PERMANENT GENETIC RESOURCES ARTICLE

BAC library construction, screening and clone sequencing of lake whitefish (Coregonus clupeaformis, Salmonidae) towards the elucidation of adaptive species divergence

J. JEUKENS,*† B. BOYLE,*‡ I. KUKAVICA-IBRULJ,*§ J. ST-CYR,* R. C. LÉVESQUE*§ and L. BERNATCHEZ*†
*Institut de biologie intégrative et des systèmes (IBIS), 1030 av. de la Médecine, Québec City, Québec, G1V 0A6, Canada,
†Quebec-Ocean, Department of biology, Université Laval, Québec City, Québec, Canada,
‡Arborea, Center for Forest Research, Université Laval, Québec City, Québec, Canada,
§Department of microbiology-infectiology and immunology, Université Laval, Québec City, Québec, Canada

Abstract

Genomic DNA sequences and other genomic resources are essential towards the elucidation of the genomic bases of adaptive divergence and reproductive isolation. Here, we describe the construction, characterization and screening of a nonarayed BAC library for lake whitefish (Coregonus clupeaformis). We then show how the combined use of BAC library screening and next-generation sequencing can lead to efficient full-length assembly of candidate genes. The lake whitefish BAC library consists of 181 050 clones derived from a single heterozygous fish. The mean insert size is 92 Kb, representing 5.2 haploid genome equivalents. Ten BAC clones were isolated following a quantitative real-time PCR screening approach that targeted five previously identified candidate genes. Sequencing of these clones on a 454 GS FLX system yielded 178 000 reads with a mean length of 358 bp, for a total of 63.8 Mb. De novo assembly and annotation then allowed retrieval of contigs corresponding to each candidate gene, which also contained up- and/or downstream noncoding sequences. These results suggest that the lake whitefish BAC library combined with next-generation sequencing technologies will be key resources to achieve a better understanding of both adaptive divergence and reproductive isolation in lake whitefish species pairs as well as salmonid evolution in general.

Keywords: BAC library, qPCR screening, 454 sequencing, coregonus, speciation

Received 8 October 2010; revision received 2 December 2010; accepted 10 December 2010

Introduction

During the past years, the identification of genes and genomic regions associated with population divergence and speciation has become a highly productive research area (e.g. Hoekstra et al. 2006; Joron et al. 2006; Schemske & Bierzychudek 2007; Storz et al. 2007; Barrett et al. 2008; Kane et al. 2009). Still, the elucidation of causative genetic variation underlying phenotypic change and reproductive isolation remains one of the major challenges of evolutionary biology (Edmands 2002; Coyne & Orr 2004; de Queiroz 2005; Storz & Wheat 2010). For many study systems, however, this undertaking is greatly hindered by the lack of extensive sequence information and other genomic resources.

In lake whitefish (Coregonus clupeaformis), the study of adaptive species divergence has been greatly aided by a growing array of molecular tools (reviewed in Bernatchez et al. 2010). Mitochondrial DNA variation has portrayed its North American phylogeographic structure in relation to Pleistocene glaciations (Bernatchez & Dodson 1990, 1991; Pigeon et al. 1997). It has also confirmed independent parallel evolution of sympatric pairs of dwarf and normal lake whitefish following secondary contact of glacial races. Then, the use of microsatellite and AFLP markers has provided molecular evidence of restricted gene flow for a small proportion of loci that might be under the influence of directional selection (Lu & Bernatchez 1999; Lu et al. 2001; Rogers et al. 2001; Campbell & Bernatchez 2004). Quantitative trait loci (QTL) mapping has led to the demonstration of genetic bases for many traits that differ between dwarf and normal lake whitefish, namely swimming behaviour (Rogers et al. 2002), growth (Rogers & Bernatchez 2005), morphology and life
history (Rogers & Bernatchez 2007). It has also provided evidence of intrinsic and extrinsic postzygotic barriers to reproduction between them (Rogers & Bernatchez 2006).

At the transcriptome level, the salmonid cDNA cGRASP microarray (Rise et al. 2004; von Schalburg et al. 2005) has allowed the identification of expression divergence for genes and key gene functions potentially implicated in the adaptive divergence of dwarf and normal lake whitefish (Derome et al. 2006; St-Cyr et al. 2008; Nolte et al. 2009). Subsequently, the integrated use of linkage, phenotypic and gene expression mapping provided insight into the genetic architecture of adaptive traits differentiating dwarf and normal whitefish, with the identification of key genomic regions which appear to have high pleiotropic effects on gene expression (Derome et al. 2008; Whiteley et al. 2008). A recent massively parallel pyrosequencing experiment has allowed whole transcriptome sequencing and efficient single nucleotide polymorphism (SNP) discovery, thus showing how next-generation sequencing and efficient single nucleotide parallel pyrosequencing experiment has allowed whole library construction was largely established according to (Di Palma et al. 2007).

DNA source
Blood cells were isolated from a single lake whitefish reared in controlled conditions at LARSA (Laboratoire régional des sciences aquatiques, Université Laval, Quebec, QC, Canada).

BAC vector
We used the CopyControl pCC1BAC HindIII Cloning-Ready Vector (Epigensoft Technologies, Madison, WI, USA). Construction with pCC1BAC EcoRI was first attempted, but yielded a high percentage (75–100%) of deleted vector bands with no inserts (data not shown). This was probably because of a high level of EcoRI enzyme star activity in the vector preparation (Di Palma et al. 2007).

HMW DNA preparation
Immediately after collection from the caudal peduncle of the fish, 300 μL volumes of blood were transferred to tubes containing 300 μL of 0.5M EDTA to avoid clotting. Cold phosphate-buffered saline (PBS) was then added to each tube (300 μL), and centrifugation was carried out at maximum speed for 5 min. Supernatants were removed, and 8 more PBS washes were performed. Blood cells were then diluted to 1 × 10^8 cells/mL (after cell counting with a haemocytometer) and combined with an equal volume of 1% low melting point agarose (InCert agarose, Thorsen et al. 2004) and held on ice for 30–60 min. HMW DNA was extracted from the cells according to the following steps: (i) 48 h at 50 °C with occasional shaking in freshly prepared lysis solution (0.2% w/v proteinase K, 80% 0.5 μ EDTA pH 8, 20% N-lauroylsarcosine 10%, replaced after 24 h); (ii) generous rinsing with distilled water; (iii) 24 h at 4 °C with gentle mixing in TE50 (10 μm Tris-HCl pH 8, 50 μm EDTA, replaced twice); (iv) 2 × 2 h at 4 °C with gentle mixing in PMSF 0.1 μm (diluted in TE50); and (v) 24 h at 4 °C with gentle mixing in TE50 (replaced twice). Agarose plugs were then kept in 0.5 μ EDTA at 4 °C.

Partial digestion of HMW DNA and size selection
Agarose plugs were dialysed in TE (10 μm Tris-HCl pH 8, 1 μm EDTA) for 24 h. Each plug was cut into four...
equal pieces. Each piece was then placed in a tube containing 500 µL of 1x HindIII Buffer (NEB buffer 2, New England Biolabs, Pickering, ON, USA) for 1 h at 4 °C. Buffer was replaced and supplemented with 15 U HindIII enzyme. Tubes were left on ice for 2 h and then at 37 °C for 35 min. The enzyme was inactivated with 20 µL of 0.5 M EDTA for 1 h on ice. Size fractionation was performed with a CHEF-DR III System (Bio-Rad, Hercules, CA, USA) in 0.5x TBE buffer at 14 °C in three steps. First, digested DNA was separated with the following conditions: 5 V/cm, 6 h, 5–20 s pulse time. Portions of the gel containing original plugs and DNA fragments <50 kb were discarded, and fresh 1% Pulse Field Certified agarose (Bio-Rad) was added to the remaining portion. Second, electrophoresis was performed with the following conditions: 5 V/cm, 19 h, 5–20 s pulse time. Bands of 0.5 cm were then cut from the portion of the gel containing fragments ranging from 100 to 250 kb. Third, the content of selected bands was concentrated by inverting each band in fresh agarose and performing electrophoresis with the following conditions: 5V/cm, 19 h, 5–20 s pulse time.

**Ligation and electroporation**

HMW DNA was retrieved from agarose bands by electroelution using SnakeSkin Pleated dialysis tubing (10 000 MWCO, Pierce, Rockland, IL, USA) with the following conditions: 0.5x TBE, 3 V/cm for 3–4 h, polarity inversion of 30 s at the end. Tubes were then dialysed in TE at 4 °C for at least 2 h. Eluted DNA was always manipulated with care using wide-bore pipette tips and kept at 4 °C for a maximum of 10 days. Ligation was carried out in individual 50-µL reactions with 50 ng DNA, 10 ng vector (approximately 4:1 vector to insert molar ratio) and 400 U ligase at 16 °C overnight. These conditions were adjusted to minimize the proportion of noninsert clones. The enzyme was heat inactivated for 10 min at 65 °C. Ligation products were placed on 0.025-µm membrane filters (Millipore, Billerica, MA, USA), dialysed on water for 2 h at room temperature and concentrated on 30% (w/v) PEG for 45 min. Transformation of the ligation products was performed using electrocompetent *E. coli* DH10B T1 phage-resistant cells (ElectroMAX DH10B T1 resistant, Invitrogen, Burlington, ON, USA) with the following conditions: 20 µL of cells + 2 µL of ligation product, 1 mm cuvettes, 200 Ω, 25 µF, 1.3 kV. After 45–60 min of shaking at 37 °C in SOC medium, cells from different ligation reactions were pooled and plated on two small Petri dishes for colony counting and large Petri dishes for growth (LB agar, 12.5 µg/mL chloramphenicol, 40 µg/mL XGal, 100 µg/mL IPTG). Finally, colonies were retrieved from the large Petri dishes by grating with LB broth, concentrated by centrifugation, supplemented with 15% glycerol, flash frozen in liquid nitrogen and stored at −80 °C.

**Insert size analysis**

Nine to 27 white colonies (recombinant colonies) from each pool of ligation reactions (depending on the total number of clones in each pool) were inoculated in brain heart infusion medium containing 20 µg/mL chloramphenicol. Clones were grown for 14–16 h, and BAC DNA was purified following a modified alkaline lysis protocol (Osoegawa et al. 2001). About 150 ng of BAC DNA was then digested with 5 U of NotI and submitted to migration on a CHEF apparatus (0.5x TBE buffer, 14 °C, 5 V/cm, 19 h, 5–20 s pulse time). Mid-range PFGE Marker (New England Biolabs, Pickering, ON, USA) and lambda DNA HindIII digest in combination with the GeneTools software (Syngene, Frederick, MD, USA) were used for molecular weight determination.

**Library screening by quantitative real-time PCR (qPCR) and regular PCR**

Five genes previously identified as potentially associated with adaptive divergence between dwarf and normal whitefish were selected as targets for library screening (Table 1). For each target gene, a set of four specific primers were designed such that two small amplicons (100–200 bp) were available for qPCR and a longer one (400–1000 bp) could be used for regular PCR and Sanger sequencing (Table 1). All primers were tested by regular PCR and qPCR using genomic DNA extracted from the fish that was used for library construction. The library was arrayed into pools of 500 white colony-forming units (CFU) per well, for a total of four 96 deep-well plates. When grown in such plates, bacteria were inoculated in 1.5 mL brain heart infusion medium containing 20 µg/mL chloramphenicol and put in a 37 °C orbital shaker at 200 rpm. A summary of the screening strategy is presented in Fig. 1. From pool plates, row and column superpools were created, followed by BAC DNA isolation.

The first step of the screening strategy was to amplify the two small amplicons for each target gene by quantitative real-time PCR (qPCR) to identify positive superpools. This allowed indirect identification of positive pools with intersections of positive row and column superpools. Fifteen-microlitre PCRs (1x QuantiTect SYBR Green master mix (Qiagen, Hilden, Germany), 0.3 µM of each primer, 10 ng DNA) were prepared using an epMotion 5075 robot (Eppendorf, Hamburg, Germany) and run on a LightCycler 480 (Roche, Basel, Switzerland) with the following conditions: 15 min activation at 95 °C followed by 40 cycles consisting of 10 s at 95 °C and 2 min
### Table 1  Screening target genes and primers

<table>
<thead>
<tr>
<th>Gene name (code)</th>
<th>Function*</th>
<th>Primer pair 1 (1–2)†</th>
<th>Primer pair 2 (3–4)‡</th>
<th>Amplicon length (bp) (1–4)§</th>
<th>References¶</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liver carboxylesterase 22 precursor (CARB)</td>
<td>Detoxification</td>
<td>AGCTTCTGAAGGGAGGAGAC  AAGGGGAGGAGCTGATACT</td>
<td>TGTGCTGGACTGTTGTGTA  GATGGCATTTGGGCAAATTT</td>
<td>377</td>
<td>3, 7, 10</td>
</tr>
<tr>
<td>Heat shock cognate 70 kDa protein (HSP)</td>
<td>Protein folding</td>
<td>AAGGTTCCTTCAAAGAATCTAACG</td>
<td>TGTGCTGGACTGTTGTGTA  GATGGCATTTGGGCAAATTT</td>
<td>377</td>
<td>3, 7, 10</td>
</tr>
<tr>
<td>Malate dehydrogenase (MDH)</td>
<td>Energy metabolism</td>
<td>CTTGTTAGTGGAAACCCTGGTAAC  GATGGCATTTGGGCAAATTT</td>
<td>TGTGCTGGACTGTTGTGTA  GATGGCATTTGGGCAAATTT</td>
<td>377</td>
<td>3, 7, 10</td>
</tr>
<tr>
<td>Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)</td>
<td>Energy metabolism</td>
<td>CCGATACTCCCTAAAGGACTCTGA  GATGGCATTTGGGCAAATTT</td>
<td>TGTGCTGGACTGTTGTGTA  GATGGCATTTGGGCAAATTT</td>
<td>377</td>
<td>3, 7, 10</td>
</tr>
<tr>
<td>MHC classII beta, exon B (MHC)</td>
<td>Immunity</td>
<td>CCGATACTCCCTAAAGGACTCTGA  GATGGCATTTGGGCAAATTT</td>
<td>TGTGCTGGACTGTTGTGTA  GATGGCATTTGGGCAAATTT</td>
<td>377</td>
<td>3, 7, 10</td>
</tr>
</tbody>
</table>

* Determined using Gene Ontology (GO) biological process terms.
† Short amplicon for qPCR (100–200 pb); Bold: left primer for regular PCR (1–4) and sequencing.
‡ Short amplicon for qPCR (100–200 pb); Bold: right primer for regular PCR (1–4) and sequencing.
§ Length of the regular PCR amplicon.
¶ Gene expression divergence between dwarf and normal lake whitefish: 1 (Derome et al. 2006), 2 (Derome et al. 2008), 3 (St-Cyr et al. 2008), 4 (Nolte et al. 2009), 6 (Whiteley et al. 2008), 7 (Jeukens et al. 2009), 10 (Jeukens et al. 2010); Highly transgressive gene in hybrid whitefish: 5 (Renaut et al. 2009); Divergent SNP(s): 8 (Renaut et al. 2010a); Evidence of balancing selection in the European whitefish: 9 (Binz et al. 2001).
**Left primer for qPCR only; the left primer for regular PCR and sequencing is: AGAACATCATCCCTGCCTCCAC.
†† 1–4 amplicon was too short to include 2 amplicons, 1–2 and 1–4 were used for qPCR.
at 62 °C. A single fluorescent read was taken immediately following the 2-min extension time. A melting cycle was performed following the acquisition cycles by melting the reaction at 95 °C for 10 s, annealing at 55 °C for 1 min and constant temperature ramping from 55 to 95 °C with data acquisition every 0.2 °C. Positive reactions were identified by the positioning of the amplification profile (Cp) and amplicon melting profile in comparison with the genomic DNA control. Quantitative PCR had many advantages compared to traditional PCR. It relieved the use of agarose gels but, foremost, prevented cross-contamination because qPCR tubes were never opened after a run. DNA was extracted from the corresponding pools, which were validated by regular PCR and sequencing of the long amplicon (primers 1–4). Two independent pools were selected for each gene.

The second step consisted in the dilution of each positive pool to 200 white CFU per well in a total of 24–32 wells. Plates with 200 or less CFU per well were grown for two nights. DNA of column superpools was extracted and tested by qPCR using the same conditions as in the first step, but on an ABI PRISM 7500 thermocycler (Applied Biosystems, Foster City, CA, USA) in 25 μL reaction volumes (in 96-well plates, reactions prepared manually), as this step was low throughput compared to the previous one. Then, 1 μL of bacterial culture from each well of a positive column was tested by regular PCR.

The third step of the screening strategy consisted in serial dilution of a positive 200 CFU well for each original selected pool such that on one plate, 16 wells had 100 white CFU each, 32 wells had 50 CFU and 48 wells had 25 CFU. DNA of column superpools was extracted, and a procedure identical to the second step was performed. If the most diluted positive well had a concentration of 100 CFU, this step was carried out a second time, starting from the positive well. If the positive well had a concentration of 50 or 25 CFU, we proceeded to clone picking. Ninety six white colonies were randomly picked, grown into plates, and a procedure identical to the second step was performed. Positive isolated clones were validated by regular PCR and sequencing of the long amplicon. Insert size was also determined (see Insert size analysis).
Sequencing, contig assembly and analysis

Fresh BAC DNA preparations were made for each isolated clone, with final elution in UltraPure 0.01 M Tris-HCl pH 8 (Invitrogen, Burlington, ON, USA). Samples were quantified using the Quant-iT Picogreen dsDNA Assay Kit (Invitrogen, Burlington, ON, USA). Then, for each target gene, the two clones were combined in a 1:1 molar ratio and treated for RNA contamination (Ribo-Shredder™ RNase Blend, Epicentre Biotechnologies, Madison, WI, USA). Per-target shotgun library preparation, tagging (454 Multiplex Identifiers, MIDs) and sequencing were carried out at the IBIS biomolecular analysis platform (Université Laval, Québec, Canada) on a 454 Genome Sequencer FLX System, with long-read GS FLX Titanium chemistry, using methods previously described (Margulies et al. 2005). Libraries were pooled and sequenced on a quarter plate. Initial quality filtering and base calling of 454 sequence data were performed using Roche proprietary analysis software Newbler (Margulies et al. 2005). Two programs were then used to assemble contigs de novo for each target gene. First, GS De novo Assembler software (Roche, Basel, Switzerland) was used with the following parameters: large or complex genome option, heterozygosity option for MHC only (because the two screening amplicons had 92% identity), trimming database of the pCC1BAC HindIII vector, screening database of the E. coli str. K12 substr. DH10B genome [GenBank (accession number CP000948)], minimum overlap length 200 bp, minimum overlap identity 98% and minimum contig length 500 bp. Second, CLC Genomics Workbench yielded three times the number of contigs assembled by GS De novo Assembler; hence, only results produced with the latter program are presented. De novo assembly resulted in a highly variable number of contigs per target gene, with HSC (9 contigs) and CARB (10 contigs) at one extreme and MHC (69 contigs) at the other (Table 3).

Contig annotation

Mean gene density was 9.8/ Mb (i.e. 1–2 genes per clone) according to blast search of the salmonid nucleotide collection and 14.7/ Mb (i.e. 1–4 genes per clone) according to blast search with omission of salmonid sequences. Contigs containing screening amplicons were identified by local blast. One contig was identified for each amplicon except for MHC (two contigs, Table 3). For all target genes except GAPDH, up- and/or downstream noncoding sequences were also present on the same contig. Moreover, blast search revealed that, except for MDH, target genes were split into up to four contigs, which could then be manually assembled using salmonid partial or complete coding sequences available in GenBank assuming a genome size of 3 Gb (Table 2, Animal Genome Size Database). According to 117 randomly picked clones, insert sizes range between 36 and 172 Kb, with an estimated average of 92 Kb. These results are comparable to those of Di Palma et al. (2007) for BAC libraries of Metriacanthus zebra (haplochromine cichlid) and Astyanax mexicanus (Mexican tetra), using a similar approach.

Sequencing and contig assembly

Pyrosequencing of 10 BAC clones pooled in a quarter plate yielded 178 000 reads with a mean length of 358 bp, for a total of 63.8 Mb (sequence data available through the Sequence Read Archive at NCBI: SRP003484). Screening and sequencing results for each target gene are summarized in Table 3. On average, de novo assembly with CLC Genomics Workbench yielded three times the number of contigs assembled by GS De novo Assembler; hence, only results produced with the latter program are presented. De novo assembly resulted in a highly variable number of contigs per target gene, with HSC (9 contigs) and CARB (10 contigs) at one extreme and MHC (69 contigs) at the other (Table 3).

Results

BAC library construction

The lake whitefish BAC library consists of a total of 181 050 clones providing 5.2 haploid genome equivalents assuming a genome size of 3 Gb (Table 2, Animal Genome Size Database). According to 117 randomly picked clones, insert sizes range between 36 and 172 Kb, with an estimated average of 92 Kb. These results are comparable to those of Di Palma et al. (2007) for BAC libraries of Metriacanthus zebra (haplochromine cichlid) and Astyanax mexicanus (Mexican tetra), using a similar approach.
as references (Fig. 2). Based on these manual assemblies, it was observed that splits between contigs generally occurred in microsatellites. Complete coding sequence was retrieved for all target genes except for \textit{MHC}.

**Discussion**

Up until recently, it was unclear how next-generation sequencing technologies would perform in large and highly repetitive genomes. Comparison of BAC clone assembly from 454 sequencing data with Sanger sequences revealed that while genic and other single-copy regions are covered at high quality by 454 sequencing, the resolution of repetitive DNA and the generation of full-length draft assemblies are only possible with a Sanger/pyrosequencing hybrid approach (Goldberg et al. 2006; Wicker et al. 2006; Quinn et al. 2008; Steuernagel et al. 2009). Here, \textit{de novo} assembly resulted in a variable number of contigs per target gene. The main reason as to why \textit{MHC} assembled into more contigs compared to the other genes is that the two clones represented different gene copies (F.-O. Gagnon-\textendash\textsc{h}e\textendash\textsc{b}ert, personal communication). Indeed, screening amplicons (\textit{MHC} class II beta exon B1) had 92% identity, which is below the 98%

---

**Table 3** Summary of screening and sequencing results

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Isolated clone insert sizes (kb)*</th>
<th>Number of reads†</th>
<th>Number of contigs (&gt;500 bp)‡</th>
<th>Sum of all contig lengths (bp)§</th>
<th>Number of microsatellites¶</th>
<th>Microsatellites/ Mb**</th>
<th>Length of target contig (bp)††</th>
<th>Mean coverage of target contig‡‡</th>
<th>GenBank accession of target contig††</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARB</td>
<td>80.2/39.4</td>
<td>29 937</td>
<td>10</td>
<td>138 146</td>
<td>32</td>
<td>231.6</td>
<td>9851</td>
<td>90.2</td>
<td>HQ287745</td>
</tr>
<tr>
<td>HSC</td>
<td>88.0/115.2</td>
<td>58 898</td>
<td>9</td>
<td>181 368</td>
<td>26</td>
<td>143.4</td>
<td>67 946</td>
<td>103.2</td>
<td>HQ287746</td>
</tr>
<tr>
<td>MDH</td>
<td>95.7/104.2</td>
<td>19 683</td>
<td>22</td>
<td>140 213</td>
<td>48</td>
<td>342.3</td>
<td>26 662</td>
<td>47.6</td>
<td>HQ287747</td>
</tr>
<tr>
<td>GAPDH</td>
<td>97.0/95.0</td>
<td>55 938</td>
<td>27</td>
<td>170 278</td>
<td>103</td>
<td>604.9</td>
<td>2160</td>
<td>119.6</td>
<td>HQ287748</td>
</tr>
<tr>
<td>MHC</td>
<td>72.7/104.9</td>
<td>13 548</td>
<td>69</td>
<td>196 701</td>
<td>61</td>
<td>310.1</td>
<td>7125/5293</td>
<td>10.5/22</td>
<td>HQ287749/HQ287750</td>
</tr>
</tbody>
</table>

CARB, Carboxylesterase; HSC, Heat shock cognate; MDH, Malate dehydrogenase; GAPDH, Glyceraldehyde-3-phosphate dehydrogenase.

*According to \textit{NotI} digestion.
†Produced with a 454 Genome Sequencer FLX System.
‡Following \textit{de novo} assembly with GS De novo Assembler software (Roche, Basel, Switzerland).
§Including contigs <500 pb.
¶Identified using WebSat (Martins et al. 2009).
**Sum of all contig lengths was used as the denominator.
††Target contig: contains the 1-4 regular PCR amplicon.

---

Fig. 2  Graphical representation of candidate gene assembly. Total length from start to stop codons is given between parentheses next to gene name. Black: exons; white: introns; grey: noncoding DNA (length between parentheses, two \textit{MHC} sequences are represented together); arrow: split between contigs in a microsatellite of unresolved length (total length is an approximation in this situation); ellipsis mark: incomplete coding sequence.
assembly threshold. The second reason is likely to be the smaller amount of data and the resulting mean coverage of 11.6 for contigs from MHC BAC clones. According to BAC clone assembly in barley, the number of contigs increases when coverage decreases below approximately 15 (Wicker et al. 2006). GAPDH had the second highest number of contigs, and this is likely due to the presence of almost twice as many microsatellites as in any other of the four datasets. Indeed, one of the main concerns regarding 454 pyrosequencing is the accuracy of individual reads for repetitive DNA, particularly in the case of monoplymer repeats (Hutchison 2007), and the tendency for repetitive DNA to collapse into single contigs during sequence assembly, leaving unresolved gaps (Wicker et al. 2006). Recent annotation of nine BAC clones spanning approximately 1 Mb of the Atlantic salmon genome revealed eight genes (Quinn et al. 2008). The higher gene density observed here is likely due to the fact that our screening approach was specifically designed to target coding regions.

This study demonstrates the efficiency of combining qPCR BAC library screening and 454 sequencing to achieve high-quality assembly of genomic DNA in a non-model organism. Previous studies have come to the conclusion that whole-genome assembly for complex genomes would necessarily have to combine 454 and Sanger technologies (e.g. Quinn et al. 2008) or an equivalent in terms of read length (Davidson et al. 2010). In this respect, salmonid genomes are particularly challenging owing to their pseudotetraploidy (Allendorf & Thorgaard 1984), which translates into the occurrence of paralogous sequence variants (Hayes et al. 2007; Moen et al. 2008). Moreover, because of the problems incurred by repetitive DNA, a BAC-by-BAC approach or the use of small pools of BAC clones are thought to be optimal solutions (Wicker et al. 2006; Steuernagel et al. 2009). Hence, BAC libraries still play a key role in the current next-generation sequencing era by making large genomes tractable.

The whitefish BAC library is currently being screened for more candidate genes. Further investigation of full-length genomic DNA sequence information that includes exons, introns as well as up- and downstream noncoding regions of candidate genes should allow a better understanding of how these genes are implicated in whitefish species divergence, namely through the study of regulatory regions (e.g. Schulte et al. 1997; Kohn et al. 2008), as many of these genes were identified on the basis of gene expression divergence (reviewed in Bernatchez et al. 2010). Furthermore, the lake whitefish BAC library combined with next-generation sequencing technologies and the upcoming Atlantic salmon genome (Davidson et al. 2010) pave the way to improved assembly and annotation of whitefish BAC sequences as well as comparative genomics among salmonid species, both of which will be valuable resources for the identification of genomic bases for adaptive divergence and reproductive isolation in lake whitefish species pairs. The complete nonarrayed whitefish BAC library and isolated clones for all target genes are available upon request at IBIS (contact the corresponding author).

Acknowledgements

Assistance was received from John MacKay (Arborea project, Université Laval) for qPCR experiments. This research was financially supported by a Natural Sciences and Engineering research Council of Canada (NSERC) and a Fonds québécois de la recherche sur la nature et les technologies (FQRNT) postgraduate scholarships to JJ; a Canadian Institutes of Health Research grant (CIHR) to RCL and a National Sciences and Engineering Council of Canada (NSERC) Discovery grant and Canadian Research Chair to LB.

References


### Data Accessibility

DNA sequences: Genbank accessions HQ287745–HQ287750; NCBI SRA: SRP003484.