MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11 (Suppl. 1),** 117–122 doi: 10.1111/j.1755-0998.2010.02967.x

SNP DISCOVERY: NEXT GENERATION SEQUENCING

Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout

PAUL A. HOHENLOHE,* STEPHEN J. AMISH,† JULIAN M. CATCHEN,* FRED W. ALLENDORF† and GORDON LUIKART†‡§

*Center for Ecology and Evolutionary Biology, 5289 University of Oregon, Eugene, OR 97403-5289, USA, †Fish and Wildlife Genomics Group, Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA, ‡Flathead Lake Biological Station, University of Montana, Polson, MT 59860, USA, §CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

Abstract

The increased numbers of genetic markers produced by genomic techniques have the potential to both identify hybrid individuals and localize chromosomal regions responding to selection and contributing to introgression. We used restriction-site-associated DNA sequencing to identify a dense set of candidate SNP loci with fixed allelic differences between introduced rainbow trout (*Oncorhynchus mykiss*) and native westslope cutthroat trout (*Oncorhynchus clarkii lewisi*). We distinguished candidate SNPs from homeologs (paralogs resulting from whole-genome duplication) by detecting excessively high observed heterozygosity and deviations from Hardy–Weinberg proportions. We identified 2923 candidate species-specific SNPs from a single Illumina sequencing lane containing 24 barcode-labelled individuals. Published sequence data and ongoing genome sequencing of rainbow trout will allow physical mapping of SNP loci for genome-wide scans and will also provide flanking sequence for design of qPCR-based TaqMan® assays for high-throughput, low-cost hybrid identification using a subset of 50–100 loci. This study demonstrates that it is now feasible to identify thousands of informative SNPs in nonmodel species quickly and at reasonable cost, even if no prior genomic information is available.

Keywords: conservation biology, genome duplication, homeolog, invasive species, population genomics, threatened salmonids

Received 13 October 2010 ; revision received 16 November 2010; accepted 17 November 2010

Introduction

Hybrid zones and invasive species provide some of the richest natural experiments in evolutionary biology. Hybridization and invasion are also important in conservation biology and are among the most serious threats to the persistence of many native species (Allendorf & Luikart 2007; Sax *et al*. 2007). Interspecific hybridization can break up co-adapted gene complexes, disrupt local adaptation and lead to genomic extinction (Rhymer & Simberloff 1996; Allendorf *et al*. 2001).

Rainbow trout (RBT, *Oncorhynchus mykiss*), the most widely introduced salmonid fish in the world (Halverson 2010), produce fertile offspring when crossed with cutthroat trout (*Oncorhynchus clarkii*). Introgression often continues until a hybrid swarm is formed and all native cutthroat genomes are lost (Allendorf & Leary 1988). Such introgression poses a serious threat to all remaining subspecies of cutthroat trout in western North America (Trotter 2008). Westslope cutthroat trout (WCT, *O. c. lewisi*) are native to the interior of north-western North America (Fig. 1a). Hybridization with non-native RBT is the leading factor contributing to the decline of genetically pure WCT populations throughout their range, and nonhybridized populations now persist in <10% of their historic range (Shepard *et al*. 2005).

Currently, only 10–15 diagnostic loci are commonly used to study introgression in these taxa, which is not sufficient for individual-level testing for admixture (Allendorf *et al*. 2010). The development of a dense genome-wide set of single-nucleotide polymorphism (SNP) markers in RBT and WCT would allow localization of chromosomal regions under selection and detection of RBT alleles rapidly spreading across WCT populations (e.g. Fitzpatrick *et al*. 2010). A smaller set of 50–100 loci to

Correspondence: Paul Hohenlohe, Fax: 541 346 2364;
E-mail: hohenlo@uoregon.edu

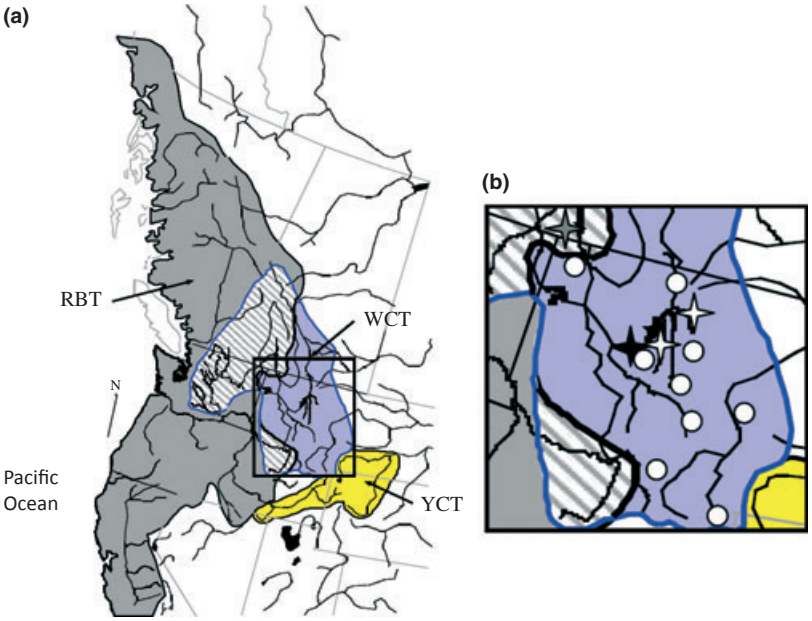

Fig. 1 Range and sampling locations of rainbow and cutthroat trout. (a) Historic range of rainbow (RBT), westslope cutthroat (WCT) and Yellowstone cutthroat (YCT) trout in western North America. (b) Enlargement of box in (a) to show sampling locations of origin for WCT (dots) and RBT (stars; grey for redband, white for introduced coastal RBT and black for hatchery RBT). Three of the locations for coastal RBT are in close proximity and appear as a single star.

be used in high-throughput qPCR assays would allow identification of individuals with no introgression, at high statistical power and low cost, for use in brood stock development and translocations with only nonhybridized ('pure') individuals.

A particular challenge in developing SNP markers in salmonids is a recent (~25–100 MYA) genome duplication, ancestral to all salmonids, including both WCT and RBT (Allendorf & Danzmann 1997; Koop *et al.* 2008). It can be difficult to distinguish true SNPs from duplicate regions, or homeologs, that differ by only one or a few nucleotide positions. Here, we use restriction-site-associated DNA (RAD) sequencing (Baird *et al.* 2008), a next-generation sequencing technique that generates short sequence reads at thousands of regions adjacent to restriction endonuclease recognition sites across the genomes of multiple individuals. We currently lack a reference genome sequence against which to align the sequence reads to identify homeologs. However, the number of nucleotide positions genotyped across multiple individuals, combined with barcoding of individual samples, allows us to screen through thousands of candidate loci and to distinguish candidate SNPs from homeologs on the basis of observed patterns of heterozygosity, prior to any further SNP validation steps. For example, homeologous regions that are fixed for one or more nucleotide differences should appear as heterozygotes in most individuals genotyped. Putative loci showing an excess of heterozygosity (e.g. proportion of heterozygotes >0.5 for a bi-allelic locus, or substantial deviation from Hardy–Weinberg proportions) can be excluded from further analysis.

We have two goals in this study: (i) to identify a set of (>1000) SNP markers that are fixed between RBT and WCT to use in further RAD sequencing studies to examine the genomic patterns of selection and introgression between RBT and WCT and (ii) to use a subset (50–100) of these markers to design qPCR-based assays (e.g. SNP chips) that can be applied at relatively low cost to a large number of samples, including degraded and historic DNA, to distinguish hybrids and quantify genome-wide average levels of introgression in multiple populations.

Methods

We collected genomic DNA from a total of 24 individuals from 15 populations in interior north-west North America (Fig. 1b): 11 WCT, 12 coastal RBT (including 7 from a hatchery population and 5 from introduced populations) and 1 inland RBT (redband trout, *Oncorhynchus mykiss gairdnerii*). We prepared RAD libraries according to Etter *et al.* (in press). Each individual sample was barcoded with a unique six-nucleotide sequence, and all samples were run in a single lane on an Illumina Genome Analyzer II.

We filtered out low-quality reads and those that lacked a correct barcode (see Emerson *et al.* 2010 for further details). We pooled the reads from all individuals and aligned them against each other, allowing a maximum of two nucleotide mismatches among reads at a putative RAD tag locus. We then applied a maximum-likelihood algorithm (Hohenlohe *et al.* 2010) to estimate the diploid genotype for each individual at each nucleotide position.

This genotyping method is designed to account for the sampling and sequencing error inherent in RAD sequencing, assigning a genotype only if a likelihood

ratio test is significant at a level of $\alpha = 0.05$. This implicitly results in a threshold for depth of sequencing coverage, so that some loci in some individuals will not be assigned a genotype simply because of the sampling variance among alleles and loci in a sequenced RAD library. For example, if all reads are identical at a nucleotide site (i.e. no apparent sequencing errors), a minimum of three reads is required to call an individual a homozygote; if the reads are evenly split between two alternative nucleotides, a minimum of two high-quality reads of each allele is required to call an individual a heterozygote (Hohenlohe *et al.* 2010). Sequencing error and sampling variance across alternative alleles increase these implicit thresholds, so that in practice read depth needs to be substantially higher to call most genotypes (see below). Software to conduct these steps of the analysis on high-throughput, short-read sequence data is available online at http://creskolab.uoregon.edu/stacks.

We then applied a number of filters to identify candidate diagnostic SNPs for detecting hybridization and other population-level studies. We focused on loci genotyped for at least 20 of the 24 individuals assayed (i.e. missing up to four individuals, thus a minimum of seven WCT or nine RBT) and on RAD tags containing only one bi-allelic, putative SNP within the 48 bp of potentially variable sequence. The presence of no more than one bi-allelic SNP ensures PCR primer and probe fidelity for qPCR assay design. It also helps screen out homeologs, as many duplicate regions are likely to have diverged at more than one nucleotide position within the 54-bp RAD tag sequence (Seeb *et al.* 2011).

At this point, we excluded additional putative SNP loci as homeologs by examining the pattern of observed heterozygosity within each species. At each putative locus, we calculated expected heterozygosity as $H_{exp} = 1-\sum n_i(n_i - 1)/n(n - 1)$, where n_i is the count of allele i in the sample and $n = \sum n_i$, and observed heterozygosity H_{obs} as the proportion of individuals that appear heterozygous at the locus. We also calculated $F_{IS} = 1 - (H_{obs}/H_{exp})$, which provides a measure of the deviation of genotype frequencies (i.e. observed heterozygosity) from Hardy–Weinberg proportions. We applied two stringent filters based on these measures: putative loci with $H_{obs} > 0.5$ within either species and those with $F_{IS} < 0.0$ within either species were eliminated.

Results

RAD sequencing generated a total of just over 40 million single-end, 60-bp reads, prior to any quality filtering. Sequence reads have been deposited in the NCBI Sequence Read Archive (Accession no. SRA026047.1). Each read provided 54 bp of genomic sequence after trimming the barcode, of which 48 bp may contain SNPs (the remaining 6 bp compose the partial restriction site, which is constant across all reads). Reads were filtered for overall quality and presence of a barcode and then aligned with each other to create a catalogue of 98 190 putative loci, or RAD tags. A total of 20.9 million reads contributed to this alignment, for a mean depth of 8.85× per individual per tag (one WCT individual had a mean coverage of 0.7× because of low DNA concentration, while the other individuals ranged from 2.6× to 20.7×; variation among individuals is often observed in a pooled RAD sequencing library).

Further filtering identified candidate SNPs (Table 1). A total of 9844 putative RAD tags had exactly one bi-allelic putative SNP and were genotyped in at least 20 of the 24 individuals. For these putative tags, the mean depth of coverage per individual was 10.9×. This list included putative SNPs that were likely homeologs, based on excess observed heterozygosity (Fig. 2). We removed 540 tags because $H_{obs} > 0.5$ in both species (242) or in one of the two species (298), plus an additional 633 tags at which $F_{IS} < 0.0$ in both species (16) or in one of the two species (617). These screens produced a total of 8671 candidate SNP markers (Table 1). Of these, 2923 were fixed within each species, providing candidate species-specific SNPs to distinguish hybrids and estimate levels of introgression (Table 1).

Design of qPCR TaqMan® assays for a subset of these loci requires additional flanking sequence of approximately 40 bp on each side of the SNP. We searched for the species-specific RAD tag loci (using both alleles of each locus) in a database of longer sequences produced by 454 sequencing of a reduced representation genomic library in RBT (Sánchez *et al.* 2009), using the alignment software bowtie (Langmead *et al.* 2009). The RBT allele for 39 candidate diagnostic RAD tag loci aligned to one

Table 1 Counts of putative loci after each step of filtering (each row incorporates all filters above), and final counts of candidate SNPs after filtering

Category	Count	% of total
Filtering steps		
(1) Total putative RAD tag loci	98 190	100.0
(2) ≥1 SNP w/in RAD tag	40 592	41.3
(3) Genotyped in ≥20 samples	19 120	19.5
(4) Exactly 1 SNP in RAD tag	9886	10.1
(5) Bi-allelic SNP	9844	10.0
(6) Observed heterozygosity ≤0.5 in each species	9304	9.5
(7) $F_{IS} \geq 0.0$ in each species	8671	8.8
Candidate SNPs		
Total	8671	100.0
Fixed between species	2923	33.7
Polymorphic within RBT	4002	46.2
Polymorphic within WCT	2007	23.1

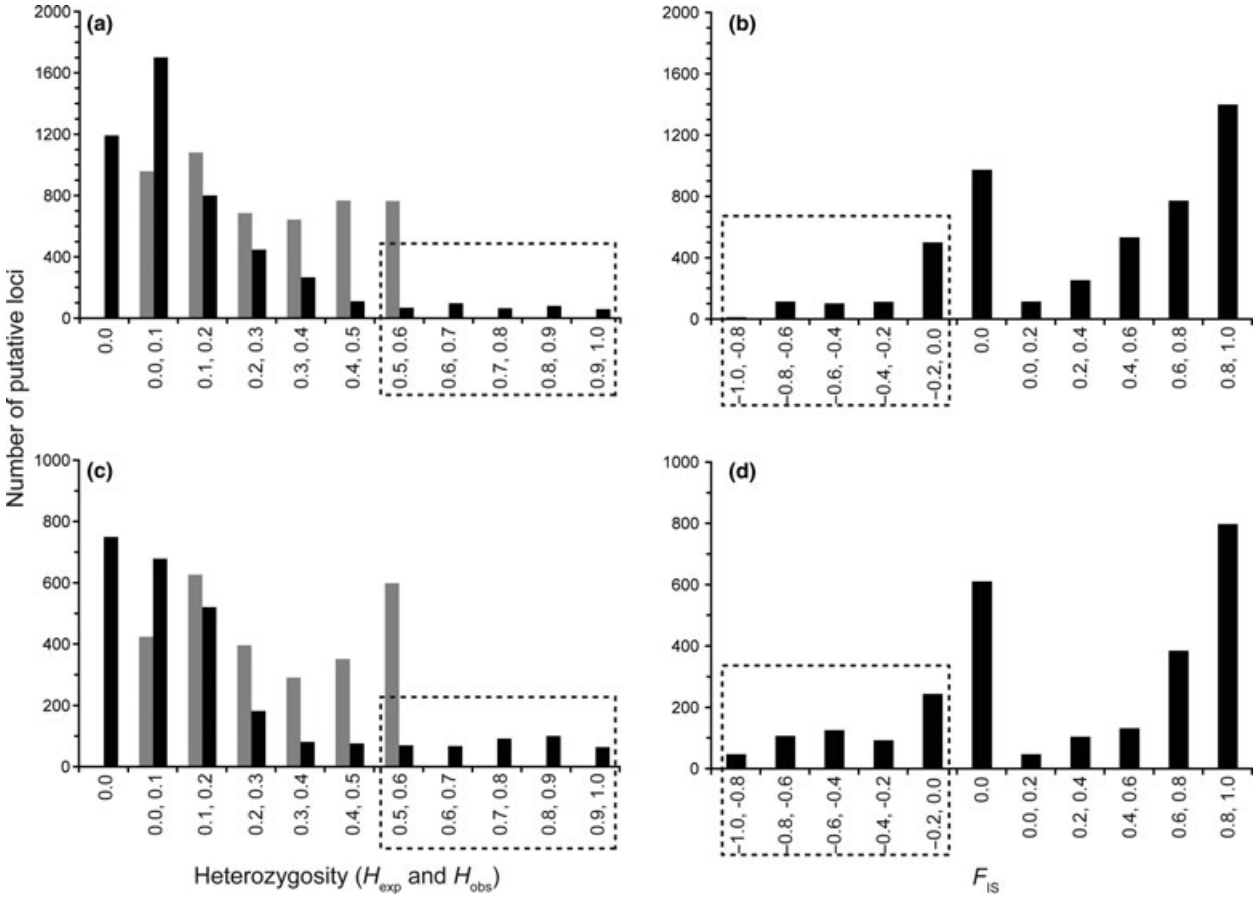

Fig. 2 Frequency histograms of population genetic statistics for putative SNP loci. Counts include only those loci that passed filtering step 5 in Table 1 and that are polymorphic within each species (4915 in RBT and 2695 in WCT). (a,b) RBT. (c,d) WCT. (a,c) Expected (grey) and observed (black) heterozygosity at each putative SNP. (b,d) F_{IS} at each putative SNP. Dashed boxes enclose likely homeologs that were excluded in steps 6 and 7 of filtering (Table 1). Numbers along the *X*-axis indicate lower and upper boundaries of each bin, but note that loci with either H_{obs} or F_{IS} exactly equal to 0.0 are grouped into a separate bin in each plot.

or more sequences from Sánchez *et al.* (2009) with no nucleotide mismatches in the 54-bp RAD tag sequence. In addition, four RBT alleles and one WCT allele aligned to these sequences with just one mismatch, and nine RBT alleles and one WCT allele aligned with two mismatches. In total, this provides flanking sequence to design primers and validate qPCR assays for 54 of our candidate diagnostic SNPs. Genomic sequence contigs from a genome sequencing project in RBT using a BAC library will also be available in 2011 (M. Miller, U. Oregon, unpublished data), providing flanking sequence for many more of our diagnostic SNPs. We also found a large number of candidate SNPs within each species useful for future studies in population genetics and landscape genomics (Schwartz *et al.* 2009) (Table 1).

Discussion

Many populations and species of salmonids are of conservation and management concern (Waples & Hendry 2008). Development of genetic tools in these fish has been hampered in part by their relatively large genome containing substantial duplicate regions and by the lack of a reference genome sequence. Our application of RAD sequencing here alleviates some of these problems and facilitates the development of SNP markers. First, RAD sequencing identifies such a large number of putative loci that stringent screens can be applied to narrow the list of candidate SNPs. Second, we were able to barcode individual samples, run all samples together in a single high-throughput Illumina lane and obtain sufficient depth of sequencing coverage to identify individual-level diploid genotypes (as in Hohenlohe *et al.* 2010). This allowed us to screen out homeologs on the basis of observed heterozygosity among individuals. Such an approach is not possible in a pooled population sample, which can be used to estimate population-level allele frequency but not observed heterozygosity (Sánchez *et al.* 2009).

The presence of homeologs is apparent in our data (Fig. 2). In both species, a significant number of putative

loci have H_{obs} = 0.5, while the maximum H_{exp} is 0.5 for bi-allelic loci in a large sample of individuals. In WCT, the distribution of H_{obs} is bimodal, with a second peak at H_{obs} =~0.8 (Fig. 2c). We do not expect loci to be in Hardy–Weinberg proportions even within either of the two species, because individuals were sampled from multiple populations. However, this should produce a deficit of heterozygotes ($F_{IS} > 0$), rather than an excess, because of the Wahlund effect. Consistent with this expectation, the majority of loci in both species show $F_{IS} > 0$ and low values of H_{obs}, including many loci at which no individuals were heterozygous ($H_{obs} = 0$) despite the detection of two alleles within the species (Fig. 2a,c).

In contrast, the observed excess of heterozygotes ($H_{obs} > 0.5$ or $F_{IS} < 0$) likely indicates homeologous pairs of loci that were incorrectly combined into a single putative locus during initial construction of the RAD tag catalogue. Other processes could also produce an excess of heterozygotes, such as selection for heterozygotes. However, this selection would have to be quite strong to produce H_{obs} values approaching 1.0, and loci under such strong selection would make poor candidates for general-purpose population genetic markers. In addition, of the 540 putative SNPs that we removed because of high observed heterozygosity, 45% had $H_{obs} > 0.5$ in both species. This is consistent with homeologous regions that diverged prior to the split between RBT and WCT, approximately 2 Ma (Allendorf & Leary 1988). Finally, our primary goal in this study was to identify SNP markers fixed between the species. This provided a further conservative screen for homeologs by eliminating candidate SNPs with any observed heterozygosity within either species.

Several factors can increase confidence in the discrimination of true single-locus SNPs from homeologs using observed heterozygosity and deviation from Hardy–Weinberg expectations. For example, increasing the number of individuals sampled would provide a better estimate of genotype and allele frequencies. Increasing the depth of coverage, and reducing variance in coverage across individuals, would improve statistical confidence in assigning diploid genotypes to each individual (Hohenlohe *et al.* 2010). Increased coverage could also allow one to use the expected double number of sequence reads from homeologs compared to true SNPs to distinguish the two. However, there is substantial variance in number of reads across loci and across individuals, and we did not attempt to use this approach here. Variance in read depth among individuals can be minimized, but not eliminated, by making every effort to use equal DNA quality and quantity among samples in a pooled sequencing run (Etter *et al.* in press). In a nonmodel species, a rough expectation of depth of coverage can be calculated using an estimate of genome size and the expected frequency of restriction cut sites given a particular restriction enzyme. For the maximum-likelihood genotyping method used here in samples from an outbred population, a mean coverage depth of 8–10× should allow genotyping at most sites across most individuals.

Applying next-generation sequencing techniques as an initial screen to identify SNPs, followed by more traditional genotyping techniques across larger samples of individuals, is likely to become commonplace in conservation genetics (Allendorf *et al.* 2010). In this study, we used single-end Illumina sequencing and have taken advantage of existing published data as well as ongoing work to gather flanking sequence for qPCR assay design. One alternative is to conduct paired-end RAD sequencing for marker development, in which overlapping paired-end sequences (e.g. 100 bp per read) at sufficient depth of coverage can provide much longer contigs (e.g. 300–400 bp) containing candidate SNPs and sufficient flanking sequence for direct PCR primer development (J. Davey, U. Edinburgh, unpublished data). Other techniques, such as 454 sequencing of RAD tags, could also be used to generate longer flanking sequence around putative diagnostic SNPs. From this point, further research is needed to validate candidate SNPs through re-sequencing or direct SNP genotyping. For example, expected patterns of Mendelian inheritance can be tested from pedigree samples.

RAD sequencing can be limiting in some respects: it requires at least 1 μg of high-quality genomic DNA per sample (Etter *et al.* in press), and like any genomic approach, bioinformatic analysis comprises a substantial portion of the work. Nonetheless, if sufficient DNA samples can be gathered for a relatively small number of individuals and used in the initial marker-development stage in as little as a single Illumina sequencing lane, as we have performed here, other techniques that require smaller amounts of low-quality DNA can be applied to a larger sample of individuals. Bioinformatic tools continue to be developed for RAD and similar genomic data types and are available online (e.g., http://creskolab.uoregon.edu/stacks). Our results suggest that it is now feasible to identify thousands of informative SNPs in nonmodel species for a reasonable price even if no prior genomic information is available.

Acknowledgements

We thank R. Leary (Montana Fish, Wildlife and Parks), C. Muhlfeld (US Geological Survey) for samples and advice, and W. Cresko and M. Miller (both U. Oregon) for helpful comments on this research. PAH received support from National Science Foundation grants IOS-0642264 and DEB-0919090 and National Institutes of Health grant R24GM079486-01A1 to W.A. Cresko. JMC

received support from NIH NRSA Ruth L. Kirschstein postdoctoral fellowship 1F32GM095213-01 and from NIH grant R01RR020833 to J. Postlethwait. GL and FWA were partially supported by the NSF grant DEB-074218. GL also was funded by the Walton Family Foundation and Portuguese science foundation grants (PTDC/BIA-BDE/65625/2006; and PTDC/CVT/69438/2006), and NSF grant (DEB DEB-1067613).

Conflict of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Allendorf FW, Danzmann RG (1997) Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. *Genetics*, **145**, 1083–1092.

Allendorf RF, Leary RF (1988) Conservation and distribution of genetic variation in a polytypic species: the cutthroat trout. *Conservation Biology*, **2**, 170–184.

Allendorf FW, Luikart G (2007) *Conservation and the Genetics of Populations*. Blackwell Publishers, Oxford, UK.

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology and Evolution*, **16**, 613–622.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics*, **11**, 697–709.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving post-glacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA*, **107**, 16196–16200.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (in press) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds Orgogozo V, Rockman MV), Humana Press, New York.

Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB (2010) Rapid spread of invasive genes into a threatened native species. *Proceedings of the National Academy of Sciences USA*, **107**, 3606–3610.

Halverson A (2010) *An Entirely Synthetic Fish: How Rainbow Trout Beguiled America and Overran the World*. Yale University Press, New Haven, CT.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomic analysis of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Koop BF, von Schalburg KR, Leong J *et al.* (2008) A salmonid EST genomic study: genes, duplications, phylogeny and microarrays. *BMC Genomics*, **9**, 545.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Rhymer JM, Simberloff D (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.

Sánchez CC, Smith TPL, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Sax DF, Stachowicz JJ, Brown JH *et al.* (2007) Ecological and evolutionary insights from species invasions. *Trends in Ecology and Evolution*, **22**, 465–471.

Schwartz MK, Luikart G, McKelvey KS, Cushman S (2009) Landscape genomics: a brief perspective. Chapter 19 In: *Spatial Complexity, Informatics and Animal Conservation* (eds Cushman SA, Huettman F), pp. 165–174. Springer, Tokyo.

Seeb JE, Pascal CE, Grau ED *et al.* (2011) Transcriptome sequencing and high-resolution melt analysis advance SNP discovery in duplicated salmonids. *Molecular Ecology Resources*, **11**, this issue.

Shepard BB, May BE, Urie W (2005) Status and conservation of westslope cutthroat within the western United States. *North American Journal of Fisheries Management*, **25**, 1426–1440.

Trotter P (2008) *Cutthroat: Native Trout of the West*, 2nd edn. University of California Press, Berkeley, CA.

Waples RS, Hendry AP (2008) Special issue: evolutionary perspectives on salmonid conservation and management. *Evolutionary Applications*, **1**, 183–188.