MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.02982.x

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 nonarrayed 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

Correspondence: Julie Jeukens, Fax: 1-418-656-7176;
E-mail: julie.jeukens.1@ulaval.ca

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 technologies can be harnessed to reach a much more comprehensive understanding of transcriptomic divergence in a young species pair (Jeukens *et al.* 2010; Renaut *et al.* 2010a). Finally, using SNP markers, genome scans of five whitefish species pairs revealed a predominant role for standing genetic variation in the repeated, independent evolution of dwarf whitefish (Renaut *et al.* 2010b). Nevertheless, while partial cDNA sequences for a number of candidate genes are available (Jeukens *et al.* 2009, 2010; Renaut *et al.* 2010a), full-length genomic DNA sequence information that includes introns as well as up- and downstream noncoding regions of genes remains crucial towards elucidating the respective roles of regulatory and structural mutations in the adaptive divergence of dwarf and normal lake whitefish. It is also essential to uncover the genomic bases of reproductive isolation in lake whitefish species pairs.

Here, we describe the construction, characterization and screening of a BAC library for lake whitefish. 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. While the rainbow trout (*Oncorhynchus mykiss*, Palti *et al.* 2004) and Atlantic salmon (*Salmo salar*, Thorsen *et al.* 2005) BAC libraries were primarily constructed in a context of resource management, the novel lake whitefish library will be key in the ongoing study of candidate genes and genomic regions thought to be implicated in whitefish species divergence.

Materials and methods

High molecular weight (HMW) DNA preparation, electroelution and insert size analysis were performed according to (Osoegawa *et al.* 2001). However, the approach for all other manipulations related to 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 *Hin*dIII Cloning-Ready Vector (Epicentre Biotechnologies, Madison, WI, USA). Construction with pCC1BAC *Eco*RI 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 *Eco*RI 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, FMC BioProducts, Rockland, ME, USA), for a final concentration of 5×10^7 cells/mL. The mixture was immediately transferred to ice-cold plug moulds (Bio-Rad, Hercules, CA, USA) 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 M 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 mM Tris-HCl pH 8, 50 mM EDTA, replaced twice); (iv) 2 × 2 h at 4 °C with gentle mixing in PMSF 0.1 mM (diluted in TE50); and (5) 24 h at 4 °C with gentle mixing in TE50 (replaced twice). Agarose plugs were then kept in 0.5 M EDTA at 4 °C.

Partial digestion of HMW DNA and size selection

Agarose plugs were dialysed in TE (10 mM Tris-HCl pH 8, 1 mM EDTA) for 24 h. Each plug was cut into four equal pieces. Each piece was then placed in a tube containing 500 μL of 1× *Hin*dIII Buffer (NEB buffer 2, New England Biolabs, Pickering, ON, USA) for 1 h at 4 °C. Buffer was replaced and supplemented with 15 U *Hin*dIII 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 0.5× 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: 0.5× 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 *Not*I and submitted to migration on a CHEF apparatus (0.5× TBE buffer, 14 °C, 5 V/cm, 19 h, 5–20 s pulse time). Mid-range PFG Marker (New England Biolabs, Pickering, ON, USA) and lambda DNA *Hin*dIII 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 [1× 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

Gene name *(code)*	Function*	Primer pair 1 (1–2)†	Primer pair 2 (3–4)‡	Amplicon length (bp) (1–4)§	References¶
Liver carboxylesterase 22 precursor *(CARB)*	Detoxification	**AGCTTCTGAGGGAGGAGGAC** AACGGGGAGGAGCTGATACT	TGTGCTGGACTGTTGTGTGA **GATGGCATATTGGGCCAATTT**	377	3, 7, 10
Heat shock cognate 70 kDa protein *(HSC)*	Protein folding	**AAGGTTTCCTCCAAGAACTCACTGG** TGCACTTCTCCAGAATCTTGGTCTT	TTTGTTTGCTACACAACAGGGAACA **TAAAGGCATTGTGACAAAGGCAGAT**	833	4, 5, 6
Malate dehydrogenase *(MDH)*	Energy metabolism	**CTTGGTAGTGGGAAACCCTGCTAAC** GGGATCCATTTAGGATTGGACCTTA	TTCTGCAAACTTTCCCAGGATTTCT **CAGCGTACACACCCATAGACATGAA**	1842	3, 8, 10
Glyceraldehyde-3-phosphate dehydrogenase *(GAPDH)*	Energy metabolism	AGAAAACCGTTGATGGTCCCTCTG** GGTTGTGAACTTGTGATATTATAATTGAGTG	GCCTTACCAATCTGCCAATACACAT **TCAACAATGGGCTTTACTGCTATCT**	725	1, 2, 3, 4, 8, 10
MHC classII beta, exon B1 *(MHC)*	Immunity	**CCGATACTCCTCAAAGGACCTGCA** CTGTGTTCCATGCTTCTGCATTCTT	na†† **CCCTGCTCACCTGTCTTATCCAGTA**	251	9

*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.

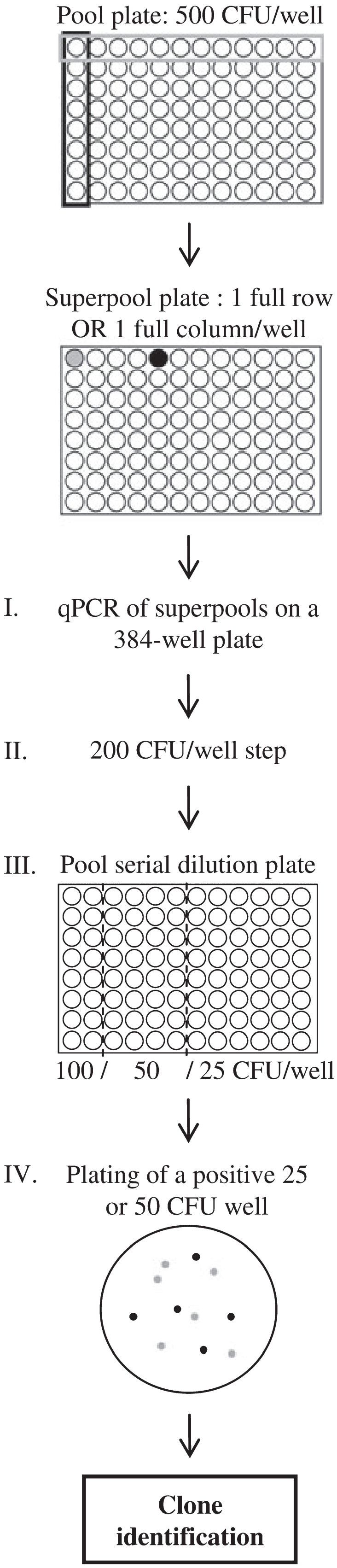
Pool plate: 500 CFU/well
Superpool plate : 1 full row
OR 1 full column/well
I. qPCR of superpools on a 384-well plate
II. 200 CFU/well step
III. Pool serial dilution plate
100 / 50 / 25 CFU/well
IV. Plating of a positive 25 or 50 CFU well
Clone identification

Fig. 1 Schematic representation of the screening approach. CFU are determined by counting white colonies only. After qPCR on superpools (step I), positive pools are identified according to the intersections of positive row and column superpools. For the 200 CFU/well and serial dilutions steps (II and III), qPCR is first performed on column superpools. Then, the positive well is identified by regular PCR. The final step is clone picking and identification (IV).

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 (RiboShredder™ 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 *Hin*dIII 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 Worbench 3.7.1 (CLC Bio, Aarhus, Denmark) was used with the following parameters: trimming databases for the BAC vector and *E. coli* genome, minimum length fraction 0.75, minimum identity 98% and minimum contig length 500 bp.

Contigs were annotated by manual blast search using 5-kb segments of contig consensus sequences to maximize the coverage of blast hits for each contig. First, contig consensus sequences were blasted (megablast) against the complete salmonid nucleotide collection in GenBank. Then, because a large proportion of the salmonid nucleotide collection is not annotated, discontiguous megablast (intended for cross-species comparisons) was performed against the complete nucleotide collection minus salmonid sequences. Microsatellites within contigs were identified using WebSat (default settings, Martins *et al.* 2009).

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 *Metriaclima 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).

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

Table 2 Whitefish BAC library summary

BAC library parameter	Estimated value
Number of recombinant clones*	181 050
Mean insert size (kb)†	92
Proportion of noninsert clones (%)†	6
Net number of genome equivalents‡	5.2

*According to white colony-forming units (CFU)/mL estimation.
†Derived from 117 *Not*I clone digests; weighted mean was used because of the varying number of clones per pool of ligation reactions.
‡Excluding the proportion of noninsert clones; 3 Gb was used as haploid genome size (Animal Genome Size Database).

Table 3 Summary of screening and sequencing results

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

CARB, Carboxylesterase; HSC, Heat shock cognate; MDH, Malate dehydrogenase; GAPDH, Glyceraldehyde-3-phosphate dehydrogenase.
*According to *Not*I digestion.
†Produced with a 454 Genome Sequencer FLX System.
‡Following *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.

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 *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, *de novo* assembly resulted in a variable number of contigs per target gene. The main reason as to why *MHC* assembled into more contigs compared to the other genes is that the two clones represented different gene copies (F.-O. Gagnon-Hébert, personnal communication). Indeed, screening amplicons (MHC class II beta exon B1) had 92% identity, which is below the 98%

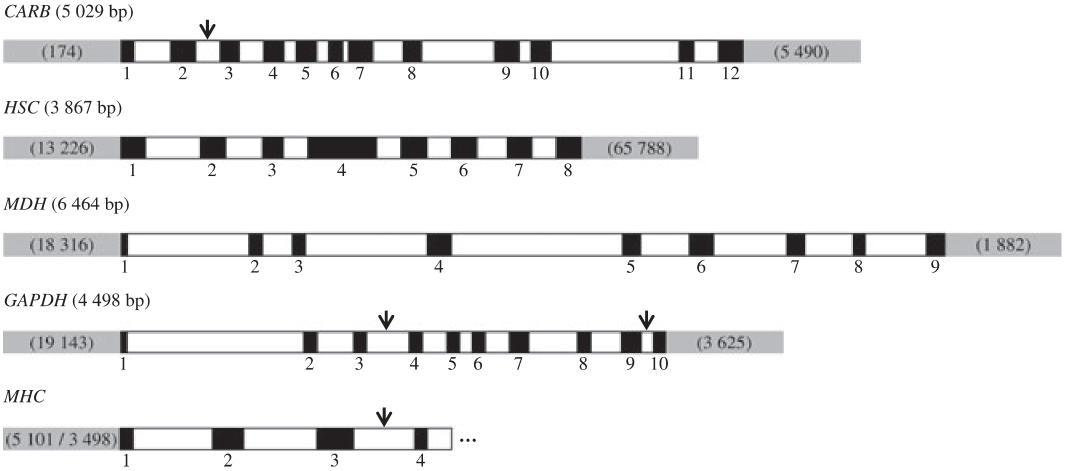

Fig. 2 Graphical representation of candidate gene assembly. Total length from start to stop codons is given between parentheses next to gene name. Black: exons; white: introns; grey: noncoding DNA (length between parentheses, two *MHC* sequences are represented together); arrow: split between contigs 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 monopolymer 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 nonmodel 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 5′ 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

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed. Turner BJ), pp. 1–53. Plenum Press, New York.

Barrett RDH, Rogers SM, Schluter D (2008) Natural selection on a major armor gene in threespine stickleback. *Science*, **322**, 255–257.

Bernatchez L, Dodson JJ (1990) Allopatric origin of sympatric populations of lake whitefish (*Coregonus clupeaformis*) as revealed by mitochondrial DNA restriction analysis. *Evolution*, **44**, 1263–1271.

Bernatchez L, Dodson JJ (1991) Phylogeographic structure in mitochondrial DNA of the lake whitefish (*Coregonus clupeaformis*) and its relation to Pleistocene glaciations. *Evolution*, **45**, 1016–1035.

Bernatchez L, Renaut S, Whiteley AR *et al.* (2010) On the origins of species: insights from the ecological genomics of whitefish. *Philosophical transactions of the Royal Society of London B-Biological Sciences*, **365**, 1783–1800.

Binz T, Largiader C, Muller R, Wedekind C (2001) Sequence diversity of Mhc genes in lake whitefish. *Journal of fish biology*, **58**, 359–373.

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role directional selection in the divergence of sympatric whitefish ecotypes. *Molecular Biology and Evolution*, **21**, 945–956.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates Inc., Sunderland.

Davidson W, Koop B, Jones S *et al.* (2010) Sequencing the genome of the Atlantic salmon (Salmo salar). *Genome Biology*, **11**, 403.

Derome N, Duchesne P, Bernatchez L (2006) Parallelism in gene transcription among sympatric lake whitefish (*Coregonus clupeaformis* Mitchill) ecotypes. *Molecular Ecology*, **15**, 1239–1249.

Derome N, Bougas B, Rogers SM *et al.* (2008) Pervasive sex-linked effects on transcription regulation as revealed by eQTL mapping in lake whitefish species pairs (*Coregonus* sp, *Salmonidae*). *Genetics*, **179**, 1903–1917.

Di Palma F, Kidd C, Borowsky R, Kocher TD (2007) Construction of Bacterial Artificial Chromosome Libraries for The Lake Malawi Cichlid (Metriaclima Zebra), And The Blind Cavefish (Astyanax Mexicanus). *Zebrafish*, **4**, 41–48.

Edmands S (2002) Does parental divergence predict reproductive compatibility? *Trends in Ecology & Evolution*, **17**, 520–527.

Goldberg SMD, Johnson J, Busam D *et al.* (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences*, **103**, 11240–11245.

Hayes B, Laerdahl JK, Lien S *et al.* (2007) An extensive resource of single nucleotide 614 polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture*, **265**, 82–90.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Hutchison CA III (2007) DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, **35**, 6227–6237.

Jeukens J, Bittner D, Knudsen R, Bernatchez L (2009) Candidate genes and adaptive radiation: insights from transcriptional adaptation to the limnetic niche among coregonine fishes (*Coregonus* spp., Salmonidae). *Molecular Biology and Evolution*, **26**, 155–166.

Jeukens J, Renaut S, St-Cyr J, Nolte AW, Bernatchez L (2010) The transcriptomics of sympatric dwarf and normal lake whitefish (Coregonus clupeaformis spp., Salmonidae) divergence as revealed by next-generation sequencing. *Molecular Ecology*, **19**, 5389–5403.

Joron M, Papa R, Beltran M *et al.* (2006) A conserved supergene locus controls colour pattern diversity in Heliconius butterflies. *PLoS Biology*, **4**, 1831–1840.

Kane NC, King MG, Barker MS *et al.* (2009) Comparative Genomic and Population Genetic Analyses Indicate Highly Porous Genomes and High Levels of Gene Flow between Divergent Helianthus Species. *Evolution*, **63**, 2061–2075.

Kohn MH, Shapiro J, Wu CI (2008) Decoupled differentiation of gene expression and coding sequence among Drosophila populations. *Genes & Genetic Systems*, **83**, 265–273.

Lu G, Bernatchez L (1999) A study of fluctuating asymmetry in hybrids of dwarf and normal lake whitefish ecotypes (*Coregonus clupeaformis*) from different glacial races. *Heredity*, **83**, 742–747.

Lu G, Basley DJ, Bernatchez L (2001) Contrasting patterns of mitochondrial DNA and microsatellite introgressive hybridization between lineages of lake whitefish (*Coregonus clupeaformis*); relevance for speciation. *Molecular Ecology*, **10**, 965–985.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Martins W, Lucas D, Neves K, Bertioli D (2009) WebSat–a web software for microsatellite marker development. *Bioinformation*, **3**, 282–283.

Moen T, Hayes B, Baranski M *et al.* (2008) A linkage map of the Atlantic salmon (Salmo salar) based on EST-derived SNP markers. *BMC Genomics*, **9**, 223.

Nolte AW, Renaut S, Bernatchez L (2009) Divergence in gene regulation at young life history stages of whitefish (Coregonus sp.) and the emergence of genomic isolation. *BMC Evolutionary Biology*, **9**, 59.

Osoegawa K, de Jong P, Frengen E, Ioannu PA (2001) Construction of bacterial artificial chromosome (BAC/PAC) libraries. In: *Current Protocols in Molecular Biology* (ed. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA & Struhl K), pp. 5.15.11–15.15.33. Wiley, New York.

Palti Y, Gahr SA, Hansen JD, Rexroad CE III (2004) Characterization of a new BAC library for rainbow trout: evidence for multi-locus duplication. *Animal Genetics*, **35**, 130–133.

Pigeon D, Chouinard A, Bernatchez L (1997) Multiple modes of speciation involved in the parallel evolution of sympatric morphotypes of lake whitefish (*Coregonus clupeaformis*, *Salmonidae*). *Evolution*, **51**, 196–205.

de Queiroz K (2005) Different species problems and their resolution. *Bioessays*, **27**, 1263–1269.

Quinn NL, Levenkova N, Chow W *et al.* (2008) Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics*, **9**, 404.

Renaut S, Nolte AW, Bernatchez L (2009) Gene Expression Divergence and Hybrid Misexpression between Lake Whitefish Species Pairs (Coregonus spp. Salmonidae). *Molecular Biology and Evolution*, **26**, 925–936.

Renaut S, Nolte AW, Bernatchez L (2010a) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. Salmonidae). *Molecular Ecology*, **19**, 115–131.

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2010b) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.). *Molecular Ecology*, **20**, 545–559.

Rise ML, von Schalburg KR, Brown GD *et al.* (2004) Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Research*, **14**, 478–490.

Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Molecular Ecology*, **14**, 351–361.

Rogers SM, Bernatchez L (2006) The genetic basis of intrinsic and extrinsic post-zygotic reproductive isolation jointly promoting speciation in the lake whitefish species complex (*Coregonus clupeaformis*). *Journal of Evolutionary Biology*, **19**, 1979–1994.

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonus* sp *Salmonidae*) species pairs. *Molecular Biology and Evolution*, **24**, 1423–1438.

Rogers SM, Campbell D, Baird SJE, Danzmann RG, Bernatchez L (2001) Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (*Coregonus clupeaformis*, Mitchill). *Genetica*, **111**, 25–41.

Rogers SM, Gagnon V, Bernatchez L (2002) Genetically based phenotype-environment association for swimming behavior in lake whitefish ecotypes (*Coregonus clupeaformis* Mitchill). *Evolution*, **56**, 2322–2329.

von Schalburg K, Rise M, Cooper G *et al.* (2005) Fish and chips: Various methodologies demonstrate utility of a 16,006-gene salmonid microarray. *BMC Genomics*, **6**, 126.

Schemske DW, Bierzychudek P (2007) Spatial differentiation for flower color in the desert annual Linanthus parryae: Was Wright right? *Evolution*, **61**, 2528–2543.

Schulte PM, Gómez-Chiarri M, Powers DA (1997) Structural and functional differences in the promoter and 5′ flanking region of Ldh-B within and between populations of the teleost *Fundulus heteroclitus*. *Genetics*, **145**, 759–769.

St-Cyr J, Derome N, Bernatchez L (2008) The transcriptomics of life-history trade-offs in whitefish species pairs (*Coregonus* sp.). *Molecular Ecology*, **17**, 1850–1870.

Steuernagel B, Taudien S, Gundlach H *et al.* (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics*, **10**, 547.

Storz JF, Wheat CW (2010) Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution*, **64**, 2489–2509.

Storz JF, Sabatino SJ, Hoffmann FG *et al.* (2007) The molecular basis of high-altitude adaptation in deer mice. *PLoS Genetics*, **3**, 448–459.

Thorsen J, Zhu B, Frengen E *et al.* (2005) A highly redundant BAC library of Atlantic salmon (Salmo salar): an important tool for salmon projects. *BMC Genomics*, **6**, 50.

Whiteley AR, Derome N, Rogers SM *et al.* (2008) The Phenomics and Expression Quantitative Trait Locus Mapping of Brain Transcriptomes Regulating Adaptive Divergence in Lake Whitefish Species Pairs (Coregonus sp.). *Genetics*, **180**, 147–164.

Wicker T, Schlagenhauf E, Graner A *et al.* (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, **7**, 275.

Data Accessibility

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

DNA sequence assembly: DRYAD entry doi:10.5061/dryad.8100.