Long-read sequence capture of the hemoglobin gene clusters across species

Siv Nam Khang Hoff*[a], Helle T. Baalsrud*[a‡], Ave Tooming-Klunderud[a], Morten Skage[a], Todd Richmond[b], Gregor Obernosterer[c], Reza Shirzadi[d], Ole Kristian Tørresen[a], Kjetill S. Jakobsen[a] and Sissel Jentoft[a‡]

[a] Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway
[b] Roche NimbleGen Inc., Madison, WI, USA
[c] Roche Diagnostics, Mannheim, Germany
[d] Roche Diagnostics, Oslo, Norway

* these authors contributed equally to this work

‡ Correspondence to: h.t.baalsrud@ibv.uio.no, sissel.jentoft@ibv.uio.no

Abstract

Combining high-throughput sequencing with targeted sequence capture has become an attractive tool to study specific genomic regions of interest. Most studies have so far focused on the exome using short-read technology. These approaches are not designed to capture intergenic regions needed to reconstruct genomic organization, including regulatory regions and gene synteny. Here, we demonstrate the power of combining targeted sequence capture with long-read sequencing technology for comparative genomic analyses of the hemoglobin (Hb) gene clusters across eight species separated by up to 70 million years. Guided by the reference genome assembly of the Atlantic cod (*Gadus morhua*) together with genome information from draft assemblies of selected codfishes, we designed probes covering the two Hb gene clusters. Use of custom-made barcodes combined with PacBio RSII sequencing led to highly continuous assemblies of the LA (~100kb) and MN (~200kb) clusters, which include syntenic regions of coding and intergenic sequences. Our results revealed an overall conserved genetic organization and synteny of the *Hb* genes within this lineage, yet with several, lineage-specific gene duplications. Moreover, for some of the species examined, we identified amino acid substitutions at two sites in the *Hbb1* gene as well as length polymorphisms in its regulatory region, which has previously been linked to temperature adaptation in Atlantic cod populations. This study highlights the use of targeted long-read capture as a versatile approach for comparative genomic studies by generation of a cross-species genomic resource elucidating the evolutionary history of the *Hb* gene family across the highly divergent group of codfishes.

Key words: Targeted sequence capture, comparative genomics, Gadiformes, PacBio sequencing, teleosts

42 # Introduction

43 The rapid advancement of high-throughput sequencing has over the last decade
44 revolutionized genomic research with the increasing numbers of whole genome
45 resources available for multiple vertebrate species, including the diverse group of
46 teleost fishes (Volff 2004; Ellegren 2014; Goodwin *et al.* 2016; Malmstrøm *et al.* 2016;
47 2017). However, whole genome sequencing (WGS) and generation of high quality
48 genome assemblies are still considered costly and time-consuming (Jones & Good
49 2015). For investigations concerning specific genomic regions there is no need for
50 complete genome information, which has spurred the development of reduction
51 complexity approaches such as targeted sequence capture (Teer *et al.* 2010; Grover *et*
52 *al.* 2012; Samorodnitsky *et al.* 2015). The basic idea of targeted sequence capture
53 involves design of specific probes covering the particular genomic area of interest
54 generating an enriched coverage of the targeted sequences (Turner *et al.* 2009; Grover
55 *et al.* 2012). Most studies using targeted sequence capture have to a large extent been
56 directed towards the exome, often supported by the existence of a reference genome
57 (Broeckx *et al.* 2014; Yoshihara *et al.* 2016), or transcriptome assemblies (Syring *et al.*
58 2016). Recent reports have, however, been focusing on off-target sequences in
59 noncoding regions (Guo *et al.* 2012; Syring *et al.* 2016; Yoshihara *et al.* 2016; Morin *et*
60 *al.* 2016), as they may contain crucial regulatory elements varying in sequence and
61 length between populations or species and could be of functional and evolutionary
62 importance (Woolfe *et al.* 2004).
63 To date, comparative studies using sequence capture have been mainly based
64 on short-read sequencing technology and probe design targeting genic regions
65 (George *et al.* 2011; Samorodnitsky *et al.* 2015; Bragg *et al.* 2015; Li *et al.* 2015).
66 Consequently, construction of continuous sequences enabling resolution of gene
67 organization across species has not yet been looked into. Comparative genetic
68 studies of gene organization or synteny requires longer, more continuous stretches of
69 DNA containing more than one gene (Huddleston *et al.* 2014). By its ability to span
70 long stretches of repeats, long-read sequencing technology has been successfully
71 applied to improve genome assembly statistics and generation of highly continuous
72 genome assemblies for a growing number of species (English *et al.* 2012; Kim *et al.*
73 2014; Tørresen *et al.* 2017; Kolmogorov *et al.* 2018; Tørresen *et al.* 2018). Incorporation
74 of long PacBio reads have for instance resulted in a significantly improved version of
75 the Atlantic cod (*Gadus morhua*) genome assembly, i.e. a 50-fold increase in sequence
76 continuity and a 15-fold reduction in the proportion of gap (Tørresen *et al.* 2017).
77 Furthermore, recent studies report the combination of capture and long read

78 sequencing as highly efficient in enriching and assembly of full length complex genes
79 as well as detailed characterization of chromosomal structural variations (Wang *et al.*
80 2015; Witek *et al.* 2016; Giolai *et al.* 2017). Correspondingly, utilizing long-read
81 sequencing technology in combination with targeted capture could yield longer
82 continuous assemblies of specific genomic regions of interest, allowing in-depth
83 comparative genetic studies including synteny analyses, in species where reference
84 genomes are not available.
85 In fishes, the hemoglobin (*Hb*) gene family, encoding the protein subunits Hba
86 and Hbb, has shown to be of importance for ecological adaptation, as environmental
87 factors such as temperature directly influences the ability of Hb to bind O_2 at
88 respiratory surfaces and its subsequent release to tissues (Wells 2005). In a recent
89 report, a characterization of the *Hb* gene repertoire by comparative draft genome
90 analysis uncovered a remarkably high *Hb* gene copy variation within the codfishes
91 (Baalsrud *et al.* 2017). Based on the gene copy number a negative correlation between
92 the number of *Hb* genes and depth of which the species occur was observed, as well
93 as signs of diversifying selection on the gene paralogues suggesting that the variable
94 environment in epipelagic waters have facilitated a larger more diverse *Hb* gene
95 repertoire (Baalsrud *et al.* 2017). However, the rather fragmented draft genomes did
96 not allow for reconstruction of the gene synteny and a deeper understanding of the
97 evolution of the Hb-clusters.
98 Moreover, two tightly linked polymorphisms at amino acid positions 55 and
99 62 of the Hbb1-globin suggested associated with thermal adaptation are
100 demonstrated in Atlantic cod populations. These polymorphisms exhibit a latitudinal
101 cline in allele frequency in populations inhabiting varying temperature and oxygen
102 regimes for Atlantic cod in the North Atlantic and Baltic Sea (Andersen *et al.* 2009).
103 Populations found in the southern regions display the Hbb1-1 variant (Met55Lys62),
104 whereas more northern populations largely display the Hbb1-2 variant (Val55Ala62)
105 (Andersen *et al.* 2009). The Hbb1-1 variant has been shown to be insensitive to
106 temperature whereas Hbb1-2 is temperature dependent with a higher O_2 affinity
107 than Hbb1-1 at colder temperatures (Andersen *et al.* 2009), however, this has been
108 questioned by Barlow *et al.* 2017. Additionally, an indel polymorphism within the
109 promoter of the *Hbb1* gene has been reported to be in linkage disequilibrium with the
110 above-mentioned polymorphisms (Star *et al.* 2011). Examination of multiple Atlantic
111 cod populations uncovered that a longer promoter variant is associated with *Hbb1-2*
112 and found to up-regulate its gene expression at higher temperatures, i.e. aiding in the
113 maintenance of total oxygen-carrying capacity (Star *et al.* 2011).

114 In teleosts, the *Hb* genes are found to reside at two distinct genomic regions,
115 the MN and LA cluster. Earlier reports have shown that there is a high evolutionary
116 turnover of *Hb* genes across teleosts, with lineage-specific duplications and losses,
117 which is in stark contrast to genes flanking the *Hb* genes, where the synteny is highly
118 conserved (Quinn *et al.* 2010; Opazo *et al.* 2012; Feng *et al.* 2014). In this study, the
119 overall goal was to elucidate the evolutionary past of the *Hb* clusters – including *Hb*
120 genes, flanking genes and intergenic sequences – within the phylogenetically diverse
121 group of codfishes (Gadiformes) by taking advantage of long read sequencing
122 technology combined with targeted sequence capture. Eight codfish species were
123 carefully selected on the basis of both phylogenetic and habitat divergence, implying
124 that they are exposed to a variety of environmental factors as well as displaying
125 distinct life-history traits. A highly continuous genome assembly of Atlantic cod
126 (Tørresen *et al.* 2017) as well as low coverage draft genome assemblies of all eight
127 species (Malmstrøm *et al.* 2017) were used in the design of the probes covering the
128 genomic regions of interest. To enable targeted sequence capture for PacBio RSII
129 sequencing, we modified the standard protocol for sequence capture offered by
130 NimbleGen, i.e. the SeqCap EZ (Roche NimbleGen), as well as generating custom-
131 made barcodes. This combined approach resulted in successful capturing and
132 assembling of the two *Hb* gene clusters across the codfishes examined. The
133 generation of highly continuous assemblies – for most of the species – enabled
134 reconstruction of micro-synteny revealing lineage-specific gene duplications and
135 identification of a relatively large and inter-species variable indel located in the
136 promoter region between the *Hbb1* and *Hba1* genes.
137 Our study demonstrates that combining sequence capture technology with
138 long-read sequencing is a highly efficient and versatile method to investigate specific
139 genomic regions of interest – with respect to micro-synteny, regulatory regions and
140 genetic organization – across distantly related species where genome sequences are
141 lacking.

Results

Capture and *de novo* assembly of the target regions

The probe design (workflow schematically shown in Figure 1) resulted in a total of 7057 capture probes based on the target region in Atlantic cod, covering 337 kbp of sequence. 26774 probes were designed for the additional codfishes, covering a total of 1.82 Mbp of target sequence. The target region and the *Hb* gene clusters were successfully captured and enriched for eight codfishes; Atlantic cod (*Gadus morhua*), haddock (*Melanogrammus aeglefinus)*, silvery pout (*Gadiculus argenteus*), cusk (*Brosme brosme*), burbot (*Lota lota*), European hake (*Merluccius merluccius*), marbled moray cod (*Muraenolepus marmoratus*), and roughhead grenadier (*Macrourus berglax*), with number of reads spanning from 35573 to 73005 (Table 1). The average read length was 3032 bp, varying from 2836 bp in European hake to 3265 bp in burbot, resulting in the capture of an average of 16.71 Mbp per species (Table 1). By mapping reads back to the capture target region we found that the average mapping depth was variable across the target region for all species (Figure 2 and 3). Because of the skewed distribution of mapping depth, we also calculated median depth, which was, as expected, the highest for Atlantic cod at 242x (Table S1). The median mapping depth was consistently high for most of the other species as well, with the lowest for roughhead grenadier (12x). Both median and average depths for the MN region were persistently higher than for the LA region for all species, with the exception of silvery pout (Table S1). Furthermore, positions with high degree of mapping corresponded to the location of the genes used in the design of the capture probes across all species (Figure 2 and Figure 3). The percentage of reads mapping to the target region ranged from 25-43%, however, the percentage of the target region covered by reads ranged from 53-100% with five species having more than 90% of the target region covered by reads (Figure 4c and Table S1).

To address factors influencing capture success we compared various capture statistics to overall genomic divergence between the Atlantic cod genome and independent WGS data for each species from (Malmstrøm *et al.* 2017) (Table S1). We found a strong negative correlation between genomic divergence to Atlantic cod and median mapping depth against the target region (r=-0.90, Figure 4a), percent of reads mapped to the target region (r=-0.90, Figure 4b), and percentage of reads mapped to the target region (r=-0.84, Figure 4c).

We constructed *de novo* assemblies with quite consistent assembly statistics across species. Contig N50 ranged from 8055 bp in burbot to 6523 bp in European hake and the total number of contigs varied from 205 in burbot to 455 in marbled

178 moray cod. However, there was some variation in the size of the largest contig,
179 which ranged from 79 kbp in Atlantic cod to 30 kbp in marbled moray cod (Table 1).
180 To evaluate whether the assemblies represent the actual target regions we mapped
181 the *de novo* assemblies for each species to the target region in Atlantic cod, for which
182 the capture design is largely based upon (Figure 2 and 3). As expected, the
183 assemblies corresponded to the regions with high coverage of reads, i.e. the areas of
184 the target region containing genes included in the probe design.

185 **Synteny of the *Hb* gene regions**

186 Our capture design combined with long-read PacBio sequencing allowed us to
187 reconstruct micro-synteny of the MN and LA regions for Atlantic cod, haddock,
188 silvery pout, cusk, burbot, European hake, marbled moray cod and roughhead
189 grenadier (Figure 5). From the *de novo* assemblies, we were able to identify the
190 majority of the *Hb* genes and all of the flanking genes, which show that our capture
191 design was successful. However, the degree of continuity varied in the different
192 assemblies. In Atlantic cod, haddock, silvery pout, cusk, burbot and European hake
193 we could infer micro-synteny revealing that *Hb* and their flanking genes organization
194 largely followed what has previously been reported for Atlantic cod (Figure 5) (Star
195 *et al.* 2011). We found *Hbb4* only to be present in Atlantic cod (Figure 5b), which is in
196 line with Baalsrud *et al.* 2017. Furthermore, the *de novo* assemblies confirmed a linage-
197 specific duplication of *Hbb2* in the roughhead grenadier (Baalsrud *et al.* 2017).
198 Additionally, we identified a complete *Hba4*-like gene in the assembly of the marbled
199 moray cod, not earlier identified in this species. However, the *Hba4*-like gene in
200 marbled moray cod is likely a pseudogene due to a frameshift mutation causing
201 multiple stop codons. Furthermore, we were able to identify most of the *Hb* genes
202 reported in the recent study by Baalsrud *et al.* 2017, however, a few are missing from
203 our dataset (Figure 5a and b). Pairwise sequence alignment of these paralogous *Hb*
204 genes from Baalsrud *et al.* 2017 revealed sequence identities up to 98 % (Table S2).

205 **Target region in the haddock and Atlantic cod genome assemblies**

206 As a proof of concept, we reconstructed synteny of the target region in the most
207 recent genome assemblies of Atlantic cod (Tørresen *et al.* 2017; gadMor2) and
208 haddock (Tørresen *et al.* 2018; melAeg). In Atlantic cod, the MN region is located on
209 linkage group 2 (Figure 5a) and LA on linkage group 18 (Figure 5b), in haddock MN
210 is located on scaffold MeA_20160214_scaffold_771 (Figure 5a) and LA on scaffold
211 MeA_20160214_scaffold_1676 (Figure 5b). The overall synteny in Atlantic cod was
212 congruent with Wetten *et al.* 2010 except for the relative direction of the genes *foxj1a*
213 and *rhbdf1.* Furthermore, the organization of *Hb*s and their flanking genes in the

214 genome assembly of haddock is conserved compared to Atlantic cod with the
215 exception of *Hbb4* in the MN region, which is absent in haddock (Figure 5).

216 Repetitive sequences in the in the *Hb* gene regions

217 Quantifying the amount of repetitive sequences in the target region(s) was only
218 possible for Atlantic cod (gadMor2) and haddock (melAeg), for which high-quality
219 genome assemblies exist. The amount of repetitive sequences in the target region
220 differed between the MN cluster and the LA cluster in Atlantic cod. The MN region
221 (214 kb) contained a total of 10.7% repeated sequences, including 1.0% retro-
222 elements, 1.3% transposons, 5.8% simple repeats, and 2.6% of various low complexity
223 and unclassified repeated sequences (Table S3). In comparison, in the smaller LA
224 region (123 kb) the proportion of repeated sequences was twice as high (20.3%),
225 which comprised of 2.8% retro-elements, 2.4% transposons, 13.8% simple repeats,
226 and 1.3% of various low complexity and unclassified repeated sequences.
227 Furthermore, the orthologous target regions in haddock followed the same pattern.
228 The MN region contained 16.3 % repeated sequences, in contrast to 19.8 % found in
229 the LA region (Table S3).

230 Insertions and deletions in the promoter region of *Hba1 – Hbb1*

231 The previously shown 73 bp indel in the bi-directional promoter region of *Hba1* and
232 *Hbb1* – discerning the cold-adapted migratory Northeast Artic (NEA) cod from the
233 more temperate-adapted southern Norwegian coastal (NC) cod (Star *et al.* 2011) –
234 was confirmed by the improved version of the NEA cod assembly (gadMor2). The
235 continuity of our capture assemblies (Figure 5) enabled location of the orthologous
236 captured regions in haddock, silvery pout and cusk. In each of the species an indel of
237 variable length were identified (Figure 6). Compared to the long promoter variant –
238 found to be linked with the *Hbb1-2* in Atlantic cod – the indel is shorter in the other
239 species by 11 bp in haddock, 22 bp in silvery pout and 56 bp in cusk (Figure 6).
240 Although the indels are varying in length, the conserved flanking sequences in the
241 alignment clearly show that they represent orthologous regions. Moreover, we found
242 the amino acid positions at 55 and 62 of the *Hbb1* gene to vary between species;
243 Haddock has Val55-Lys62, silvery pout has Met55-Gln62, while cusk has Met55-
244 Lys62 similarly to NEA cod (Figure 6). Additionally, we investigated amino acid
245 positions 55 and 62 in the *Hbb1* gene across a number additional codfish species for
246 which we have available gene sequences from Baalsrud *et al.* 2017, revealing these
247 sites to be variable across this lineage (Table S4). Ancestral reconstruction of *Hbb1*
248 demonstrated that the ancestral state in position 55 was Met in codfishes, and in

249 position 62 was Lys in all codfishes except *Bregmaceros cantori* (Supplementary
250 Figures S1 and S2).

Discussion

Capture of *Hb* gene clusters with 70 million years divergence time reveal conserved synteny and lineage-specific *Hb* duplications

We here demonstrate a successful in-solution targeted sequence capture and assembling of coding and noncoding sequences of the *Hb* clusters from codfish species separated by up to 70 million years (My) of evolution. Two features make our approach unique from earlier studies. First, the target regions consisted of both coding and noncoding genomic sequences. Second, we designed custom-made probes in order to utilize the long-read PacBio sequencing platform. In contrast to previous targeted capture sequencing studies based on short-read sequencing technologies (George *et al.* 2011; Mascher *et al.* 2013) our approach enabled the generation of highly continuous assemblies of the *Hb* clusters across distantly related codfishes.

The organization and orientation of the *Hb* flanking genes that we identified were conserved across all species (Figure 5a and b). However, in concordance with earlier studies of the *Hb* region, we found significant variation in copy numbers of the *Hb* genes, with linage specific duplications and losses (Star *et al.* 2011; Opazo *et al.* 2012; Feng *et al.* 2014; Baalsrud *et al.* 2017). We only found *Hbb4* in Atlantic cod, supporting earlier studies showing that *Hbb4* is the result of a recent duplication in this species (Borza *et al.* 2009; Baalsrud *et al.* 2017). Interestingly, the presence of two copies of *Hbb2* on the same contig in the roughhead grenadier *de novo* assembly confirmed a lineage-specific gene duplication of *Hbb2* in this species. This duplication was postulated in recent study of *Hbs* in codfishes (Baalsrud *et al.* 2017), but due to the lack of synteny in the draft genome assemblies, it was not possible to determine with certainty. Additionally, a copy of the *Hba4* was found in the *de novo* assembly of the marbled moray cod not found in the previous study by Baalsrud *et al.* (2017). The presence of a frame-shifting mutation that is causing multiple stop codons indicated that this *Hba4* gene is most likely a pseudogene. *Hba4* is also a pseudogene in the closely related species *Mora moro*, *Trachyrincus scabrus*, *T. murrayi* and *Melanonus zugmayeri* (Baalsrud *et al.* 2017). Although we identified most of the *Hb* genes from Baalsrud *et al.* (2017), a few were absent from this dataset (Figure 5a and b), which we suspect may be due to collapse of paralogous *Hb* genes, as they may have as high as 98% sequence identity (Table S2).

Length variation in the bi-directional *Hba1-Hbb1* promoter within the codfishes

The discovery of a promoter of variable length between *Hba1* and *Hbb1* in different species (Figure 6) was concordant with earlier findings of length variation in the homologous region in different populations of Atlantic cod (Star *et al.* 2011). The migratory NEA cod population has been shown to harbor the 73 bp longer variant at a higher frequency compared to coastal cod populations (see Figure 6 and Star *et al.* 2011). Interestingly, we found relatively long promoters with high sequence similarity to the NEA cod indel in haddock and silvery pout. In contrast, cusk displayed a relatively short promoter, however, still 17 bp longer than in NC cod (Figure 6). Furthermore, we found the amino acid positions 55 and 62 in *Hbb1,* known to be polymorphic in Atlantic cod, to be variable across all codfishes included in this study (Figure 6). Investigations of the same positions in a number additional codfishes for which we have available gene sequences (Baalsrud *et al.* 2017), revealed that these positions are highly variable across this linage (Table S4). Notably, the most likely ancestral state of codfish *Hbb1* is Met55Lys62 (Supplementary Figures S1 and S2). Cusk and the coastal/southern Atlantic cod thus both display the ancestral state as well as a short promoter, although the cusk promoter was 17 bp longer (Figure 6). Collectively, these results suggest two different scenarios for promoter length evolution. Scenario 1: The short promoter represents the ancestral state of the Gadidae-family (including cusk and Atlantic cod; see Malmstrøm *et al.* 2016) and that silvery pout and some populations of Atlantic cod have evolved a longer promoter. Scenario 2: The long promoter is the ancestral state with independent deletions of variable lengths in cusk, silvery pout, haddock and coastal/southern Atlantic cod (*Hbb1-1*). To disentangle this, we would need to obtain promoter sequences from additional gadiform species. Regardless, the short-long promoter polymorphism has been maintained throughout speciation events based on the presence of both variants in Atlantic cod. Moreover, in both scenarios, cusk and coastal/southern Atlantic cod (*Hbb1-1*) have maintained the ancestral Met55Lys62, while silvery pout, haddock and NEA cod (*Hbb1-2*) have acquired substitutions at these positions due to similar selection pressures or genetic drift. In this regard, it could be mentioned that the NEA cod, haddock and silvery pout display migratory behavior (e.g. diurnally feeding movements as well as seasonal spawning migrations) compared to the more stationary cusk and coastal cod (Eschemeyer & Fricke 2017) which could mean that they have a higher O_2 demand and are exposed to greater temperature variation, which in turn has selected for a temperature-dependent long promoter. Furthermore, given that promoter length and positions 55/62 at *Hbb1* are important genetic

321 components of temperature adaptation in Atlantic cod populations (Star *et al.* 2011),
322 they most likely play a role in temperature adaptation in the other codfishes.

323 Assembly success affected by probe design and repeat content

324 In some species, nearly the complete target region is assembled in large contigs
325 containing multiple genes including cusk, whereas in other species such as the more
326 distantly related roughhead grenadier, the cluster is more fragmented (Figure 5). In
327 all species, the areas of the target regions that harbor genes of which probes are
328 designed for, as well as any areas containing repeated sequences, have very high
329 depths in comparison to the areas of intergenic sequences (Figure 2 and 3). This
330 poses a challenge for the assembly software, which is based in the assumption of
331 uniform depth over the sequencing data (Miller *et al.* 2010).
332 Overall, the MN cluster seems to be more successfully assembled than the LA
333 cluster, which is more fragmented (Figure 5). Differences in assembly completeness
334 between the two regions might be a result of several factors. Firstly, the MN region
335 has more flanking genes in closer proximity to the *Hb* region, which results in a
336 higher density of probes. Secondly, the overall repeat content of the LA region is one
337 order of magnitude larger than in the MN region, largely due to the larger
338 proportion of simple repeats. Repeat content is a major interference in capture
339 experiments because unwanted repetitive DNA may be enriched for, especially if
340 there are repeated sequences included in the probe design. Furthermore, if probes
341 were not completely covered by target DNA they get single-stranded sticky ends that
342 can hybridize to repetitive or other non-target DNA (Newman & Austin 2016).
343 Lastly, unless there were some longer reads that bridged such areas, this would in
344 turn have led to gaps in the downstream *de novo* assemblies. Following that the
345 assembly success was possibly a result of read length, we reason that a future
346 increase of the average read length from 3 kbp to 5-10 kbp, would be sufficient to
347 substantially increase the completeness of the assemblies. Due to the current circular
348 consensus (CCS) PacBio sequencing technology, however, which is a trade-off
349 between accuracy and length of reads, longer reads with sufficient accuracy are not
350 feasible.

351 Long-read sequencing capture across species harbors new potential for 352 comparative genomic studies

353 The number of reads mapping to the target region was in the range of 23-43%, which
354 may be seemingly low compared with other capture studies. For instance, a whole
355 exome capture study on humans reported 56.1% of reads mapped to the target
356 region (Guo *et al.* 2012) and a similar study in rats reported to have 78.3% of reads on

357 target (Yoshihara *et al.* 2016). In contrast to our study however, these capture
358 experiments enriched either the exome or ultra-conserved elements within a single
359 species, allowing for more efficient capture of conserved sequences. We were
360 however, able to cover up to 98% of the target region with a sequencing depth of >10
361 reads across species (Table S1) which is similar to what mentioned experiments
362 within human and rat exomes reported (Guo *et al.* 2012; Yoshihara *et al.* 2016) and the
363 main difference is the higher percentage of non-target sequences in our study.
364 We were able to capture complete genes for species with 70 My divergence
365 time from the Atlantic cod (Figure 5). As expected, we found that capture success
366 declines with increased sequence divergence between the reference genome of which
367 we chiefly based our capture probes and the genomes of the included codfishes
368 (Figure 4). It has been reported that orthologous exons were successfully captured in
369 highly divergent frog species (with 200 My of separation), nevertheless the capture
370 success greatly decreased with increased evolutionary distance (Hedtke *et al.* 2013).
371 Similarly, it has been demonstrated that it is possible to capture >97% of orthologous
372 sequences in four species of primates that diverged from humans 40 My ago, using
373 probes entirely based on the human exome (George *et al.* 2011). Further, exomes were
374 effectively captured from skink species that diverged up to 80 My from the reference,
375 yet reporting a substantial decline in capture efficiency for sequences >10 % different
376 from the reference species (Bragg *et al.* 2015). Our study stands out from previous
377 capture experiments because intergenic, noncoding sequences in addition to genes
378 were captured across distantly related species. Efficient capture of intergenic
379 sequences requires less divergence time, as these regions usually evolve faster than
380 genes (Koonin & Wolf 2010). Thus, the most distantly related species from Atlantic
381 cod for which we captured both coding and noncoding sequences was burbot, which
382 diverged from Atlantic cod 46 My (Figure 5). We argue, in line with (Schott *et al.*
383 2017), that sequence divergence may be a more exact predictor of capture success
384 than evolutionary distance, as the sequence capture process is mainly influenced by
385 the difference between the probe sequence and the target sequence. European hake,
386 marbled moray cod and roughhead grenadier all diverged from cod about 70 My
387 ago, however, the European hake *Hb* regions was more successfully captured and
388 assembled (Table 1; Figure 2). This could be due to European hake having a lower
389 genome-wide divergence to Atlantic cod than marbled moray cod and roughhead
390 grenadier (809k vs 879k and 907k SNPs; Table S1).
391 Finally, it should be mentioned that cusk – which diverged from Atlantic cod
392 39 My ago – was added to the experimental design after the probes were generated.

393 Thus, the successful capture of cusk was therefore solely based on cross-species
394 target enrichment, demonstrating the power of heterologous probe targeting.

395 Concluding remarks and future perspectives

396 Here, we have successfully demonstrated that combining targeted sequence capture
397 with long-read sequencing technology is as an efficient approach to obtain high
398 quality sequence data of a specific genomic region, including both coding and
399 noncoding sequences, across evolutionary distant species. We show that genome-
400 wide divergence is of importance for capture success across species. Furthermore, the
401 use of long-read sequencing augmented the *de novo* assembly of regions containing
402 repeated sequences that would otherwise fragment assemblies based on short-read
403 sequencing. This is crucial for capturing complete intergenic sequences that may be
404 highly divergent compared to genic regions even among fairly closely related
405 species. Given the rapid development in sequencing technologies future methods
406 will enable read-through of repeated regions and thus further increase the
407 completeness of assemblies. Moreover, a less stringent hybridization protocol should
408 make it possible to capture sequences across even deeper evolutionary time. In sum,
409 our approach has generated a cross-species genomic resource across distantly related
410 codfishes and show the potential of enhancing comparative genomic studies of
411 continuous genic and intergenic regions between any eukaryotic species-group
412 where genomic resources are scarce.

413 # Material and methods

414 ## Defining target region and probe design

415 The probe design was chiefly based on the high-quality genome of Atlantic cod,
416 known as gadMor2 (Tørresen *et al.* 2017). In addition, species-specific probes were
417 designed based on low-coverage assembled genomes (Malmstrøm *et al.* 2016) for ten
418 selected species representing six families in the Gadiformes order. These species
419 were Atlantic cod (*Gadus morhua*), Alaskan Pollock (*Gadus chalcogrammus*), polar cod
420 (*Boreogadus saida*), haddock (*Melanogrammus aeglefinus*), Silvery pout (*Gadiculus*
421 *argenteus*), burbot (*Lota lota*), European hake (*Merluccius merluccius*), roughhead
422 grenadier (*Macrourus berglax*), roughsnout grenadier (*Trachyrincus scabrus*) and
423 marbled moray cod (*Muraenolepus marmoratus*).
424 To retrieve relevant sequence data for the probe design, the MN and LA *Hb*
425 regions were extracted from gadMor2 (Figure 1). These sequences, hereby known as
426 the target region, were then used as queries in BLAST (Altschul *et al.* 1990) searches
427 with an E-value threshold of <0.1 against the genome assembly data of all ten species.
428 In total, 5604 sequences from the chosen species were supplied to NimbleGen
429 probe design. Protein coding genes from the ENSEMBL database were used to define
430 the regions to be tiled in the probe design (Table S5) within the target region of the
431 Atlantic cod, and the unitigs for each of the additional codfishes.
432 NimbleGen SeqCap EZ capture probes were designed by NimbleGen (Roche,
433 Madison, USA) using a proprietary design algorithm. NimbleGen offers an in-
434 solution sequence capture protocol, which includes custom made probes. Uniquely,
435 the capture probes from NimbleGen are tiled to overlap the target area. 50 – 100 bp
436 (average 75 bp) probes where designed tiled over the target region (subset of
437 gadMor2) resulting in each base, on average, being covered by two probes.
438 Additionally, raw reads from Illumina sequencing from (Malmstrøm *et al.* 2016; 2017)
439 were used for each species to estimate repetitive sequences in each of the species
440 genomes, aiming to discard probes containing any repeats. A more detailed
441 description of the probe design is provided in Supplementary Materials and
442 Methods and Table S6.

443 ## Sample collection and DNA extraction

444 Our goal working with animals is always to limit any harmful effects of our research
445 on populations and individuals. Whenever possible we try to avoid animals being
446 euthanized to serve our scientific purpose alone by collaborating with commercial
447 fisheries or museums. The tissue samples used in this study are either from

448 commercially fished individuals intended for human consumption or museum
449 specimen. The commercially caught fish were immediately stunned, by bleeding
450 following standard procedures by a local fisherman. There is no specific legislation
451 applicable to this manner of sampling in Norway, however it is in accordance with
452 the guidelines set by the 'Norwegian consensus platform for replacement, reduction
453 and refinement of animal experiments' (www.norecopa.no).
454 DNA was extracted from tissue samples using High Salt DNA Extraction
455 method by Phill Watts (https://www.liverpool.ac.uk/~kempsj/IsolationofDNA.pdf ,
456 last day accessed: December 2017). The concentration and purity of the DNA samples
457 were quantified using NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA) and
458 a Qubit fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA). Due
459 to poor DNA quality, three species included in the probe design; Alaskan Pollock,
460 polar cod and roughsnout grenadier were excluded from further analysis. In total,
461 eight species were sequenced; seven of these species were included in the probe
462 design and one closely related species (cusk), which serves as a cross species capture
463 experiment without species-specific probes.

464 **Capture, library preparation and sequencing**

465 The sequencing libraries were prepared following a modified Pacific Biosciences
466 SeqCap EZ protocol. As multiplexing of the samples before capture was required,
467 barcodes were designed at the Norwegian Sequencing Centre
468 (http://www.sequencing.uio.no) using guidelines from Pacific Biosciences, for more
469 information see Supplementary Materials and methods, Table S7 and Figure S3.
470 Genomic DNA was sheared to 5 kb fragments using MegaRuptor (Diagenode,
471 Seraing (Ougrée), Belgium). Due to poorer DNA quality, fragmenting was not done
472 for European hake. For this sample together with fragmented DNA from roughhead
473 grenadier, short fragments were removed using BluePippin (Sage Science, Beverly,
474 MA, USA) before library preparation. Illumina libraries were prepared using KAPA
475 Hyper Prep kit (Kapa Biosystems, Wilmington, MA, USA) and barcoded using
476 different Illumina barcodes. PacBio barcodes were implemented during pre-capture
477 amplification of libraries. After amplification, fragment length distribution was
478 evaluated using Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and
479 samples were pooled in equimolar ratio. During hybridization, SeqCap EZ
480 Developer Reagent (universal repeat blocker for use on vertebrate genomes) and
481 oligos corresponding to Illumina and PacBio barcodes were used for blocking.
482 Captured gDNA was amplified to ensure that sufficient amount of DNA was
483 available for PacBio library preparation. Size selection of the libraries was performed

484 using BluePippin. Final libraries were quality checked using Bioanalyzer and Qubit
485 fluorometer (Invitrogen, Thermo Fisher Scientific, Waltham, MA, USA) and
486 sequenced on RS II instrument (PacBio, Menlo Park, CA, USA) using P6-C4
487 chemistry with 360 minutes movie time. In total, 9 SMRT cells were used for
488 sequencing.

489 *De novo* assemblies

490 Reads were filtered and de-multiplexed using the 'RS_reads of insert.1' pipeline on
491 SMRT Portal (SMRT Analysis version smrtanalysis_2.3.0.140936.p2.144836). Each set
492 of reads corresponding to a given species was crossed-checked with their respective
493 six-nucleotide Illumina adapter. Reads containing an incorrect Illumina adapter were
494 removed. Adapter sequences were then trimmed using the application Prinseq-lite
495 v0.20.4 (Schmieder & Edwards 2011). The trimmed reads were assembled *de novo*
496 using Canu v1.4 +155 changes (r8150 c0a988b6a106c27c6f993dfe586d2336282336a6)
497 (Berlin *et al.* 2015). The Canu software is optimized for assembling single molecule
498 high noise sequence data. We specified genome size as the size of the target region
499 (300 kbp). Additionally, we ran PBJelly (English *et al.* 2012) on the Canu *de novo*
500 assemblies, using the raw reads to possible bridge gaps between scaffolds, settings
501 given in Supplementary Materials and Methods.
502 We assessed the assemblies by running Assemblathon 2 (Bradnam *et al.* 2013),
503 which reports assembly metrics such as the longest contig, the number of contigs,
504 and the N50 value. *De novo* assemblies of the MN and LA regions of Atlantic cod and
505 haddock were aligned and compared to their reference genomes, gadMor2 and
506 melAeg respectively, using BLAST and BWA v0.7.10 (Li & Durbin 2009) to determine
507 syntenic similarities and assembly completeness.

508 Estimating capture success

509 PacBio reads for all the species were mapped back to the Atlantic cod genome
510 assembly (gadMor2) in order to determine sequence capture success and target
511 mapping depths. Mapping was done using BWA-MEM v0.7.10 (Li & Durbin 2009).
512 Target-area read depth for all the species based on mapping against gadMor2, were
513 calculated using Samtools v1.3.1 (Li *et al.* 2009). We calculated both average and
514 median mapping depth against the target region as a whole and for the MN and LA
515 region separately. We also calculated percentage of reads that mapped to the target
516 region, and the percentage of the target regions covered by reads to a minimum
517 depth of 10x. To compare assemblies to the target region we additionally mapped the
518 assemblies to the target region. In order to verify the sequence capture process,
519 sequence data for Atlantic cod and haddock were mapped back to their reference

520 genomes using BWA-mem v0.7.10 (Li & Durbin 2009). The results were visualized
521 using Integrative Genome Viewer (Robinson *et al.* 2011).
522 To obtain an independent measure of divergence between species in the
523 capture experiment we calculated genome wide level of divergence of each species to
524 the reference genome of Atlantic cod using low-coverage whole-genome sequence
525 data from (Malmstrøm *et al.* 2016; 2017). We mapped raw reads to Atlantic cod using
526 BWA-MEM (Li & Durbin 2009) and called SNPs using the Freebayes variant caller
527 (Garrison & Marth 2012). Some species are more closely related to Atlantic cod than
528 others, which could introduce a bias in mapping. To avoid this, we only looked at
529 genomic regions where all species mapped. The number of SNPs was then used as an
530 estimate of genome-wide divergence of each species to Atlantic cod. We also mapped
531 a low-coverage genome of Atlantic cod to the Atlantic cod reference genome as a
532 control.
533 In pursuance of factors explaining capture success we tested for correlations
534 and plotted the relationship between the genome wide level of divergence and the
535 following variables; median mapping depth against the target region (for total, LA
536 and MN, respectively); percentage of reads that mapped to the target region; and the
537 percentage of the target region covered by reads. All tests and plots were done using
538 R version 3.2.5 (R Core Team 2016).
539 Assembly continuity is very often hampered by the presence of repeats, which
540 create gaps. We therefore quantified repeat-content in the target region extracted
541 from gadMor2 and orthologous regions in haddock using Repeatmasker Open 3.0
542 (Smit A *et al.* 2010) for the MN region and the LA region separately.

543 Identifying gene location and synteny

544 In order to identify the genes of interest and their location in the assembly we used
545 local sequence alignment algorithm BLAST v2.4.0 (Altschul *et al.* 1990) with protein
546 sequences of the genes of interest (Table S5) as queries. tblastn was used with an e-
547 value of 0.1. Investigation of *Hbb1-Hba1* promoter region was done for four species,
548 Atlantic cod, haddock, silvery cod and cusk. Sequences were aligned with ClustalW
549 default settings using MEGA7 (Kumar *et al.* 2016). Ancestral sequence reconstruction
550 was carried out for *Hbb-1* gene sequences from 24 species of codfishes from (Baalsrud
551 *et al.* 2017) using a maximum likelihood method implemented in MEGA7 (Kumar *et*
552 *al.* 2016).
553 Additionally, we estimated sequence identity using EMBOSS Needle (Rice *et*
554 *al.* 2000) with default settings, between *Hbb* gene sequences from (Baalsrud *et al.*

2017) that were missing and present in the *de novo* assemblies to evaluate similarity (Table S2).

Acknowledgements

We would like to thank Marianne H. Selander Hansen and Alexander J. Nederbragt for help with the initial design of this project. All computational work was performed on the Abel Supercomputing Cluster (Norwegian Metacenter for High-Performance Computing (NOTUR) and the University of Oslo), operated by the Research Computing Services group at USIT, the University of Oslo IT Department. Sequencing library creation and high-throughput sequencing were carried out at the Norwegian Sequencing Centre (NSC), University of Oslo, Norway. This work was funded by a grant awarded to K.S.J. from the Research Council of Norway (RCN grant 222378).

Author contributions

H.T.B. and S.J. initially conceived and designed the study, with input from S.N.K.H, A.T.-K., M.S., G.O., R.S., and K.S.J. Tissue samples were provided by S.J. and H.T.B. Probe design was carried out by T.R. with assistance from S.N.K.H and H.T.B. DNA extraction and sequence library preparation was performed by S.N.K.H and A.T.-K, respectively. Sequence capture was carried out by S.N.K.H, A.T.-K., M.S. and G.O. Filtering, mapping of sequences and *de novo* assemblies was done by S.N.K.H., assisted by O.K.T and H.T.B. Annotation of genes, synteny analyses, statistical analyses and construction of all figures and tables was done by S.N.K.H and H.T.B. The manuscript was written by S.N.K.H and H.T.B. with input from S.J. and K.S.J.

Competing interests

The authors declare that they have no competing interests.

Data and materials availability

All reads and assemblies (unitigs) reported on here, and the target region, subset of gadMor2, as well as the relevant sequence data for the probe design from the chosen species supplied to NimbleGen and the probe sequences have been deposited at figshare under doi/xxx.

585 **References**

586 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment
587 search tool. *Journal of Molecular Biology*, **215**, 403–410.
588 Andersen O, Wetten OF, De Rosa MC *et al.* (2009) Haemoglobin polymorphisms
589 affect the oxygen-binding properties in Atlantic cod populations. *Proceedings of*
590 *the Royal Society B: Biological Sciences*, **276**, 833–841.
591 Baalsrud HT, Voje KL, Tørresen OK *et al.* (2017) Evolution of Hemoglobin Genes in
592 Codfishes Influenced by Ocean Depth. *Scientific Reports*, **7**, 1–10.
593 Barlow SL, Metcalfe J, Righton DA, Berenbrink M (2017) Life on the edge: O2 binding
594 in Atlantic cod red blood cells near their southern distribution limit is not
595 sensitive to temperature or haemoglobin genotype. *The Journal of Experimental*
596 *Biology*, **220**, 414–424.
597 Berlin K, Koren S, Chin C-S *et al.* (2015) Assembling large genomes with single-
598 molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, **33**, 623–
599 630.
600 Borza T, Stone C, Gamperl AK, Bowman S (2009) Atlantic cod (*Gadus morhua*)
601 hemoglobin genes: multiplicity and polymorphism. *BMC Genetics*, **10**, 51–14.
602 Bradnam KR, Fass JN, Alexandrov A *et al.* (2013) Assemblathon 2: evaluating de
603 novo methods of genome assembly in three vertebrate species. *Giga Science,* **2**, 10.
604 Bragg JG, Potter S, Bi K, Moritz C (2015) Exon capture phylogenomics: efficacy across
605 scales of divergence. *Molecular Ecology Resources*, **16**, 1059–1068.
606 Broeckx BJG, Coopman F, Verhoeven GEC *et al.* (2014) Development and
607 performance of a targeted whole exome sequencing enrichment kit for the dog
608 (Canis Familiaris Build 3.1). *Scientific Reports*, **4**, 1522–4.
609 Ellegren H (2014) Genome sequencing and population genomics in non-model
610 organisms. *Trends in Ecology & Evolution*, **29**, 51–63.
611 English AC, Richards S, Han Y *et al.* (2012a) Mind the Gap: Upgrading Genomes with
612 Pacific Biosciences RS Long-Read Sequencing Technology (Z Liu, Ed,). *PLoS*
613 *ONE*, **7**, e47768–12.
614 Eschemeyer WN, Fricke R. Catalog of fishes.
615 http://research.calacademy.org/researchichthyology/catalog/fishcatmain.asp.2017
616 . (Last accessed December 2017)
617 Feng J, Liu S, Wang X *et al.* (2014) Channel catfish hemoglobin genes: Identification,
618 phylogenetic and syntenic analysis, and specific induction in response to heat
619 stress. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, **9**,
620 11–22.

621 Garrison E, Marth G (2012) Haplotype-based variant detection from short-read
622 sequencing. *bioRxiv*, 1207.3907.
623 George RD, McVicker G, Diederich R *et al.* (2011) Trans genomic capture and
624 sequencing of primate exomes reveals new targets of positive selection. *Genome*
625 *Research*, **21**, 1686–1694.
626 Giolai M, Paajanen P, Verweij W *et al.* (2017) Comparative analysis of targeted long
627 read sequencing approaches for characterization of a plant's immune receptor
628 repertoire. *BMC Genomics*, **18**, 521.
629 Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-
630 generation sequencing technologies. *Nature Reviews Genetics*, **17**, 333–351.
631 Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful
632 tool for evolutionary analysis1. *American Journal of Botany*, **99**, 312–319.
633 Guo Y, Long J, He J *et al.* (2012) Exome sequencing generates high quality data in
634 non-target regions. *BMC Genomics*, **13**, 194.
635 Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM (2013) Targeted Enrichment:
636 Maximizing Orthologous Gene Comparisons across Deep Evolutionary Time (U
637 Joger, Ed,). *PLoS ONE*, **8**, e67908–10.
638 Huddleston J, Ranade S, Malig M *et al.* (2014) Reconstructing complex regions of
639 genomes using long-read sequencing technology. *Genome Research*, **24**, 688–696.
640 Jones MR, Good JM (2015) Targeted capture in evolutionary and ecological
641 genomics. *Molecular Ecology*, **25**, 185–202.
642 Kim KE, Peluso P, Babayan P *et al.* (2014) Long-read, whole-genome shotgun
643 sequence data for five model organisms. *Scientific Data*, **1**, 140045–10.
644 Kolmogorov M, Yuan J, Lin Y, Pevzner P (2018) Assembly of Long Error-Prone
645 Reads Using Repeat Graphs. *bioRxiv*, 247148.
646 Koonin EV, Wolf YI (2010) Constraints and plasticity in genome and molecular-
647 phenome evolution. *Nature Reviews Genetics* , **11**, 487–498.
648 Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics
649 Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, **33**,
650 1870–1874.
651 Li C, Corrigan S, Yang L *et al.* (2015) DNA capture reveals transoceanic gene flow in
652 endangered river sharks. *Proceedings of the National Academy of Sciences*, **112**,
653 13302–13307.
654 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler
655 transform. *Bioinformatics*, **25**, 1754–1760.
656 Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and
657 SAMtools. *Bioinformatics*, **25**, 2078–2079.

658 Malmstrøm M, Matschiner M, Tørresen OK *et al.* (2016) Evolution of the immune
659 system influences speciation rates in teleost fishes. *Nature Genetics*, **48**, 1204–1210.
660 Malmstrøm M, Matschiner M, Tørresen OK, Jakobsen KS, Jentoft S (2017) Whole
661 genome sequencing data and de novo draft assemblies for 66 teleost species.
662 *Scientific Data*, **4**, 160132–13.
663 Mascher M, Muehlbauer GJ, Rokhsar DS *et al.* (2013) Anchoring and ordering NGS
664 contig assemblies by population sequencing (POPSEQ). *The Plant Journal*, **76**, 718–
665 727.
666 Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation
667 sequencing data. *Genomics*, **95**, 315–327.
668 Morin A, Kwan T, Ge B *et al.* (2016) Immunoseq: the identification of functionally
669 relevant variants through targeted capture and sequencing of active regulatory
670 regions in human immune cells. *BMC Medical Genomics*, **9**, 1–12.
671 Newman CE, Austin CC (2016) Sequence capture and next-generation sequencing of
672 ultraconserved elements in a large-genome salamander. *Molecular Ecology*, **25**,
673 6162–6174.
674 Opazo JC, Butts GT, Nery MF, Storz JF, Hoffmann FG (2012) Whole-Genome
675 Duplication and the Functional Diversification of Teleost Fish Hemoglobins.
676 *Molecular Biology and Evolution*, **30**, 140–153.
677 Quinn NL, Boroevich KA, Lubieniecki KP *et al.* (2010) Genomic organization and
678 evolution of the Atlantic salmon hemoglobin repertoire. *BMC Genomics*, **11**, 539–
679 17.
680 Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open
681 Software Suite. *Trends in Genetics*, **16**, 276–277.
682 Robinson JT, Thorvaldsdóttir H, Winckler W *et al.* (2011) Integrative genomics
683 viewer. *Nature Biotechnology*, **29**, 24–26.
684 Samorodnitsky E, Datta J, Jewell BM *et al.* (2015) Comparison of Custom Capture for
685 Targeted Next-Generation DNA Sequencing. *The Journal of Molecular Diagnostics*,
686 **17**, 64–75.
687 Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic
688 datasets. *Bioinformatics*, **27**, 863–864.
689 Schott RK, Panesar B, Card DC *et al.* (2017) Targeted Capture of Complete Coding
690 Regions across Divergent Species. *Genome Biology and Evolution*, **9**, 398–414.
691 Smit A, Hubley R, Green P. RepeatMasker Open 3.0. http://www.repeatmasker.org.
692 2010. (Last accessed December 2017)
693 Star B, Nederbragt AJ, Jentoft S *et al.* (2011) The genome sequence of Atlantic cod
694 reveals a unique immune system. *Nature*, **477**, 207–210.

695 Syring JV, Tennessen JA, Jennings TN *et al.* (2016) Targeted Capture Sequencing in
696 Whitebark Pine Reveals Range-Wide Demographic and Adaptive Patterns
697 Despite Challenges of a Large, Repetitive Genome. *Frontiers in Plant Science*, **7**,
698 403–15.
699 Team RC. R Core Team. R: A language and environment for statistical computing. R
700 Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL
701 http://www. R-project. org/.(3.3. 1) Software Vienna, Austria: R Foundation for
702 Statistical Computing; 2013.
703 Teer JK, Bonnycastle LL, Chines PS *et al.* (2010) Systematic comparison of three
704 genomic enrichment methods for massively parallel DNA sequencing. *Genome*
705 *Research*, **20**, 1420–1431.
706 Turner EH, Ng SB, Nickerson DA, Shendure J (2009) Methods for Genomic
707 Partitioning. *Annual Review of Genomics and Human Genetics*, **10**, 263–284.
708 Tørresen OK, Brieuc MSO, Solbakken MH *et al.* (2018) Genomic architecture of
709 haddock (*Melanogrammus aeglefinus*) shows expansions of innate immune genes
710 and short tandem repeats. *BMC Genomics*, **19**, 51.
711 Tørresen OK, Star B, Jentoft S *et al.* (2017) An improved genome assembly uncovers
712 prolific tandem repeats in Atlantic cod. *BMC Genomics*, **18**, 95.
713 Volff J-N (2004) Genome evolution and biodiversity in teleost fish. *Heredity*, **94**, 280–
714 294.
715 Wang M, Beck CR, English AC *et al.* (2015) PacBio-LITS: a large-insert targeted
716 sequencing method for characterization of human disease-associated
717 chromosomal structural variations. *BMC Genomics*, **16**, 214.
718 Wells RMG (2005) Blood-Gas Transport and Hemoglobin Function in Polar Fishes:
719 Does Low Temperature Explain Physiological Characters? In: *Physiology of Polar*
720 *Fishes* Fish Physiology. pp. 281–316. Academic Press.
721 Wetten OF, Nederbragt AJ, Wilson RC *et al.* (2010) Genomic organization and gene
722 expression of the multiple globins in Atlantic cod: conservation of globin-
723 flanking genes in chordates infers the origin of the vertebrate globin clusters.
724 *BMC Evolutionary Biology*, **10**, 315.
725 Witek K, Jupe F, Witek AI *et al.* (2016) Accelerated cloning of a potato late blight–
726 resistance gene using RenSeq and SMRT sequencing. *Nature Biotechnology*, **34**,
727 656–660.
728 Woolfe A, Goodson M, Goode DK *et al.* (2004) Highly Conserved Non-Coding
729 Sequences Are Associated with Vertebrate Development (Sean Eddy, Ed,). *PLoS*
730 *Biology*, **3**, e7–15.

Yoshihara M, Saito D, Sato T *et al.* (2016) Design and application of a target capture sequencing of exons and conserved non- coding sequences for the rat. *BMC Genomics*, **17**, 1–11.

Figure legends

Figure 1: Flowchart of sequence capture approach. a) Sequence data from the Atlantic cod genome (Tørresen *et al.* 2017; gadMor2) combined with gene sequences of target genes and sequences from low coverage genomes of the additional codfishes are combined to generate probes. b) Isolated DNA is multiplexed with Illumina and PacBio barcodes. c) Raw reads for each species are used to score all probes, ensuring that no repeated sequences are present. DNA Probes are used in solution on isolated DNA for all of the included species, hybridizing to the target sequences. Target sequences are then captures and sequences on the PacBio RSII sequencing platform. d) Downstream bioinformatics includes de-multiplexing of reads and trimming, making the reads ready for downstream analysis such as mapping and *de novo* assembly.

Figure 2: Mapping of reads and assemblies against the MN target region. Each panel shows the reads and *de novo* assembly mapped against the MN target region in grey and orange, respectively, for species a.) Atlantic cod, b) haddock, c) silvery pout, d) cusk, e) burbot, f) European hake, g) marbled moray cod and h) roughhead grenadier. The positions of genes in the target region are indicated at the top.

Figure 3: Mapping of reads and assemblies against the LA target region. Each panel shows the reads and *de novo* assembly mapped against the LA target region in grey and orange, respectively, for species a) Atlantic cod, b) haddock, c) silvery pout, d) cusk, e) burbot, f) European hake, g) marbled moray cod and h) roughhead grenadier. The positions of genes in the target region are indicated at the top.

Figure 4: The relationship between capture success and genomic divergence to Atlantic cod. Linear regression of the relationship between the genomic divergence to Atlantic cod (SNPs x 10^5) and a) median mapping depth for the MN region (blue), LA region (red) and the combined target region (black); b) the percentage of reads mapping to the target region; c) the percentage of the target region covered by reads to a minimum depth of 10x. For each regression the correlation coefficient, r, is shown along with a p-value. Each data point is labeled by species according to this code: Ac=Atlantic cod, H=haddock, Sp=silvery pout, C=cusk, B=burbot, Eh=European hake, Mm=marbled moray cod and Rg=roughhead grenadier.

Figure 5: Synteny of the Hb gene clusters. Genomic synteny of the hemoglobin gene clusters shown at the top for the genomes of Atlantic cod (Tørresen *et al.* 2017; gadMor2) and haddock (Tørresen *et al.* 2018; melAeg). Below, the genomic synteny inferred from the *de novo* assemblies for all of the species included in the capture experiment. Stippled lines indicate assembly gaps – here we assume that the orientation of genes corresponds to the genomes of Atlantic cod and haddock. Gray boxes indicate genes that have been identified in Baalsrud *et al.* (2017) but are absent in the *de novo* assemblies. a) Synteny across the MN region b) Synteny across the LA region.

Figure 6: Polymorphisms in the bi-directional promoter between *Hba1* and *Hbb1* for five species in the Gadidae family.
A schematic representation of *Hba1* and *Hbb1* with the promoter region between them. The region contains an indel polymorphism of variable length across the five species, as indicated by gaps. For each species/variant the alignment is shown along with amino acid substitutions at positions 55 and 62 in the translated part of the *Hbb1* gene.

Figures

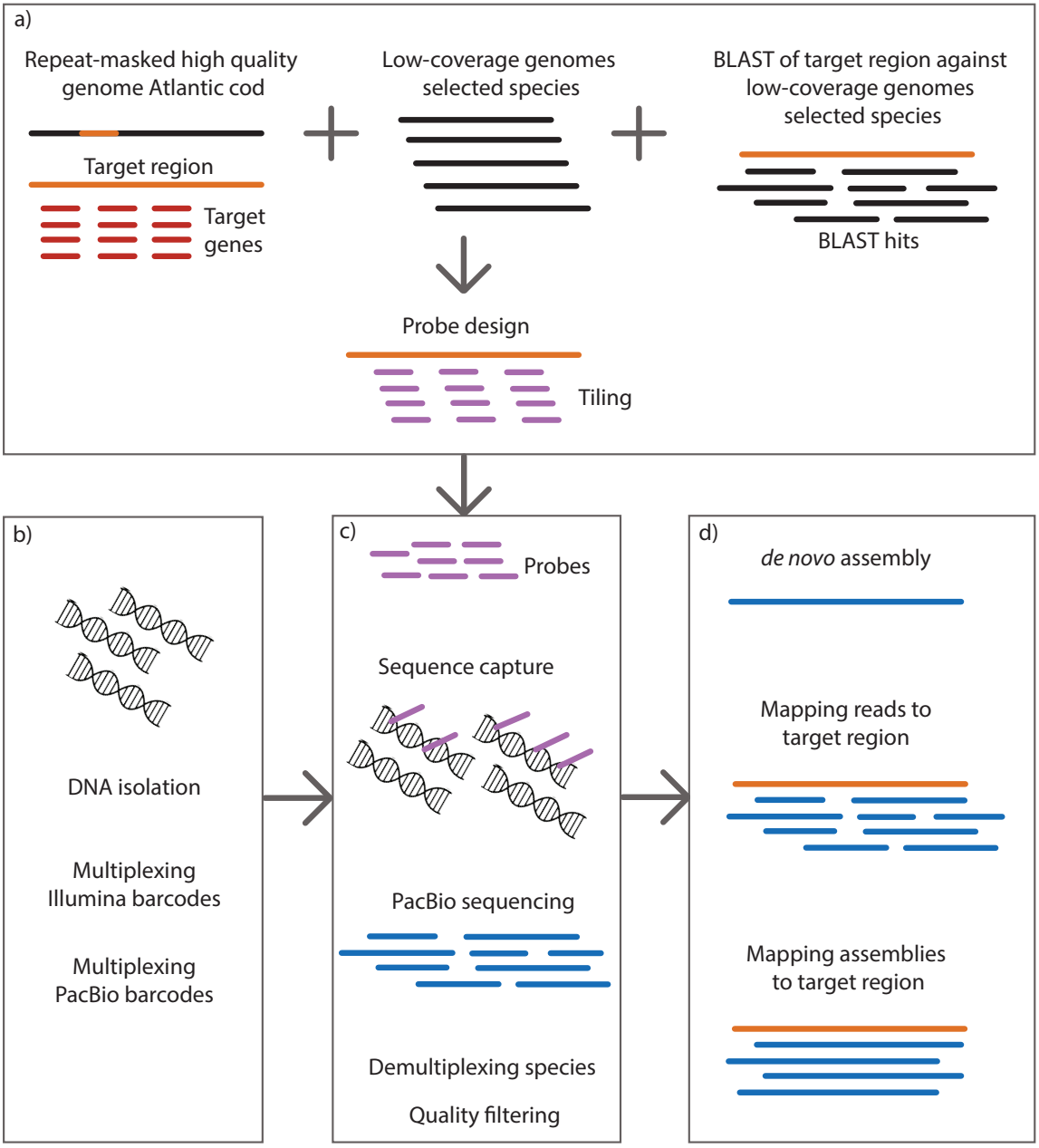

Figure 1

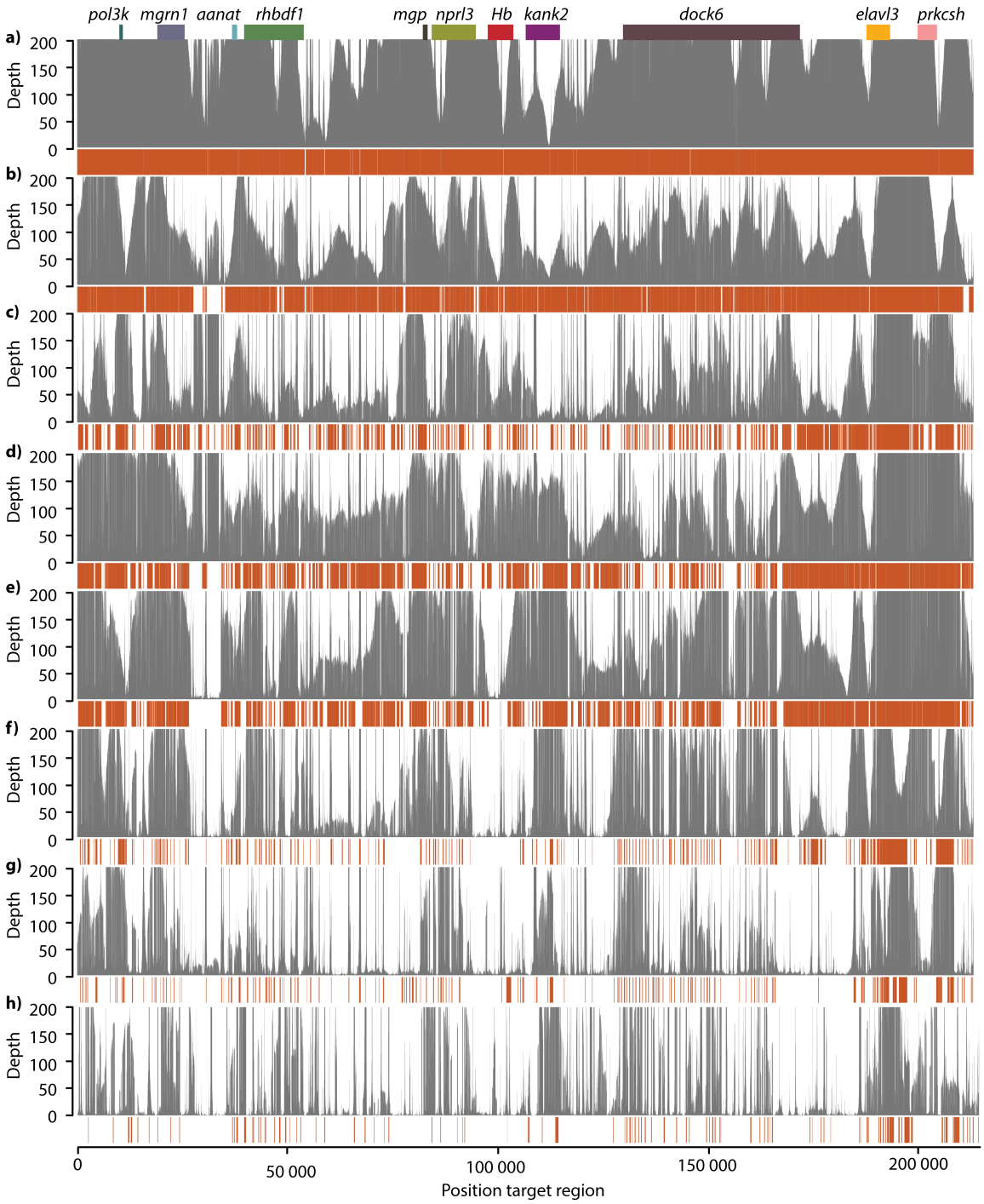
pol3k
mgrn1
aanat
rhbdf1
mgp
nprl3
Hb
kank2
dock6
elavl3
prkcsh
a)
b)
c)
d)
e)
f)
g)
h)
Depth
200
150
100
50
0
0
50 000
100 000
150 000
200 000
Position target region

Figure 2

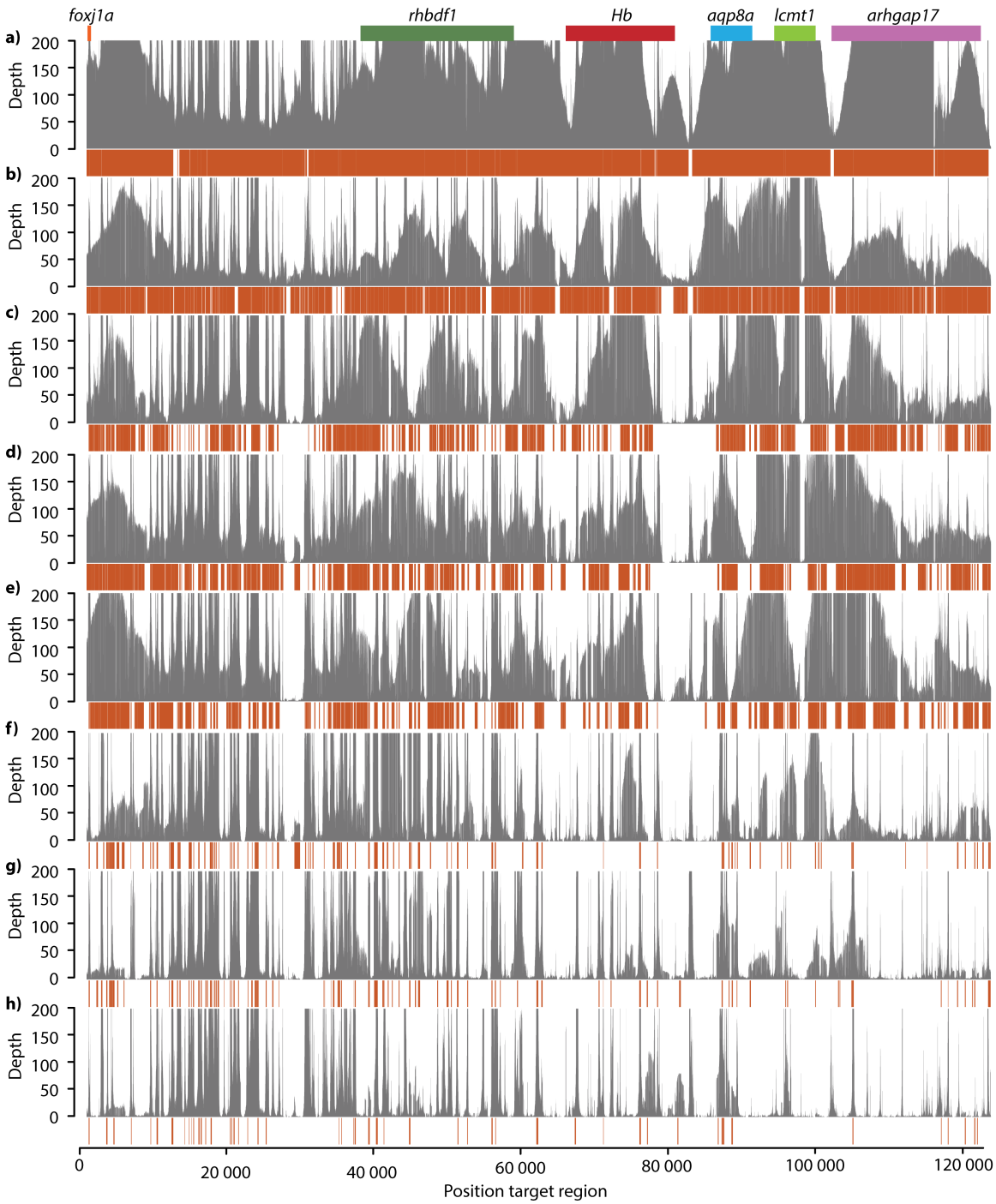
foxj1a
rhbdf1
Hb
aqp8a
lcmt1
arhgap17
a)
b)
c)
d)
e)
f)
g)
h)
Depth
200
150
100
50
0
0
20 000
40 000
60 000
80 000
100 000
120 000
Position target region

Figure 3

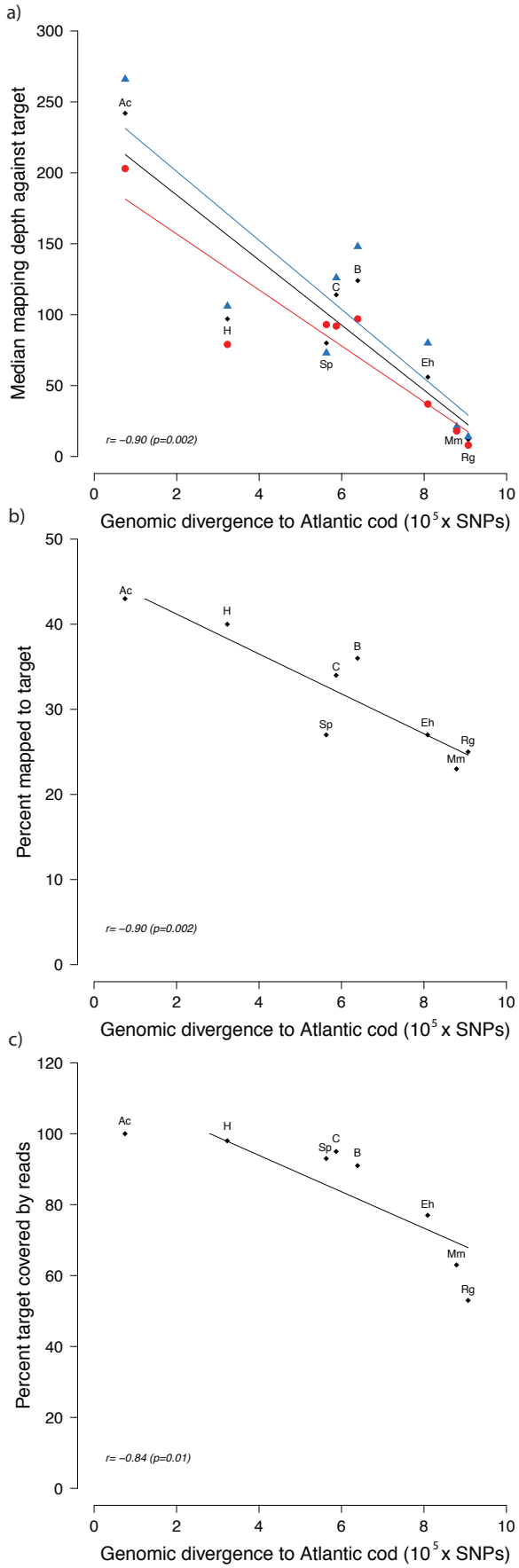
a)
Median mapping depth against target
300
250
200
150
100
50
0
Ac
H
C
B
Sp
Eh
Mm
Rg
r= −0.90 (p=0.002)
0
2
4
6
8
10
Genomic divergence to Atlantic cod (10^5 x SNPs)
b)
Percent mapped to target
50
40
30
20
10
0
Ac
H
B
C
Sp
Eh
Rg
Mm
r= −0.90 (p=0.002)
0
2
4
6
8
10
Genomic divergence to Atlantic cod (10^5 x SNPs)
c)
Percent target covered by reads
120
100
80
60
40
20
0
Ac
H
Sp
C
B
Eh
Mm
Rg
r= −0.84 (p=0.01)
0
2
4
6
8
10
Genomic divergence to Atlantic cod (10^5 x SNPs)

Figure 4

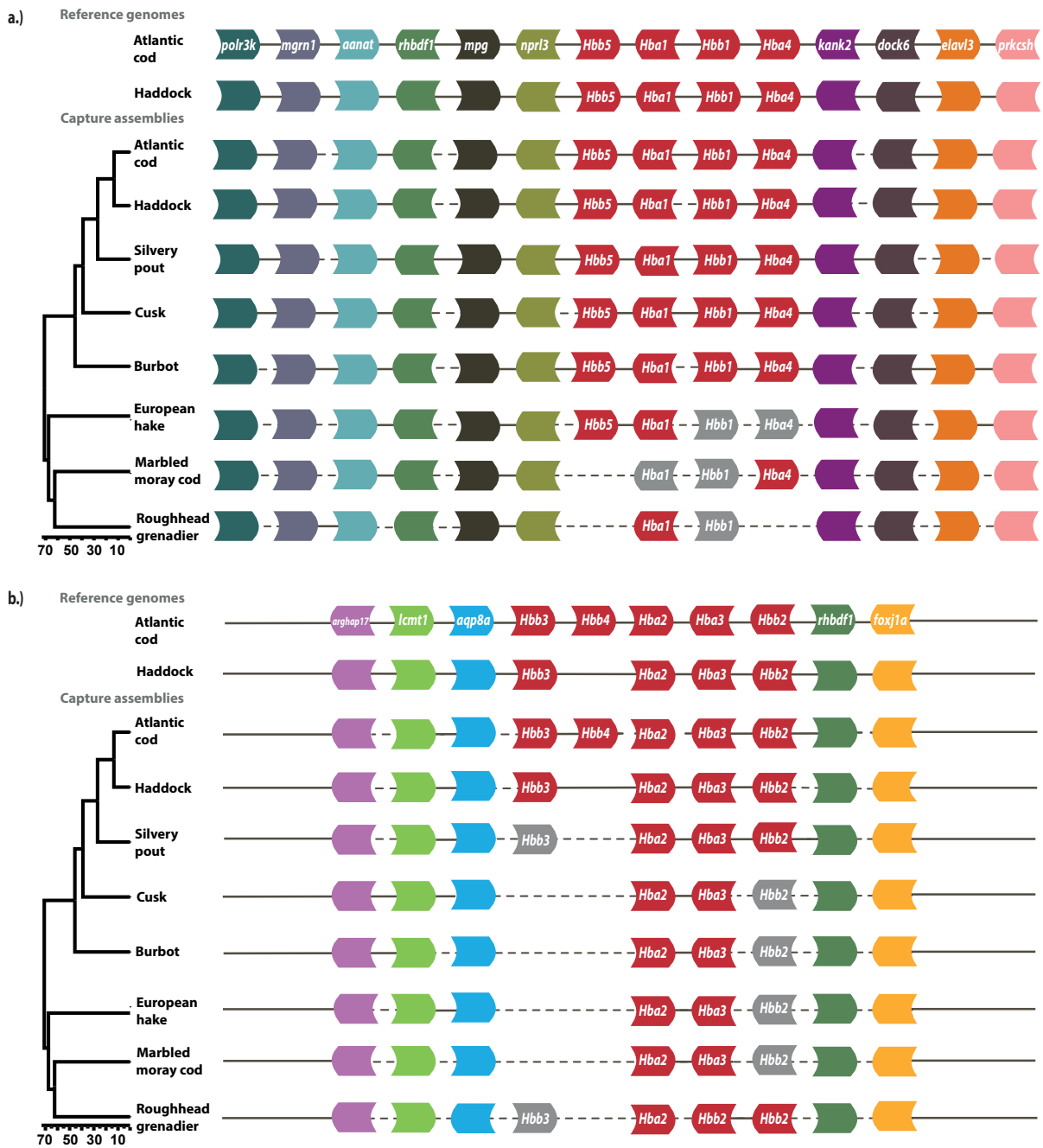
a.)
Reference genomes
Atlantic cod
Haddock
Capture assemblies
Atlantic cod
Haddock
Silvery pout
Cusk
Burbot
European hake
Marbled moray cod
Roughhead grenadier
70 50 30 10
polr3k
mgrn1
aanat
rhbdf1
mpg
nprl3
Hbb5
Hba1
Hbb1
Hba4
kank2
dock6
elavl3
prkcsh
b.)
Reference genomes
Atlantic cod
Haddock
Capture assemblies
Atlantic cod
Haddock
Silvery pout
Cusk
Burbot
European hake
Marbled moray cod
Roughhead grenadier
70 50 30 10
arghap17
lcmt1
aqp8a
Hbb3
Hbb4
Hba2
Hba3
Hbb2
rhbdf1
foxj1a

Figure 5

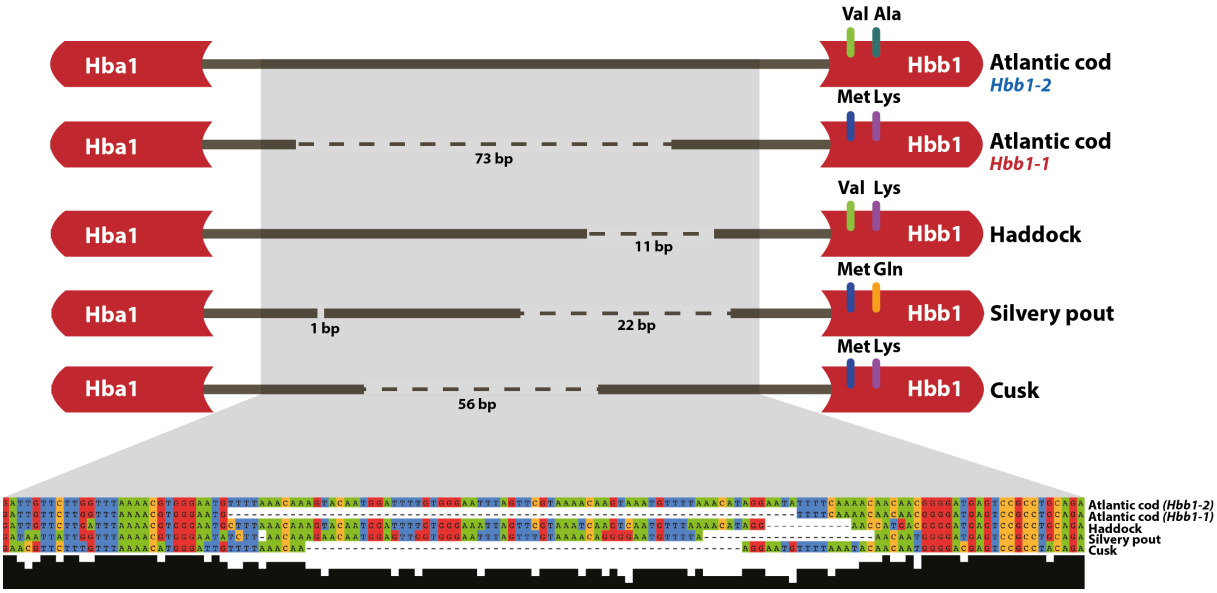
Val Ala
Hba1
Hbb1
Atlantic cod
Hbb1-2
Met Lys
Hba1
73 bp
Hbb1
Atlantic cod
Hbb1-1
Val Lys
Hba1
11 bp
Hbb1
Haddock
Met Gln
Hba1
1 bp
22 bp
Hbb1
Silvery pout
Met Lys
Hba1
56 bp
Hbb1
Cusk
Atlantic cod (Hbb1-2)
Atlantic cod (Hbb1-1)
Haddock
Silvery pout
Cusk

Figure 6

806 # Tables

807 **Table 1.** Number of reads and bases captured and sequenced for each species, and
808 number of utgs, largest utg and N50 in the assemblies.

Species	Latin name	Number of reads	Number of bases	Number of utgs	Largest utg (bp)	N50 (bp)
Atlantic cod	*Gadus morhua*	73005	217252583	278	79 020	7 728
Haddock	*Melanogrammus aeglefinus*	35573	107839552	227	52 433	7 227
Silvery pout	*Gadiculus argenteus*	69775	212519845	410	35 801	7 098
Cusk	*Brosme brosme*	55348	175883008	394	64 145	7 322
Burbot	*Lota lota*	56155	165360828	205	70 602	8 055
European hake	*Merluccius merluccius*	65661	180558336	311	31 558	6 523
Marbled moray cod	*Muraenolepus marmoratus*	52076	148100933	455	30 019	6 632
Roughhead grenadier	*Macrourus berglax*	46195	129085001	325	35 216	7122

809
810
811
812
813
814