number | 357479 |
| | Mean length (bp) | 196.7 |
| | Median length (bp) | 135 |
| | Min/max (bp) | 2/17274 |

Table 3. Summary of synteny between whitefish scaffolds (WFSs) and Northern Pike (PK) chromosomes, and between WFSs. Homology between WFSs and PK chromosomes, and between WFSs was inferred using SyMAP. In the first column WFSs are underlined if they are thought to be collapsed. In the second column corresponding PK chromosomes are identified with a cross (†) if they were found to have homology with only one WFS. The third column indicates homeologous WFSs, which were identified when two WFSs showed synteny with the same PK chromosome (see Figure S7). An asterisk (*) highlights homeologous relationships that were not confirmed with our analysis of synteny between WFSs.

| WFS | Corresponding Pike - PK chromosome (secondary PK) | Homeologous WFS (secondary homeolog) |
|-----|--|---|
| 1 | 6 (5) | 37 (31) |
| 2 | 4 | 3 |
| 3 | 4 | 2 |
| 4 | 16 (23†) | 23 |
| 5 | 1 | 6 |
| 6 | 1 | 5 |
| 7 | 21 | 34 |
| 8 | 10 (24) | 17 (39*) |
| 9 | 7 | 13 |
| 10 | 17 | 12 |
| 11 | 19 | 19 |
| 12 | 17 | 10 |
| 13 | 7 | 9 |
| 14 | 3 | 21 |
| 15 | 14 | 27 |
| 16 | 13 | 18 |
| 17 | 10 | 8 |
| 18 | 13 | 16 |
| 19 | 19 | 11 |

| | | |
|----|-----------|-------|
| 20 | 8 | 26 |
| 21 | 3 | 14 |
| 22 | 22† (25†) | - (-) |
| 23 | 16 | 4 |
| 24 | 12 | 30 |
| 25 | 15 | 29 |
| 26 | 8 | 20 |
| 27 | 14 | 15 |
| 28 | 20† | - |
| 29 | 15 | 25 |
| 30 | 12 | 24 |
| 31 | 5 | 1 |
| 32 | 2† | - |
| 33 | 18 | 36 |
| 34 | 21 | 7 |
| 35 | 11† | - |
| 36 | 18 | 33 |
| 37 | 6 | 1 |
| 38 | 9† | - |
| 39 | 24 | 8* |
| 40 | - | - |