1 **Title:** A *de novo* chromosome-level genome assembly of *Coregonus sp. "Balchen"*: one

2 representative of the Swiss Alpine whitefish radiation

3

4 **Running title:** A chromosome-level whitefish genome assembly

5

6 **Authors:** Rishi De-Kayne [1,2], Stefan Zoller [3] and Philine G. D. Feulner [1,2]

7

8 **Affiliations:**

9 1. Department of Fish Ecology and Evolution, Centre of Ecology, Evolution and

10 Biogeochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology,

11 Switzerland

12 2. Division of Aquatic Ecology and Evolution, Institute of Ecology and Evolution, University

13 of Bern, Switzerland

14 3. Genetic Diversity Centre (GDC), ETH Zürich, Zürich, Switzerland

15 **Keywords:**

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

17

18 **Corresponding author:**

19 Philine Feulner

20 EAWAG Swiss Federal Institute of Aquatic Science and Technology

21 Seestrasse 79, 6047 Kastanienbaum, Switzerland

22 Phone: +41 (0)58 765 21 06

23 Fax: +41 (0)58 765 21 68

24 philine.feulner@eawag.ch

Abstract:

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/Brienz (Aare system) and Walen/Zurich (Limmat system), respectively. Likewise, in the same lakes *C. zugensis*, *C. albellus*, and *C. heglingus*, 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,

78 2019). Further, it was demonstrated that genetic differentiation across the genome is widespread when

79 comparing sympatric species from contrasting habitats (Feulner & Seehausen, 2019). However, the

80 low density and uncertainty of positioning of markers along the genome currently still limits a true

81 genome-wide view of adaptation and speciation within these radiations.

82

83 The Salmonidae are a focal family in which to study genome evolution, specifically the

84 rediploidization process following whole genome duplication. As part of the Salmonidae family,

85 Coregonids share a common ancestor with the Salmoninae and Thymallinae. Before these sub-

86 families split from one another, the whole lineage experienced a whole genome duplication 80-100

87 Mya (Lien et al., 2016; Macqueen & Johnston, 2014; Near et al., 2012). Recent studies have

88 determined that different Salmonid genomes are uniquely shaped by rediploidization following this

89 whole genome duplication, referred to as the salmonid-specific fourth vertebrate whole-genome

90 duplication, Ss4R (Robertson et al., 2017). It has been shown that whilst many regions of Salmonid

91 genomes rediploidized prior to the diversification of the three sub-families, and thus are shared across

92 the family, each lineage also has unique patterns of rediploidization for some genomic regions leading

93 to substantial variation in genome structure between lineages (Robertson et al., 2017). To fully

94 understand the impact of whole genome duplication and subsequent rediploidization on genome

95 structure in the Salmonidae, high quality genome assemblies for all major lineages are needed.

96

97 Although many salmonid species now have suitably well assembled and curated reference genomes,

98 including Atlantic Salmon (*Salmo salar*; Lien et al., 2016), Rainbow Trout (*Oncorhynchus mykiss*;

99 Berthelot et al., 2014; Pearse et al., 2019), Chinook Salmon (*Oncorhynchus tshawytscha*; Christensen,

100 Leong, et al., 2018), Coho Salmon (*Oncorhynchus kisutch*; NCBI BioProject: PRJNA352719), Arctic

101 Charr (*Salvelinus alpinus*; Christensen, Rondeau, et al., 2018), and European grayling (*Thymallus*

102 *thymallus*; Sävilammi et al., 2019; Varadharajan et al., 2018), genomic resources for the Coregoninae

103 subfamily are largely limited. To date, the best curated genomic resources for the Coregoninae are

104 next-generation sequencing linkage maps, one for the North American whitefish *Coregonus*

105 *clupeaformis* (Gagnaire, Normandeau, Pavey, & Bernatchez, 2013), one for the cisco *Coregonus*

artedi (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.

134

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

162 represent alternative alleles at heterozygous sites in the subject fish. For downstream processing of the

163 Falcon assembly this primary contigs assembly was combined with the reads identified as haplotigs

164 by Falcon Unzip. This allowed us to find misidentified primary contigs (which may rather represent

165 haplotigs or mitochondrial DNA) as well as misidentified haplotigs (which could in fact be low

166 coverage contigs or repetitive/duplicated regions).

167

168 *Canu*

169 Assembly following the Canu (v.1.6) pipeline (Koren et al., 2017) was carried out using the same raw

170 PacBio data. Canu assembly includes three main steps, error correction followed by read trimming

171 and finally, assembly. Canu read correction was carried out using the default settings regarding

172 minimum read length (1000 bp) and minimum read overlap (500 bp) whilst specifying a genome size

173 of 4 Gb (aimed at minimizing the potential collapsing of sequence similar homeologous regions of the

174 genome). The same parameters were used for the trimming step however, for the assembly step

175 minimum read length was extended to 1200 bp and minimum read overlap to 600 bp. Similar to the

176 Falcon/Falcon Unzip assembly, the final step involved polishing contigs using ~33x Illumina reads in

177 the Pilon program v.1.22 (one round; (Walker et al., 2014)).

178

179 *wtdbg2*

180 Finally, an assembly was carried out with the least computationally intensive of the three programs -

181 wtdbg2 (v.2.2; Ruan & Li, 2019). wtdbg2 involves two steps, the first of which assembles long reads

182 and the second derives a consensus sequence. For long read assembly kmer psize was set to 21 (-p

183 21), and 1/3 kmers were subsampled (using -S 3), the maximum length variation of two aligned

184 fragments was set to 0.05 (-s 0.05) and the minimum length of alignment was 5000 bp (-L 5000).

185 After the consensus was derived one round of polishing was carried out using Arrow (v.5.0.1)

186 followed by two rounds of polishing with Pilon v.1.22 (Walker et al., 2014).

187

188 **Haplotig purging**

189 Following each assembly, we used purge_haplotigs v.1 (Roach, Schmidt, & Borneman, 2018) to

190 identify contigs that were more likely to represent alternative alleles (from heterozygous regions of

191 the genome) or mitochondrial DNA rather than the haploid nuclear genome. In each case, all raw

192 PacBio data was mapped against the assembly and a read-depth histogram was produced. A low, mid,

193 and high value of coverage was then selected from this histogram to flag suspect haplotigs and

194 regions with exceptionally high coverage which should minimize the likelihood of removing sequence

195 similar homeologous regions (all thresholds and histograms can be found in Table S1). Suspect

196 haplotigs were then mapped to the rest of the assembly to identify their allelic partner before the

197 contigs with good matches were reassigned as haplotigs. To assess the gene-level completeness of

198 each assembly after running purge_haplotigs each assembly was again compared against the core

199 gene set from ray-finned fishes (actinopterygii_odb9) using BUSCO v.3.0.2 (Simão et al., 2015).

200

201 **Genome scaffolding**

202 Hi-C scaffolding of the purged assemblies, including tissue processing, library preparation, and

203 sequencing, was carried out by Phase Genomics (Seattle, WA). Chromatin conformation capture data

204 was generated using a Phase Genomics (Seattle, WA) Proximo Hi-C Animal Kit, which is a

205 commercially available version of the Hi-C protocol (Lieberman-Aiden et al., 2009). Following the

206 manufacturer's instructions for the kit, intact cells from the same whitefish female were crosslinked

207 using a formaldehyde solution, digested using the Sau3AI restriction enzyme, and proximity ligated

208 with biotinylated nucleotides to create chimeric molecules composed of fragments from different

209 regions of the genome that were physically proximal *in vivo*, but not necessarily genomically

210 proximal. Continuing with the manufacturer's protocol, molecules were pulled down with streptavidin

211 beads and processed into an Illumina-compatible sequencing library. Sequencing was performed on

212 an Illumina HiSeq 4000, generating a total of 249,544,461 100 bp read pairs.

213

214 Reads were aligned independently to each of the three draft assemblies (Canu, Falcon and wtdbg2).

215 Briefly, reads were aligned using BWA-MEM v.0.7.17 (Li & Durbin, 2010) with the -5SP and -t 8

216 options specified, and all other options default. SAMBLASTER (Faust & Hall, 2014) was used to flag

217 PCR duplicates, which were later excluded from analysis. Alignments were then filtered with

218 samtools v.1.9 (Li et al., 2009) using the -F 2304 filtering flag to remove non-primary and secondary

219 alignments and matlock (https://github.com/phasegenomics/matlock) using default options. Putative

220 misjoined contigs were broken using Juicebox (Durand et al., 2016) based on the Hi-C alignments; 11

221 breaks were introduced into the Canu assembly, 42 breaks into the Falcon assembly, and 11 breaks

222 into the wtdbg2 assembly.

223

224 Phase Genomics' Proximo Hi-C genome scaffolding platform was used to create scaffolds from each

225 draft assembly in a method similar to that described by Bickhart et al. (2017). As in the LACHESIS

226 method (Burton et al., 2013), this process computes a contact frequency matrix from the aligned Hi-C

227 read pairs, normalized by the number of Sau3AI restriction sites (GATC) on each contig, and

228 constructs scaffolds in such a way as to optimize expected contact frequency. For each of the four

229 assemblies, approximately 100,000 separate Proximo runs were performed to optimize the number of

230 scaffolds and the scaffold construction in order to make the scaffolds as concordant with the observed

231 Hi-C data as possible. Finally, each set of scaffolds was polished an additional time using Juicebox

232 (Durand et al., 2016).

233

234 The differential log-likelihood of each set of scaffolds was calculated and examined in the same

235 manner demonstrated by LACHESIS. A threshold of 100 was used to identify contigs scaffolded in a

236 position and orientation in which the log-likelihood (base e) of the chosen orientation was more than

237 100 times greater than the alternative, a method shown by Burton et al. (2013) to be effective for

238 identifying contigs that are well ordered and oriented in their region of a scaffold. Following

239 scaffolding each of the assemblies, BUSCO v.3.0.2 (Simão et al., 2015) was run again on each

240 complete assembly as well as the 40 scaffolds (denoted by WFSs, CFs, and FSs for the wtdbg2, Canu,

241 and Falcon assemblies, respectively).

242

243 **Validation of whitefish assemblies**

244 *Illumina short read mapping*

245   To assess the qualitative differences between the three scaffolded assemblies we used two

246   independent data sets, the Illumina short reads and a previously published *Coregonus sp. "Albock"*

247   linkage map (see next section). Mapping the Illumina data helped to assess the composition of each of

248   the scaffolds. The Illumina data, collected from the same individual from which the genome was

249   sequenced, was mapped back to each of the reference assemblies. In this way we assessed the

250   consistency of coverage across the assembly and identify potentially duplicated regions of the

251   whitefish genome which may have been collapsed into one sequence during the assembly process.

252   Illumina reads were mapped to each assembly using BWA-MEM v.0.7.17 (Li & Durbin, 2010; with

253   default parameters). A summary of this mapping was produced using samtools (Li et al., 2009;

254   samtools flagstat). Coverage was then calculated in 30 Kb windows using bedtools v.2.27.1 (Quinlan

255   & Hall, 2010) and a custom perl script cov.per.window.pl. Coverage statistics were then calculated in

256   R (R Core Team, 2017).

257

258   *Linkage map synteny*

259   In addition, we were able to assess the reliability of scaffolding by investigating the synteny between

260   the 40 scaffolds from each assembly and the *C. sp. "Albock"* linkage groups (De-Kayne & Feulner,

261   2018). RAD loci (90 bp containing a marker) with a known position in the linkage map were mapped

262   to the 40 scaffolds constituting each of the three assemblies using BWA-MEM v.0.7.17 (Li & Durbin,

263   2010; with default parameters). Synteny plots were then visualized using the *circlize* package in R

264   (Gu et al., 2014; R Core Team, 2017).

265

266   **Repeat masking and genome annotation**

267   To characterize the repeat landscape of the whitefish genome we first produced a repeat library using

268   RepeatModeler v.1.0.11 (Arian F. A. Smit & Hubley, 2008) for each of the haplotig-purged

269   assemblies. These libraries were then combined with a Salmonidae repeat library (from RepeatMasker

270   repeat database; queryRepeatDatabase.pl -species Salmonidae) to produce one concatenated reference

271   library. Each of the scaffolded assemblies was then repeat masked using this concatenated library

272   with RepeatMasker v.4.0.7 (A. F. A. Smit, Hubley, & Green, 2015). An interspersed repeat landscape

273   was then produced for the best assembly, from wtdbg2, using the RepeatMasker scripts

274   calcDivergenceFromAlign.pl and createRepeatLandscape.pl.

275

276   Annotation of the wtdbg2 assembled genome was carried out using a three-pass iterative approach

277   with MAKER2 v.2.31 (Holt & Yandell, 2011). First, an initial gene model was made using our repeat

278   library (described above), protein evidence from *Salmo salar* (UPID: UP000087266) and

279   *Oncorhynchus mykiss* (UPID: UP000193380) proteomes from Uniprot and the Swissprot vertebrate

280   database (uniprot_sprot_vertebrates), a recently published whitefish transcriptome (Carruthers et al.,

281   2018) and alternative transcriptome evidence from a *Danio rerio* transcriptome (TSA:

282   GDQQ01000001:GDQQ01083602). Next, this gene model was used to produce HMMs with SNAP

283   (Korf, 2004) and Augustus v.3.2.1 (Stanke, Diekhans, Baertsch, & Haussler, 2008). A second pass of

284   MAKER2 was then carried out using these ab initio gene prediction models and the models were

285   optimized before a final third MAKER2 pass was carried out. These final MAKER2 gene models

286   were filtered to remove spurious genes with AED scores < 0.6 (in accordance to Campbell, Holt,

287   Moore & Yandell 2014). Finally, functional annotation of this gene set was carried out using

288   Pannzer2 (Törönen, Medlar, & Holm, 2018) and the accuracy of the annotation was determined using

289   BUSCO v.3.0.2 (Simão et al., 2015; with the core gene set from ray-finned fishes

290   actinopterygii_odb9) on the final gene set. To further evaluate our gene set we used Orthofinder

291   v2.3.11 (Emms & Kelly 2015, 2019) to construct orthologous gene sets. This analysis included

292   protein sequences from whitefish (42,695 genes annotated with an AED < 0.6 and positioned on the

293   40 wtdbg2 whitefish scaffolds) and three other salmonids (*Hucho hucho* (ASM331708v1;

294   GCA_003317085.1; submitted by University Of Aberdeen on July 2018), *Salmo salar* (ICSASG_v2;

295   GCA_000233375.4; submitted by International Cooperation to Sequence the Atlantic Salmon

296   Genome on June 2015), *Salmo trutta* (fSalTru1.1; GCA_901001165.1; submitted by SC on June

297   2019) and the outgroup (which did not go through the Ss4R whole genome duplication) *Esox lucius*

298   (Eluc_V3; GCA_000721915.3; submitted by Ben F Koop, Jong S Leong on January 2017). All

299   protein files were downloaded from ENSMBL (https://www.ensembl.org/index.html; 21.Feb

300   2020).

301

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.

315

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.

327

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

357   dropped slightly (by 0.3%) in the wtdbg2 assembly (Table S1). The high completeness percentage of

358   BUSCOs for each of the assemblies prior to scaffolding suggests that we have succeeded in capturing

359   a large proportion of the whitefish genome sequence during the assembly process.

360

361   **Genome scaffolding**

362   Hi-C scaffolding of contigs into scaffolds resulted in a set of 40 scaffolds, many of which were

363   chromosome-scale, for each of the three assemblies, containing 2.38 Gb (96% of all sequence;

364   Falcon), 2.41 Gb (98% of all sequence; Canu) and 2.07 Gb (94% of all sequence; wtdbg2). The

365   differential log-likelihood calculation, which was used identify the length of confidently ordered and

366   orientated scaffolds, resulted in 1.08 Gb (45.41%) for the Falcon scaffolds (FSs), 1.21 Gb (50.1%) for

367   the Canu scaffolds (CSs), and 1.74 Gb (84.1%) for the wtdbg2 scaffolds (WFSs), meeting this

368   criterion. These results are in agreement with the patterns observable in the final scaffold heatmaps

369   for each assembly, in which the patterns observable for the wtdbg2 scaffolds are more in alignment

370   with a priori expectations about Hi-C linkage density patterns (Figure 3; contact plots for the Falcon

371   and Canu assemblies are displayed in Figures S1 and S4; van Berkum et al., 2010), yielding

372   qualitative confirmation of the quantitative scaffold quality assessment. The percentage of complete

373   BUSCOs went up for each assembly following Hi-C scaffolding, with the Falcon, Canu, and wtdbg2

374   assemblies now having 91.8%, 93.7%, and 93.3% (Table 1). When considering only scaffolds, the

375   Canu assembly retained the highest complete percentage of BUSCOs at 93.7% with the Falcon and

376   wtdbg2 assemblies dropping only slightly to 91.5% and 93.0% each. Based on having the highest

377   length of confidently scaffolded contigs and the high number of complete BUSCOs, the Hi-C

378   scaffolded wtdbg2 assembly was selected as the best of the three assemblies and was uploaded to the

379   European Nucleotide Archive (Accession: GCA_902810595.1). Falcon and Canu assemblies are

380   available on Dryad (https://doi.org/10.5061/dryad.xd2547ddf).

381

382   **Validation of whitefish assemblies**

383   *Illumina short read mapping*

384   Summaries of the mapping of Illumina reads to each of the assemblies can be seen in Table S2. The

385   wtdbg2 assembly had the highest number of mappings over MAPQ 30 (71.6%) with the Canu and

386   Falcon assemblies having slightly lower proportions of high-quality mappings (69.5% and 60.7%

387   each). When considering the proportion of read mates mapped to a different scaffold, however, the

388   Canu assembly looks the best of the three with only 1.4% of mates mapped to a different scaffold

389   with a MAPQ > 5, compared to 2% for Falcon and 2.4% for wtdbg2. As a result of our coverage

390   analysis the highest mean coverage (in 30 Kb windows) was 17.3x observed in the wtdbg2 assembly.

391   The Falcon and Canu assemblies had lower mean coverages of 15.0x and 15.9x, respectively. Plots of

392   coverage across the 40 wtdbg2 WFSs are shown in Figure 4 and the equivalent plots for the Falcon

393   and Canu assemblies in Figures S2 and S5. Based on these coverage plots we identified regions which

394   are likely to represent collapsed duplicated regions, spread across each genome assembly (Table S3

395   and Table S4). As expected in these regions, the coverage was approximately double that of the rest

396   of the assembly. In other salmonid genome assemblies, which have successfully resolved each copy

397   of a duplicated region, these duplicated regions typically span whole chromosome arms or even

398   chromosomes. Similarly, in the whitefish assemblies we identify collapsed blocks which encompass

399   whole scaffolds or parts of scaffolds. In the wtdbg2 assembly some WFSs (WFS4, 7, 14 and 37)

400   likely represent collapsed regions which span chromosome arms, and in other WFSs (WFS22, 28, 32,

401   36, and 38) the whole chromosome appears collapsed (a BED file containing the locations of these

402   wtdbg2 collapsed duplicates is included in Table S3). We estimate that in total 309 Mb of the wtdbg2

403   assembly (representing 14% of the assembly) is collapsed compared to 413 Mb and 517 Mb in the

404   Canu and Falcon assemblies, respectively (representing 17% and 21% of each assembly; Table S4).

405

406   *Linkage map synteny*

407   Out of the 5395 markers from the *C. sp. "Albock"* linkage map which were mapped to each assembly,

408   only high-quality mappings (MAPQ > 30) were retained, resulting in a mapping position for 3648,

409   4494, and 4744 markers in the sequence of the Falcon, Canu, and wtdbg2 assemblies respectively

410   (Table S2). For all three assemblies, concordance between sequence and recombination position

411   across the majority of markers was very high, suggesting a high synteny between linkage groups and

412   scaffolds. In the wtdbg2 assembly 95% of markers (4489/4744) showed strong synteny between one

413   linkage group and one WFS (38 out of the 40 linkage groups; Figure 5; equivalent to figures S3 and

414   S6 for the Falcon and Canu assemblies, respectively). Only two scaffolds, (WFS38 and 40) could not

415   be matched to any linkage group. We also identified a series of substantial deviations from the

416   broader pattern, where a number of markers from a linkage group also mapped to a second, alternative

417   scaffold. This was the case for markers from Calb01 – WFS37, Calb02 – WFS32, Calb08 – WFS38,

418   Calb13 – WFS35, Calb16 – WFS06, Calb20 – WFS28, Calb34 – WFS07, and Calb36 – WFS04.

419   Strikingly the mapping location of seven of these deviations (WFS04, 07, 28, 32, 35, 37, and 38) also

420   represent seven of the nine scaffolds we identified as collapsed duplicates showing inflated coverage

421   (shown in grey in Figures 4, 5 and 7). Although part of WFS17 looked like a collapsed duplicate

422   based on coverage, no significant deviations of markers mapped to this scaffold. Additionally, despite

423   having an unusual mapping pattern with markers from Calb16 in addition to those from Calb33,

424   WFS06 showed a consistent coverage of Illumina reads mapping. Out of all deviating markers, 83%

425   (212/255) mapped to regions identified as collapsed duplicates. No markers from the linkage map

426   were successfully mapped to WFS40, the smallest of the scaffolds, at only 1.1 Mb long. Additionally,

427   in two cases markers from two linkage groups predominantly mapped to one WFS. Markers from

428   both Calb35 and Calb40 mapped to WFS31 and from Calb38 and Calb39 to WFS22. The few

429   deviations from the patterns of synteny between linkage groups and scaffolds could be caused by a

430   number of interacting factors. These include the potential collapse of some scaffolds or parts of

431   scaffolds, small mistakes in either the linkage map or the sequence assembly, and the distribution of

432   repetitive sequence similar regions which all reduce the accuracy of mapping the short (100 bp) RAD

433   loci from the linkage map. However, collapsed regions likely play the most significant role in driving

434   deviations in the observed patterns of synteny as evidenced by the fact that the majority of deviating

435   RAD loci map to regions that are thought to be collapsed.

436

437   **Repeat masking and genome annotation**

438   Around 52% of each assembly was masked with the most abundant repetitive elements being DNA

439   elements followed by LINEs and then unclassified repeats (Table S1). DNA elements alone made up

440   nearly a quarter of each assembly (24.65% of Falcon, 23.79% of Canu, and 24.41% of the wtdbg2

441   assembly). The resulting landscape (Figure 6) identified the Class II TE superfamily Tc1-*mariner* as

442   the most abundant in the *Coregonus sp. "Balchen"* genome, making up 18% of the interspersed

443   repeats. The most abundant Class I TEs were LINE-2 elements, however, these only made up 4.2% of

444   the interspersed repeats. The three-pass MAKER2 annotation resulted in the identification of 44,525

445   protein-coding genes (42,695 on scaffolds and 1,830 on unscaffolded contigs) and included 357,479

446   identified exons (Table 2), with the final set of genes 81.8% complete for BUSCOs

447   (C:81.8%[S:54.8%,D:27.0%],F:9.3%,M:8.9%,n:4584). Functional annotation with Pannzer2 allowed

448   the assignment of gene ontology terms to 29,046 genes. Across a total of 415,276 genes in the five

449   species, we identified 41,042 orthogroups. This includes 7,673 species-specific orthogroups, 725

450   single-copy orthologs, and 16,599 orthogroups with all five species present. Out of the 42,695

451   whitefish genes (on scaffolds) 38,219 could be assigned to 22,311 orthogroups. The number of

452   whitefish genes annotated on scaffolds (42,695) is similar to the number of genes annotated in the

453   diploid outgroup (*Esox lucius* 43,143 genes), but substantially lower than the ENSMBL annotation of

454   any of the other three salmonids (*Hucho hucho* 91,817 genes, *Salmo salar* 121,064 genes, *Salmo*

455   *trutta* 116,557 genes). This suggests that more transcriptomic work on various tissues and

456   developmental stages is warranted to further improve the annotation of the whitefish genome. We

457   found 4,504 orthogroups that were duplicated in whitefish (present with more than one copy) that

458   were present in a single copy in pike (*Esox lucius*). A total of 2,746 orthogroups show duplication

459   patterns relative to the pike consistent across all four salmonids. However, we also identify 3,459

460   orthogroups where only the other three salmonids show duplications and 474 orthogroups where only

461   whitefish appear duplicated. This could reflect true biological differences in duplication loss in the

462   different salmonid lineages, however technical artefacts during the assembly (collapsing of highly

463   identical regions) or gene annotation differences likely contribute as well.

464

465   **Identification of homeologous regions in the whitefish genome**

466   Using SyMAP we identified 55 syntenic links between 34 of the 40 WFSs (Figure 7; Table S5).

467   Sequence similarity calculations for each link (110 mappings, one in each direction, for each of the 55

468 homeologous blocks) showed that the majority of identified syntenic blocks had sequence similarity

469 ≥ 90% (shown in orange and red in Figure 7). Slightly lower sequence similarity was observed for

470 syntenic links between WFS02 and WFS03 and multiple smaller links between WFS01 and WFS31,

471 WFS07 and WFS34, and WFS08 and WFS17. The WFSs for which no syntenic links were identified

472 were WFS22, WFS28, WFS32, WFS38, WFS39, and WFS40. Of those, WFS22, WFS28, WFS32,

473 and WFS38 have been identified as fully collapsed based on their unusual coverage patterns, hence

474 likely represent residual tetraploid regions or a collapse of two very sequence similar homeologs. It is

475 likely that the shortest scaffold, WFS40, constitutes part of a chromosome, explaining why no

476 homeologous scaffold was found.

477

478 **Ancestral chromosome identification**

479 As above, SyMAP was used to identify the corresponding non-duplicated ancestral chromosome for

480 each syntenic WFS pair. This comparison of the WFSs and the Northern Pike genome identified that

481 the majority of pike chromosomes (18 out of 25) had two equivalent WFSs, the remaining seven pike

482 chromosomes match to only one WFS (Table 3; Figure S7). These seven pike chromosomes included

483 PK2, PK9, PK11, PK20, and PK23 which corresponded to WFS32, WFS38, WFS35, WFS28, and

484 WFS4. Additionally, PK22 and PK25 are both in single copy in whitefish but appear to have been

485 rearranged into a single WFS, WFS22.

486

487 **Discussion**

488 To enable future studies to investigate both the genetic basis of adaptation and speciation within

489 Coregoninae and genome evolution following whole genome duplication across the Salmonidae

490 family we have assembled the first whitefish reference genome. Assembling > 90x PacBio data from

491 one female *Coregonus sp. "Balchen"* with three of the most commonly used assemblers resulted in

492 three high quality assemblies, each with >90% complete BUSCOs and 40 scaffolds, many of which

493 are chromosome-scale. Out of these three assemblies we judge the assembly produced by wtdbg2 as

494 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

523   of confidently mapped linkage map markers. Although the complete BUSCO scores were slightly

524   lower than that of the Canu assembly (93% in the wtdbg2 assembly compared to 93.7% in the Canu

525   assembly), the increased confidence of scaffolding in the wtdbg2 assembly, the superior mapping

526   metrics, the lowest proportion of the genome being collapsed, and highest synteny with the linkage

527   map, all led to us selecting this as the best assembly.

528

529   The scaffold N50 of the wtdbg2 *C. sp. "Balchen"* assembly was 51.9 Mb, which is higher than a

530   number of recently published salmonid genomes including Chinook salmon (1.138 Mb; Christensen,

531   Leong, et al., 2018), Arctic charr (1.02 Mb; Christensen, Rondeau, et al., 2018) and grayling (33 Mb;

532   Sävilammi et al., 2019). The characterization of the repeat landscape of the whitefish genome also

533   highlighted the broad similarity in the proportion of many families of transposable elements between

534   salmonid species. We identified that around 52% of the whitefish genome is repetitive, a similar

535   proportion to that of Chinook salmon (56%; Christensen, Leong, et al., 2018), Arctic charr (56%;

536   Christensen, Rondeau, et al., 2018), and European grayling (47%; Sävilammi et al., 2019). The

537   relative abundances of different types of repetitive element are similar to those reported in other

538   salmonid assemblies including that of Atlantic Salmon (Lien et al., 2016) and Chinook Salmon

539   (Christensen, Leong, et al., 2018). The relatively high abundance of Class II TE superfamily Tc1-

540   *mariner* and LINE-2 elements amongst the youngest elements suggest that these families are still

541   expanding and potentially diversifying in the whitefish genome. Conversely, the lack of new LTR

542   elements suggests that their abundance and diversity peaked in the past and they are no longer

543   diversifying in the genome. Annotation of the wtdbg2 assembly identified 44,525 genes, similar to

544   those reported in the publications associated with the rainbow trout genome (46,585 by Berthelot et

545   al., 2014; 53,383 by Pearse et al., 2019) but higher than the 37,206 genes identified by Lien et al.

546   (2016) in the Atlantic salmon genome and the 36,216 identified by Christensen, Leong et al, (2018) in

547   the Chinook salmon genome.

548

549   Although whole genome duplications have punctuated the tree of life, few have occurred recently

550   enough to allow investigations into the subsequent rediploidization process at the genomic level.

551 Salmonids are therefore an ideal family in which to study rediploidization since the genomic signals

552 of whole genome duplication and genomic rearrangements which followed have not yet been

553 confounded by other genomic processes such as mutations (including small point mutations and large

554 structural changes such as inversions and deletions; Macqueen & Johnston, 2014). One recent

555 investigation into rediploidization within Salmonidae already identified substantial genomic

556 differences between 16 salmonid species, which evolved independently as rediploidization proceeded

557 (Robertson et al., 2017). However, with high-quality genomes available for an increasing number of

558 salmonid species, the resolution with which we can identify differences which have occurred in

559 genome structure and composition following whole genome duplication is vastly increasing.

560 Therefore, highly contiguous reference genomes, particularly for underrepresented groups such as the

561 salmonid subfamily Coregoninae, are invaluable to fill gaps in the genomic resources currently

562 available. Here, we have been able to determine pairs of whitefish scaffolds that represent homeologs

563 and their corresponding ancestral chromosome (using Northern Pike chromosome numbering). This

564 will facilitate future comparisons across the salmonid family (similar to those by Blumstein et al.,

565 2020) to assess the independent rediploidization process in different salmonid lineages. We also

566 identified a number of whitefish scaffolds for which no homeolog was present. By combining synteny

567 data with our coverage-based validation of the new *C. sp. "Balchen"* assembly we showed that some

568 of these regions were due to the collapsing of highly sequence similar regions (e.g. WFS22 and

569 WFS28). In other instances, we rather identified potential genomic rearrangements as the driver of

570 this pattern, for example, pike PK24 showed homology with WFS39 but also a part (potentially a

571 chromosome arm) of WFS08 (Table 3; Figure S7). More complex patterns were also identified such

572 as the merging of two different PK chromosomes (22 and 25) into one whitefish scaffold WFS22 that

573 showed inflated coverage estimates across the scaffold suggesting that the both homeologs of the

574 fused PK chromosomes were collapsed in our assembly due to their high sequence similarity.

575 Although their high sequence similarity makes duplicated regions difficult to assemble and

576 subsequently sometimes causes their collapse during assembly, we have assembled a highly

577 continuous reference genome. This is despite a high sequence similarity between large parts of the

578 genome, with at least 80% of the genome being $\geq$ 90% sequence similar. Our estimates of sequence

579   similarity are comparable to observations in Atlantic salmon where it has been shown that 94% of

580   chromosome sequence is duplicated, with 26% of the genome having a duplicate region with

581   sequence similarity >90% (Lien et al., 2016). Future work should aim to investigate the partially

582   collapsed scaffolds WFS04, WFS07, WFS17 and WFS37 and collapsed regions which span the length

583   of scaffolds WFS22, WFS28, WFS32, WFS35, and WFS38, which we identified (Figure 4; Figure 7;

584   Table S3), to determine the evolutionary history of these duplicates and the process by which the non-

585   duplicated regions of the whitefish genome may have rediploidized independently, or not, compared

586   to other salmonid species. Specifically, it should be determined whether the lack of identified

587   homeologs for WFS22, WFS28, WFS32, WFS35 and WFS38 is an artefact of our assembly (as

588   indicated by their increased coverage; Figure 4; for example for WFS28, WFS32, WFS38), the result

589   of genomic rearrangements resulting in coregonid-specific arrangement (such as WFS39), or a

590   combination of the two (such as WFS22).

591

592   In addition to facilitating the investigation of salmonid genome evolution, the highly contiguous

593   whitefish assembly presented here will also support future genomic studies within the subfamily

594   Coregoninae. Coregoninae are distributed across the Northern hemisphere (North America and

595   Eurasia), widely fished and of economic importance, and exhibit an extraordinary ecological

596   diversity. Studying whitefish diversification is of fundamental scientific interest to understand the

597   processes driving and facilitating such diversification and to assist in the conservation of this diverse

598   group. We anticipate that the whitefish genome assembly presented here will aid future investigations

599   into the ecology and evolution of all whitefish. Specifically, it will facilitate investigations into the

600   genetic basis of adaptation across the Alpine whitefish radiation including determining the level of

601   parallelism across multiple pre-alpine lake systems at a genome-wide resolution.

602

603

604

605

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 populatinos 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 artedi*) 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.

687 Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E.
688 L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell*
689 *Systems*, *3*, 99–101.
690
691 Emms, D.M., & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative
692 genomics. *Genome Biology* 20:238
693
694 Emms, D.M., & Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome
695 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16:157
696
697 Faust, G. G., & Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read
698 extraction. *Bioinformatics*, *30*, 2503–2505.
699
700 Feulner, P. G. D., & Seehausen, O. (2019). Genomic insights into the vulnerability of sympatric
701 whitefish species flocks. *Molecular Ecology*, *28*, 615–629.
702
703 Gagnaire, P.-A., Normandeau, E., Pavey, S. A., & Bernatchez, L. (2013). Mapping phenotypic,
704 expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish
705 *(Coregonus clupeaformis)*. *Molecular Ecology*, *22*, 3036–3048.
706
707 Gu Z., Gu L., Eils R., Schlesner M., Brors B. (2014) circlize implements and enhances circular
708 visualization in R. *Bioinformatics 30*, 2811–2812.
709
710 Hardie, D. C., & Hebert, P. D. (2003). The nucleotypic effects of cellular DNA content in
711 cartilaginous and ray-finned fishes. *Genome*, *46*, 683-706.
712
713 Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management
714 tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491.
715
716 Hudson, A. G., Lundsgaard-Hansen, B., Lucek, K., Vonlanthen, P., & Seehausen, O. (2017).
717 Managing cryptic biodiversity: Fine-scale intralacustrine speciation along a benthic gradient in Alpine
718 whitefish (*Coregonus* spp.). *Evolutionary Applications*, *10*, 251–266.
719
720 Hudson, A. G., Vonlanthen, P., & Seehausen, O. (2010). Rapid parallel adaptive radiations from a
721 single hybridogenic ancestral population. *Proceedings of the Royal Society B: Biological Sciences*,
722 *278*, 58-66.
723
724 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu:
725 scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
726 *Genome Research*, *27*, 722–736.
727
728 Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59.
729
730 Kottelat, M., & Freyhof, J. (2007). *Handbook of European freshwater fishes*. Cornol, Switzerland:
731 Publications Kottelat
732
733 Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., …
734 Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of
735 the human genome. *Science*, *326*, 289–293.
736
737 Lien, S., Koop, B. F., Sandve, S. R., Miller, J. R., Kent, M. P., Nome, T., … Davidson, W. S. (2016).
738 The Atlantic salmon genome provides insights into rediploidization. *Nature*, *533*, 200–205.
739
740 Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform.
741 *Bioinformatics*, *26*, 589–595.

742
743 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data
744 Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*,
745 *25*, 2078–2079.
746
747 Lockwood, S. F., Seavey, B. T., Dillinger, R. E., Jr., & Bickham, J. W. (1991). Variation in DNA
748 content among age classes of broad whitefish (*Coregonus nasus*) from the Sagavanirktok River delta.
749 *Canadian Journal of Zoology*, *69*, 1335–1338.
750
751 Lu, G., & Bernatchez, L. (1999). Correlated trophic specialization and genetic divergence in
752 sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): Support for the ecological speciation
753 hypothesis. *Evolution*, *53*, 1491–1505.
754
755 Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid
756 whole genome duplication reveals major decoupling from species diversification. *Proceedings of the*
757 *Royal Society B: Biological Sciences*, *281*, 20132881.
758
759 Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of
760 occurrences of k-mers. *Bioinformatics*, *27*, 764–770.
761
762 Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., ... & Smith, W. L.
763 (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the*
764 *National Academy of Sciences*, *109*, 13698-13703.
765
766 Østbye, K., Bernatchez, L., Naesje, T. F., Himberg, K.-J. M., & Hindar, K. (2005). Evolutionary
767 history of the European whitefish *Coregonus lavaretus* (L.) species complex as inferred from mtDNA
768 phylogeography and gill-raker numbers. *Molecular Ecology*, *14*, 4371–4387.
769
770 Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., ... & Moen, T.
771 (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature ecology &*
772 *evolution*, 3, 1731-1742.
773
774 Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic
775 features. *Bioinformatics*, *26*, 841–842.
776
777 R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R
778 Foundation for Statistical Computing. Retrieved from https://www.R-project.org/
779
780 Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment
781 for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*, 460.
782
783 Robertson, F. M., Gundappa, M. K., Grammes, F., Hvidsten, T. R., Redmond, A. K., Lien, S., …
784 Macqueen, D. J. (2017). Lineage-specific rediploidization is a mechanism to explain time-lags
785 between genome duplication and evolutionary diversification. *Genome Biology*, *18*, 111.
786
787 Ruan, J., & Li, H. (2019). Fast and accurate long-read assembly with wtdbg2. *BioRxiv*, 530972.
788
789 Sävilammi, T., Primmer, C. R., Varadharajan, S., Guyomard, R., Guiguen, Y., Sandve, S. R., ... &
790 Lien, S. (2019). The chromosome-level genome assembly of European grayling reveals aspects of a
791 unique genome evolution process within salmonids. *G3*, *9*, 1283-1294.
792
793 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015).
794 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.
795 *Bioinformatics*, *31*, 3210–3212.
796

797   Smit, A. F. A., & Hubley, R. (2008). RepeatModeler Open-1.0. *Available from Http://www.*
798   *repeatmasker.org*
799
800   Smit, A. F. A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. *Available from Http://www.*
801   *repeatmasker.org*
802
803   SMRT Link v.5.0.1 – Pacific Biosciences SMRT Tools Reference Guide.
804   https://www.pacb.com/support/software-downloads/
805
806   Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3.4: a turnkey synteny system with
807   application to plant genomes. *Nucleic acids research*, 39, e68-e68.
808
809   Soderlund, C., Nelson, W., Shoemaker, A., & Paterson, A. (2006). SyMAP: A system for discovering
810   and viewing syntenic regions of FPC maps. *Genome research*, 16, 1159-1168.
811
812   Sutherland, B. J., Gosselin, T., Normandeau, E., Lamothe, M., Isabel, N., Audet, C., & Bernatchez, L.
813   (2016). Salmonid chromosome evolution as revealed by a novel method for comparing RADseq
814   linkage maps. *Genome Biology and Evolution*, 8, 3600-3617
815
816   Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped
817   cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*, 637–644.
818
819   Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: a rapid functional annotation web server.
820   *Nucleic Acids Research*, *46*, W84–W88.
821
822   van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., ... &
823   Lander, E. S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE*
824   *(Journal of Visualized Experiments)*, (39), e1869.
825
826   Varadharajan, S., Sandve, S. R., Gillard, G. B., Tørresen, O. K., Mulugeta, T. D., Hvidsten, T. R., ...
827   & Jakobsen, K. S. (2018). The grayling genome reveals selection on gene expression regulation after
828   whole-genome duplication. *Genome biology and evolution*, *10*, 2785-2800.
829
830   Vonlanthen, P., Bittner, D., Hudson, A. G., Young, K. A., Müller, R., Lundsgaard-Hansen, B., …
831   Seehausen, O. (2012). Eutrophication causes speciation reversal in whitefish adaptive radiations.
832   *Nature*, *482*, 357–362.
833
834   Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz,
835   M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*,
836   *33*, 2202–2204.
837
838   Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014).
839   Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly
840   improvement. *PloS One*, *9*, e112963.
841
842
843
844
845
846

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

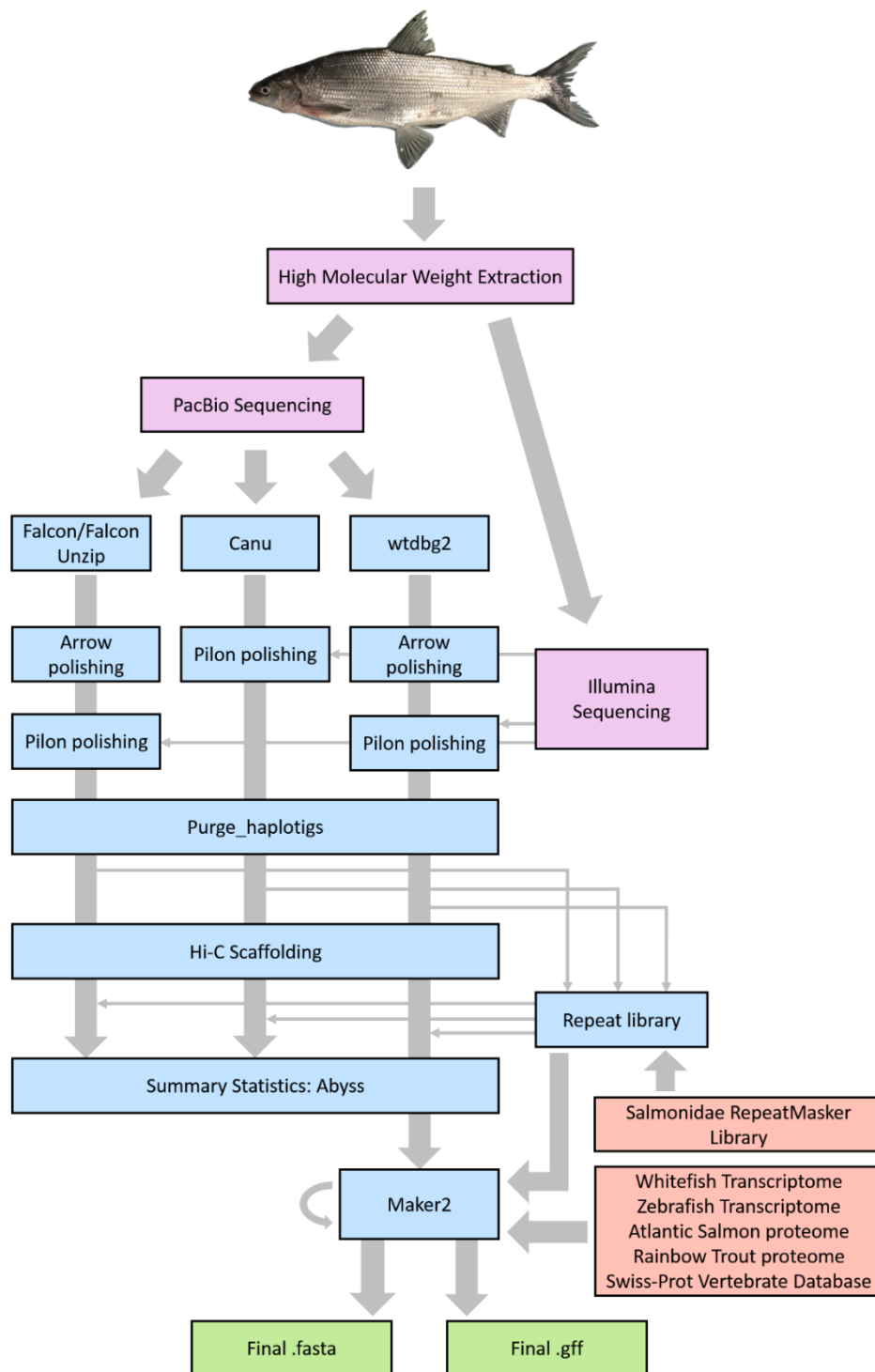the final version for submission.

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
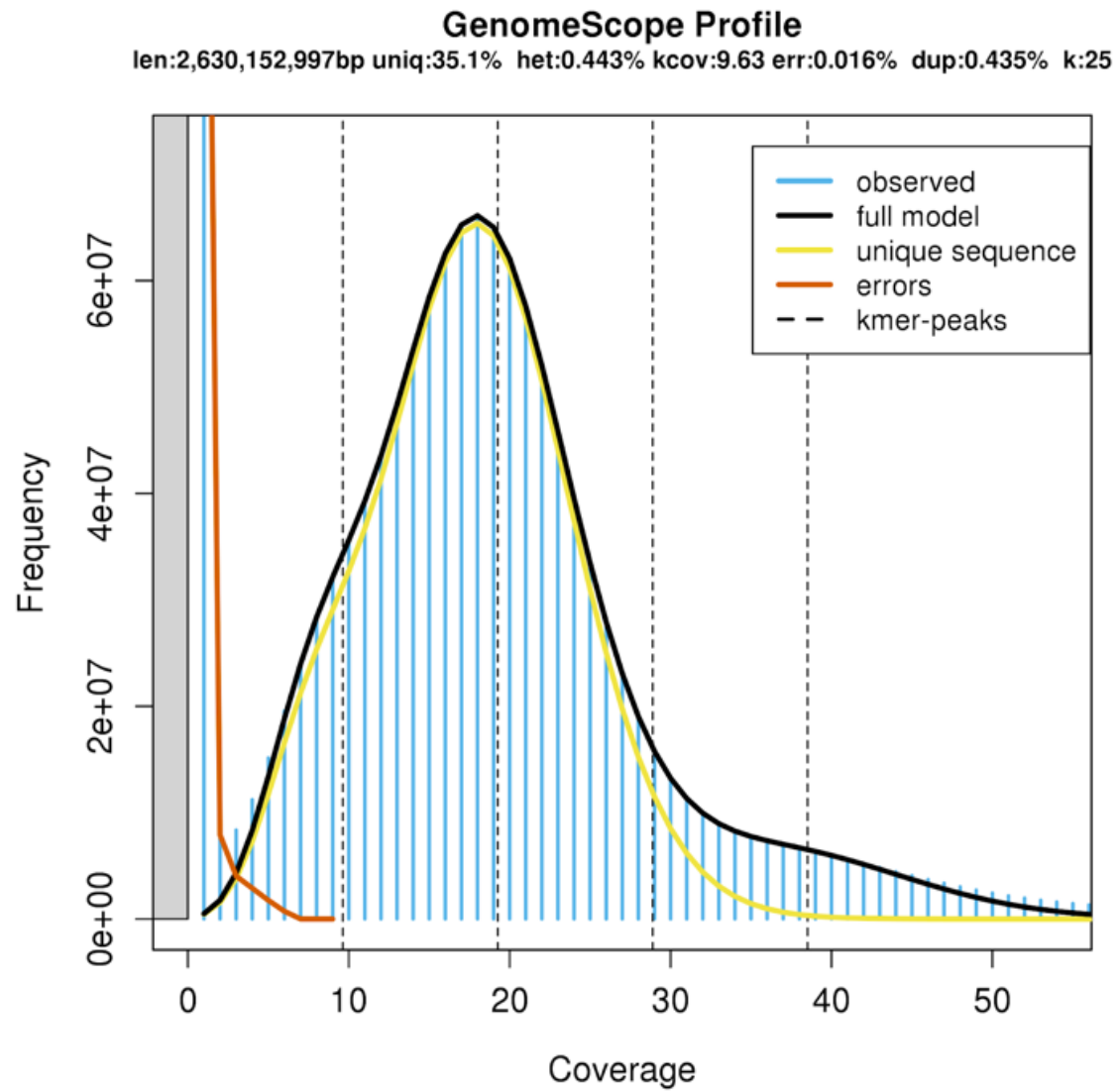outputs are shown in green.

GenomeScope Profile

len:2,630,152,997bp uniq:35.1% het:0.443% kcov:9.63 err:0.016% dup:0.435% k:25

888
889
890   Figure 2. GenomeScope profile established based on short read data, which estimates the
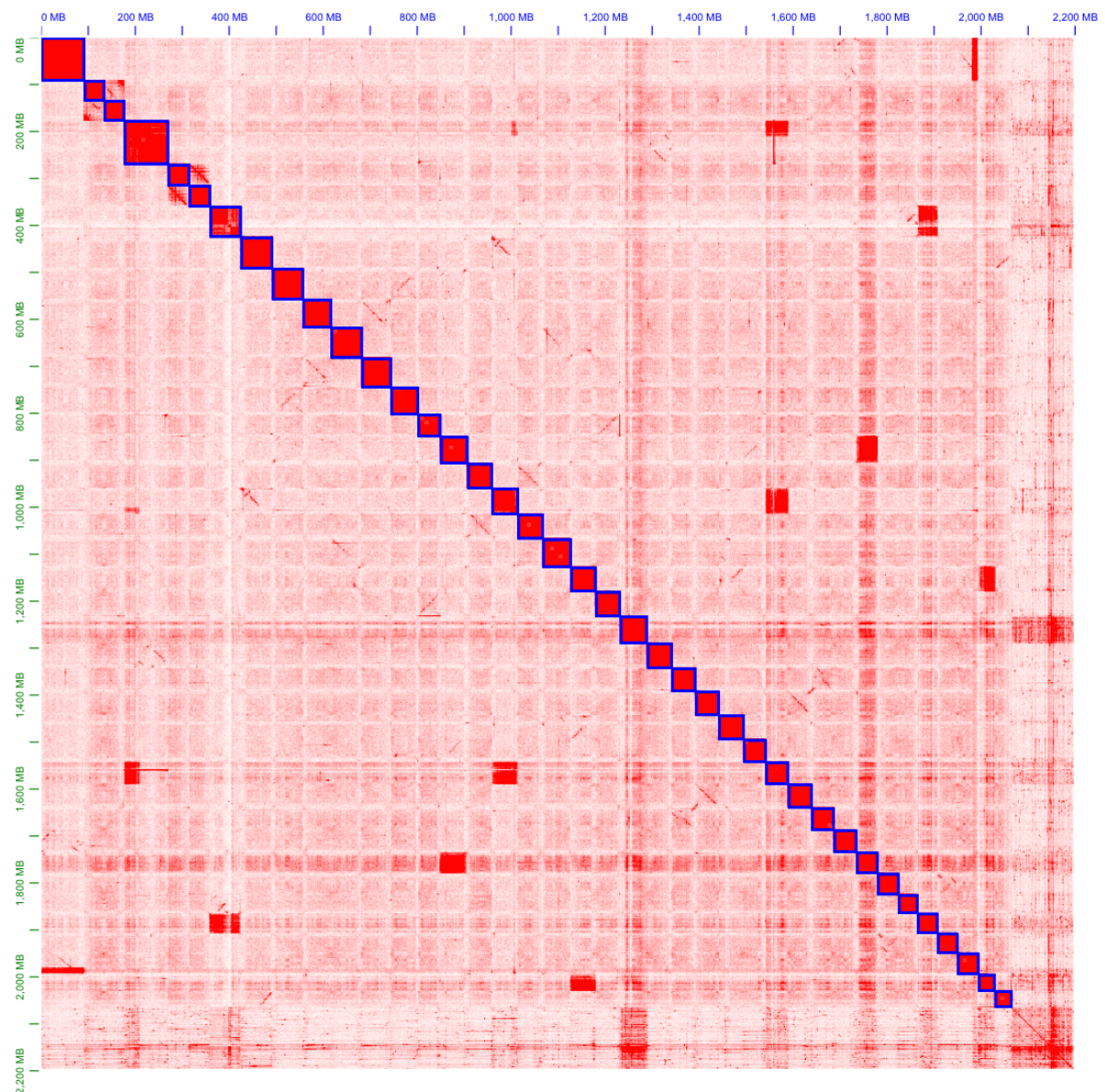891   genome size of *Coregonus sp. "Balchen"* to be 2.6 Mb.
892

893
894 Figure 3. *Coregonus sp. "Balchen"* contig contact map from Hi-C scaffolding of the wtdbg2
895 assembly. The intensity of red represents the relative contact density between contigs. The
896 highest contact density is found within whitefish scaffolds (WFSs), which are outlined in
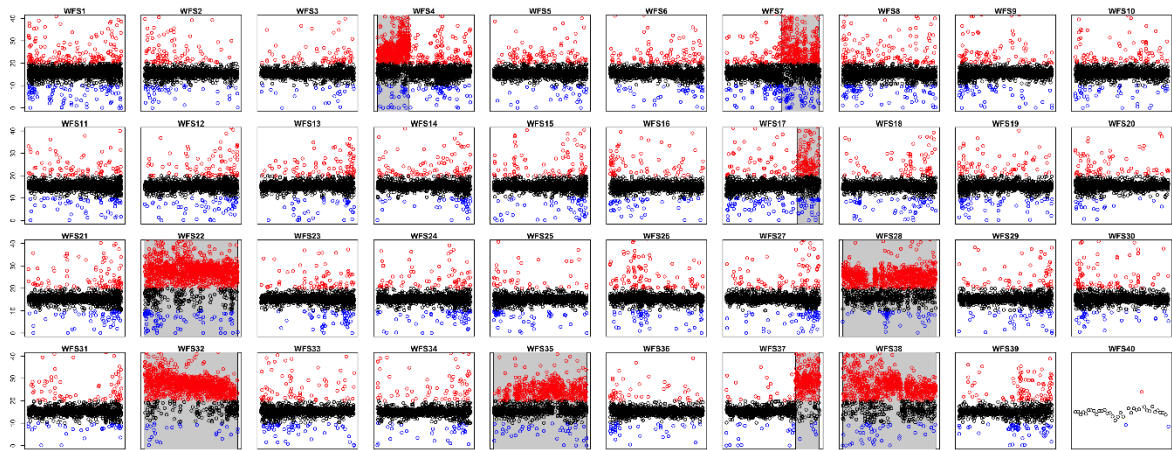897 blue.
898
899

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.
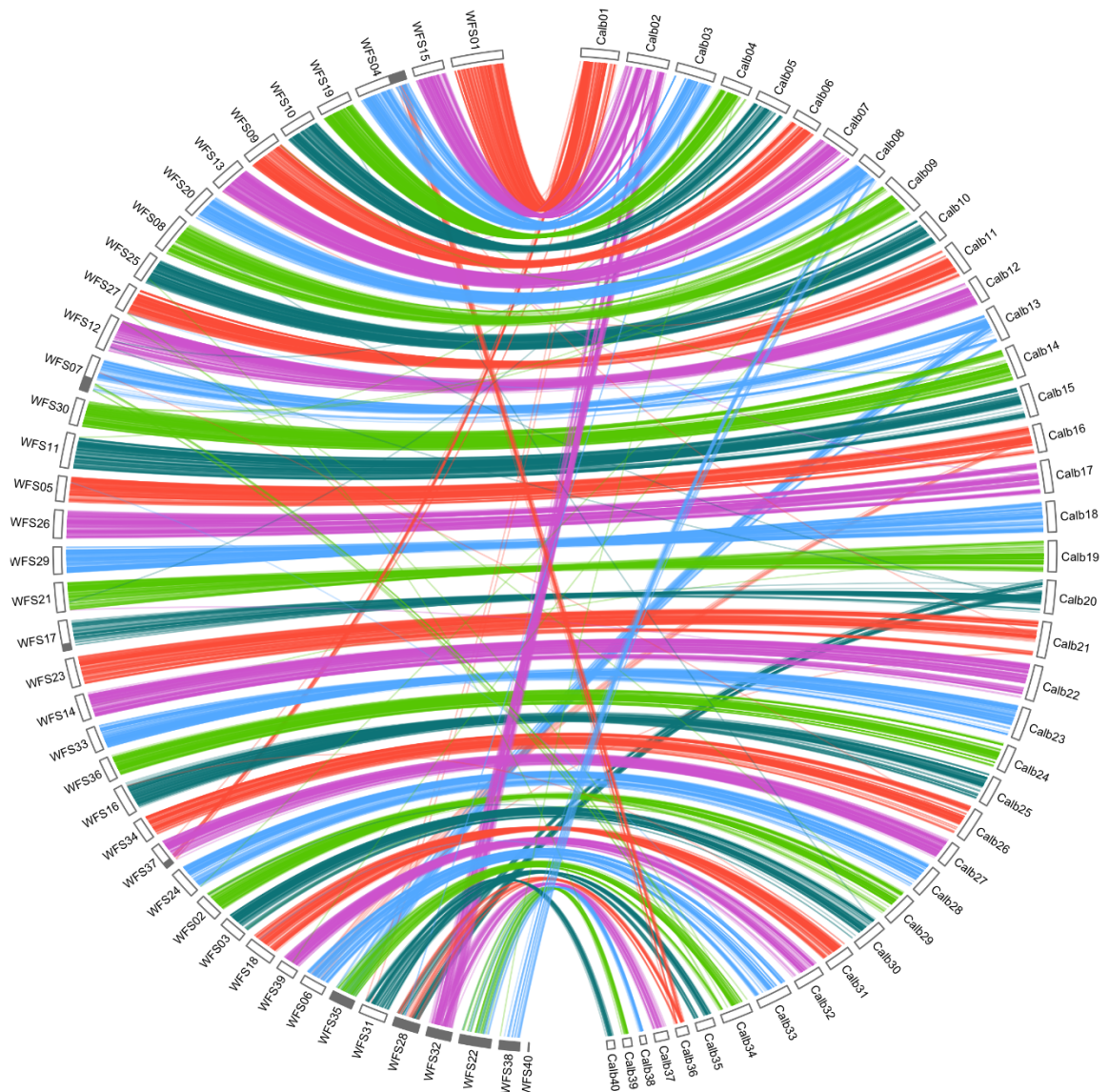
907
908 Figure 5. Circos plot comparing the structure of the *C. sp. "Albock"* linkage map (right; De-
909 Kayne & Feulner, 2018) and the 40 whitefish scaffolds (WFSs) of the wtdbg2 *C. sp.*
910 *"Balchen"* assembly (left). Lines indicate mapping locations of RAD loci from the linkage
911 map in the genome assembly. Most mappings suggest a good match between linkage map
912 and genome assembly (high synteny between linkage groups and WFSs) and only few lines
913 map discordantly. Genome assembly regions which represent collapsed duplicate regions are
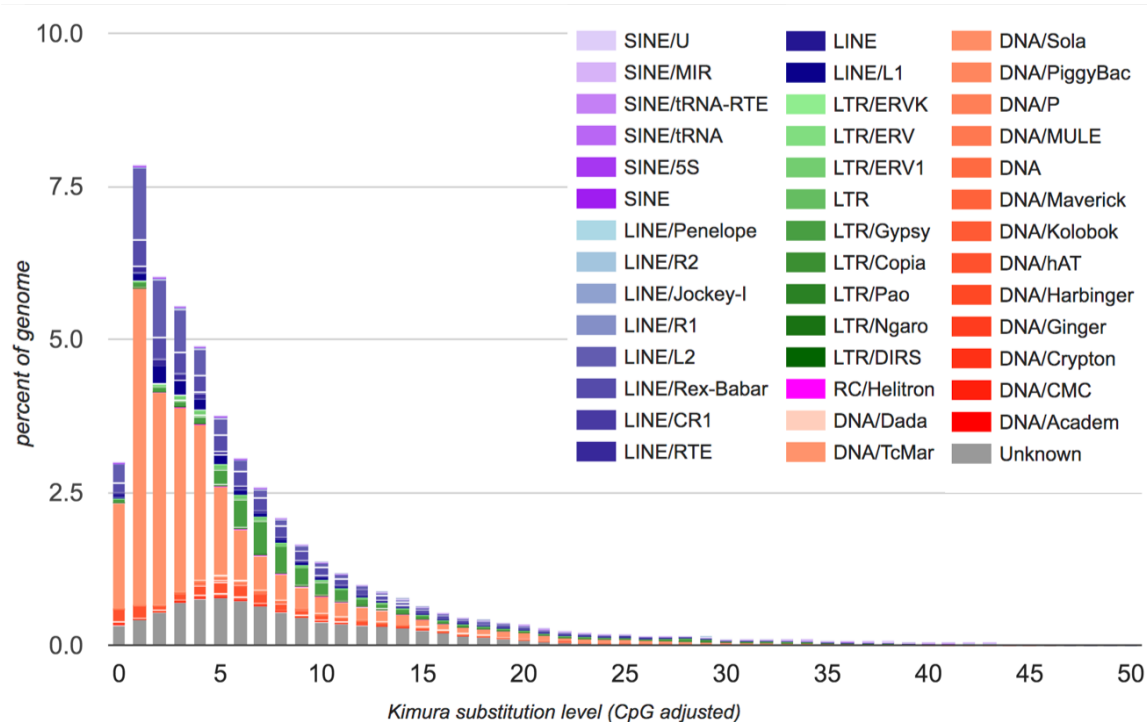914 identified in grey around the left perimeter.
915
916

917
918 Figure 6. *Coregonus sp. "Balchen"* transposable element divergence landscape.
919 Transposable elements within the whitefish genome have been characterised (different
920 classes represented by distinct colours). The plot shows the relative abundance of each class
921 and their relative age (molecular clock estimate). Note the ongoing DNA element
922 diversification within the whitefish genome, particularly in DNA elements and LINEs.
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
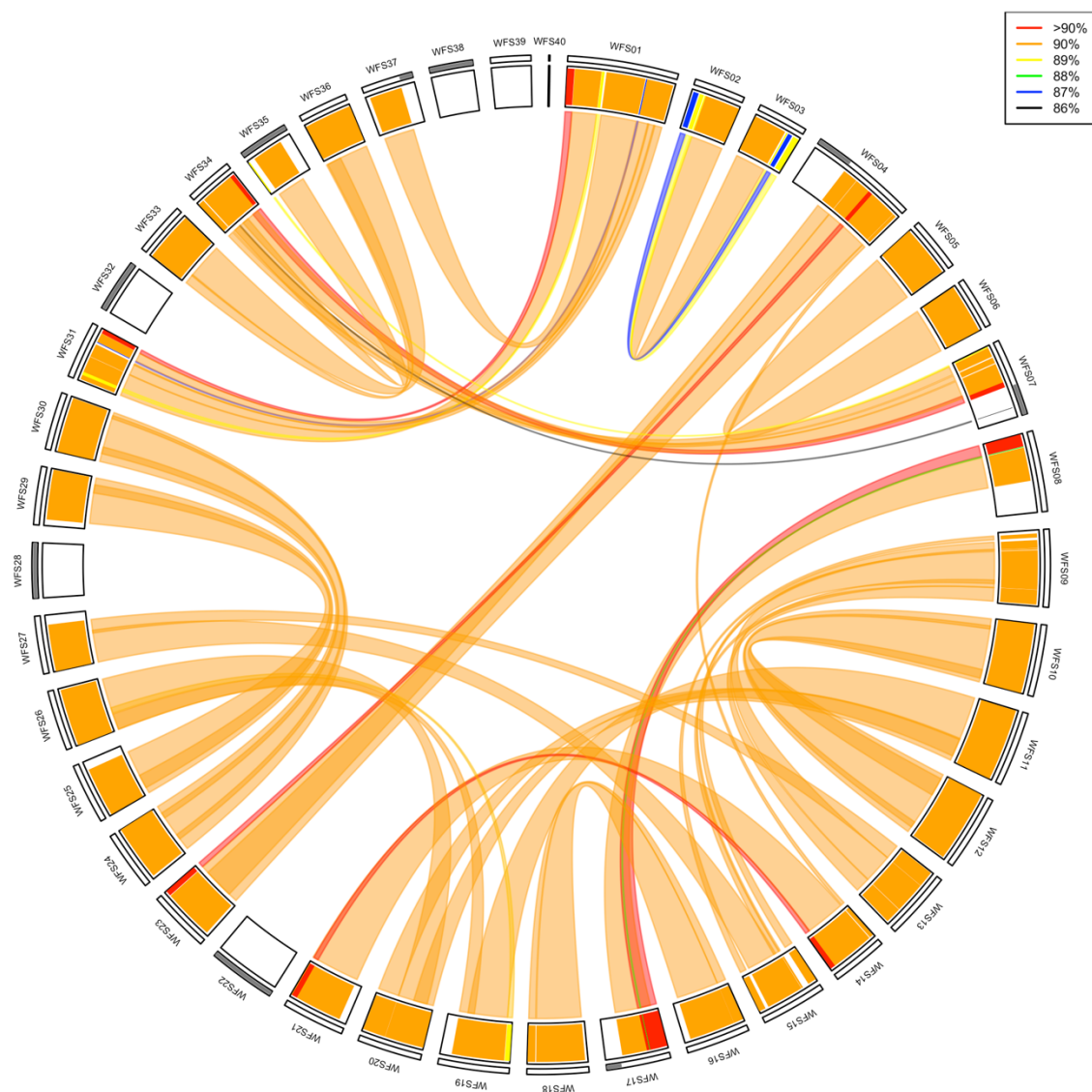939
940
941
942
943
944
945
946
947

948
949 Figure 7. Homeologous whitefish scaffolds (WFSs) within the whitefish genome identified
950 using SyMAP. Links between homeologs are coloured according to their mean sequence
951 similarity based on Lastz alignments for each syntenic link in both directions. Genome
952 assembly regions which are thought to be collapsed are identified in grey on the outer most
953 track.
954
955
956
957
958
959
960
961
962
963
964

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%)

980

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

982 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

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

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	-	-