MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 3002–3013 doi: 10.1111/mec.12239

Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing

PAUL A. HOHENLOHE,* MITCH D. DAY,* STEPHEN J. AMISH,† MICHAEL R. MILLER,‡ NICK KAMPS-HUGHES,‡ MATTHEW C. BOYER,§ CLINT C. MUHLFELD,¶** FRED W. ALLENDORF,† ERIC A. JOHNSON‡ and GORDON LUIKART**

*Department of Biological Sciences, Institute of Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844-3051, USA, †Fish and Wildlife Genomics Group, Division of Biological Sciences, University of Montana, Missoula, MT 59812, USA, ‡Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA, §Montana Fish, Wildlife and Parks, Kalispell, MT 59901, USA, ¶U.S. Geological Survey, Northern Rocky Mountain Science Center, Glacier National Park, West Glacier, MT 59936, USA, **Flathead Lake Biological Station, Fish and Wildlife Genomics Group, Division of Biological Sciences, University of Montana, Polson, MT 59860, USA*

Abstract

Rapid and inexpensive methods for genomewide single nucleotide polymorphism (SNP) discovery and genotyping are urgently needed for population management and conservation. In hybridized populations, genomic techniques that can identify and genotype thousands of species-diagnostic markers would allow precise estimates of population- and individual-level admixture as well as identification of 'super invasive' alleles, which show elevated rates of introgression above the genomewide background (likely due to natural selection). Techniques like restriction-site-associated DNA (RAD) sequencing can discover and genotype large numbers of SNPs, but they have been limited by the length of continuous sequence data they produce with Illumina short-read sequencing. We present a novel approach, overlapping paired-end RAD sequencing, to generate RAD contigs of >300–400 bp. These contigs provide sufficient flanking sequence for design of high-throughput SNP genotyping arrays and strict filtering to identify duplicate paralogous loci. We applied this approach in five populations of native westslope cutthroat trout that previously showed varying (low) levels of admixture from introduced rainbow trout (RBT). We produced 77 141 RAD contigs and used these data to filter and genotype 3180 previously identified species-diagnostic SNP loci. Our population-level and individual-level estimates of admixture were generally consistent with previous microsatellite-based estimates from the same individuals. However, we observed slightly lower admixture estimates from genomewide markers, which might result from natural selection against certain genome regions, different genomic locations for microsatellites *vs.* RAD-derived SNPs and/or sampling error from the small number of microsatellite loci (n = 7). We also identified candidate adaptive super invasive alleles from RBT that had excessively high admixture proportions in hybridized cutthroat trout populations.

Keywords: adaptive introgression, conservation genomics, hybridization, invasive species, natural selection, next-generation sequencing, salmonids, super invasive genes

Received 31 July 2012; revision received 5 December 2012; accepted 11 December 2012

Introduction

Hybridization between native and introduced taxa is an increasing concern for conservation and legal asse ssments of threatened species (Allendorf *et al.* 2001).

Correspondence: Paul A. Hohenlohe, Fax: 208 885 7905; E-mail: hohenlohe@uidaho.edu

Hybridization can reduce fitness through outbreeding depression (Muhlfeld *et al.* 2009a), cause genomic extinction (Allendorf *et al.* 2001) and destroy important genetic and ecological adaptations (Muhlfeld *et al.* 2009b; Kelly *et al.* 2010). The loci most responsible for the genetic effects of hybridization may be outliers in their degree of introgression because of natural selection in admixed populations ('super invasive alleles'; Gompert & Buerkle 2009; Fitzpatrick *et al.* 2010; Teeter *et al.* 2010; Miller *et al.* 2012). As a result, estimates of admixture averaged across loci at the individual or population level may miss important genetic factors in conservation and management of native taxa. Current high-throughput sequencing techniques now allow genome scans for invasive alleles in natural populations of nonmodel species.

Anthropogenic hybridization is especially widespread in freshwater fishes due to decades of fish translocations and hatchery supplementation of wild populations. Rainbow trout (RBT, *Onchorhynchus mykiss*) is the most widely translocated and problematic invasive fish worldwide (Halverson 2010). RBT hybridize with cutthroat trout (*O. clarkii*), including the subspecies westslope cutthroat trout (WCT, *O. c. lewisi*). WCT is the most widely distributed of 12 extant cutthroat subspecies, and hybridization is the leading threat to persistence of genetically pure WCT populations (Shepard *et al.* 2005).

Management of WCT populations would benefit from detection of hybridization and introgression at low levels and from the ability to precisely estimate individual-level admixture proportion. Previous work has used microsatellites and other loci to assess levels of admixture from RBT into native WCT populations (Hitt *et al.* 2003; Boyer *et al.* 2008; Muhlfeld *et al.* 2009a,c). Muhlfeld *et al.* (2009c) found that levels of RBT admixture were negatively related to distance from the source of RBT hybridization (Abbot Creek; see Fig. 1) and positively related to mean summer water temperature, suggesting potential for the existence of RBT alleles that are adaptive to warm water temperatures (Perry *et al.* 2001; Narum *et al.* 2010). However, the low number of diagnostic markers available with microsatellites typically allows precise admixture estimates only at the population level, not at the individual or genome-scan level.

Single nucleotide polymorphisms (SNPs) are ideal markers for hybridization assessment and monitoring because hundreds of SNPs can be rapidly, reliably and cheaply genotyped using new genotyping platforms (Morin *et al.* 2004; Seeb *et al.* 2009, 2011a; Angeloni *et al.* 2011; Twyford & Ennos 2012). Much recent effort has been committed to assembling a set of diagnostic SNP loci for RBT and WCT (Finger *et al.* 2009; McGlauflin *et al.* 2010; Harwood & Phillips 2011; Kalinowski *et al.* 2011; Amish *et al.* 2012; Campbell *et al.* 2012; Pritchard *et al.* 2012).

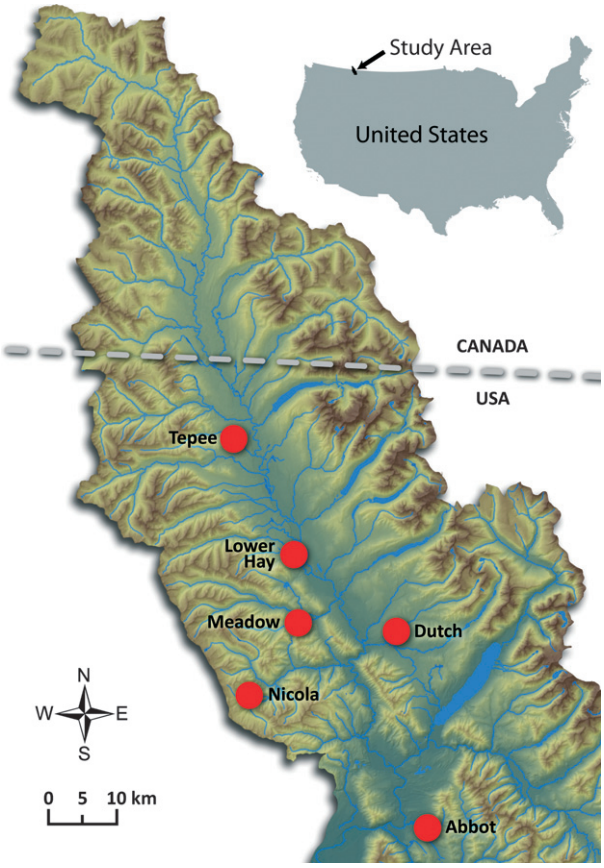

Fig. 1 Map of the North Fork Flathead River study area, showing the five admixed westslope cutthroat trout populations examined here plus the initial source of introduced rainbow trout individuals (Abbot Creek; see Boyer *et al.* 2008; and Muhlfeld *et al.* 2009c for more information on these populations).

A high density of markers across the genome promises individual-level estimates of admixture proportion, as well as detection of super invasive alleles. However, SNP discovery in salmonid fish is especially challenging due to a recent genome duplication event, making it difficult to distinguish true SNPs from fixed sequence differences between homeologous duplicate chromosomal regions (Allendorf & Danzmann 1997; Everett *et al.* 2011; Seeb *et al.* 2011b) as well as more typical tandem-duplicated paralogous regions. One way to filter out both paralogs and homeologs is to gather more sequence data around candidate SNP markers to resolve between next-generation sequence reads that come from one locus *vs.* two different loci.

We previously used restriction-site-associated DNA (RAD) sequencing (Baird *et al.* 2008) to identify several thousand WCT diagnostic SNPs (Hohenlohe *et al.* 2011). Those candidate diagnostic markers have shown a high rate of subsequent validation in microfluidic PCR-based genotyping assays (Amish *et al.* 2012). However, primer

design for those genotyping assays required >50 bp of flanking sequence on each side of each SNP, which we obtained from previously published sequence data, reducing the number of candidate markers for which assays could be designed (Amish *et al.* 2012). In addition, our ability to distinguish duplicate sequence based on the flanking sequence was limited to the 54 bp single-end Illumina read length in that study. The approach we present here can be used to simultaneously identify and genotype SNP markers, as well as gather substantial flanking sequence, in a single RAD sequencing experiment. The amount of flanking sequence is more than sufficient for primer design and also allows better discrimination of paralogous loci.

Restriction-site-associated DNA sequencing is one of a family of genomic approaches that provide sequence data adjacent to restriction enzyme recognition sites (Davey *et al.* 2011). The primary difference between RAD and related techniques is that RAD incorporates a random shearing step in library preparation. As a result, while the forward reads are anchored at the restriction site, the reverse reads produced by paired-end Illumina sequencing of RAD libraries are staggered over a local genomic region (of several hundred base pairs). These staggered paired-end reads can be assembled into a 'mini-contig', a continuous stretch of genomic sequence that is longer than each individual read and potentially up to 1 kb (Baxter *et al.* 2011; Etter *et al.* 2011a; Willing *et al.* 2011; Etter & Johnson 2012). Here, we designed our RAD libraries so that a substantial fraction of DNA fragments would produce overlapping paired-end reads, allowing assembly of contigs containing both the forward and reverse reads of each pair. These 'RAD contigs' are anchored at one end by the restriction enzyme recognition site and contain several hundred base pairs of continuous genomic sequence data across dozens of individuals.

The goals of this study were to: (i) assemble a large set of RAD contigs from a sample of low-admixture WCT populations; (ii) provide flanking sequence for finer filtering of candidate diagnostic SNP markers between RBT and WCT; (iii) genotype filtered diagnostic SNPs across five WCT populations to assess the ability of RAD sequencing compared with microsatellites to provide precise individual-level estimates of admixture; and (iv) identify outlier loci exhibiting the signature of super invasive alleles.

Methods

Study system

We focus on WCT populations in tributaries to the North Fork of the Flathead River in northwestern Montana (Fig. 1). The North Fork Flathead River originates in Canada and forms the western border of Glacier National Park before joining the main-stem Flathead River, which flows into Flathead Lake. The presence of hybridization and RBT admixture was previously estimated in several populations using seven diagnostic microsatellite loci (Boyer *et al.* 2008; Muhlfeld *et al.* 2009c).

Here, we use five of these populations (Meadow, Nicola, Dutch, Lower Hay and Tepee) for which estimates of the mean population-level admixture based on microsatellite loci ranged from 1.3% to 13.0% (see Boyer *et al.* 2008; Muhlfeld *et al.* 2009c for further information on these populations). We chose populations without F1 hybrids as identified in previous studies with the goal of using later-generation admixed populations to detect specific loci with elevated levels of introgression. We used preserved DNA samples, collected from 18 to 22 individuals in each population during 2003 to 2004 for the study by Boyer *et al.* (2008), to allow individual-level comparisons between SNP-based and microsatellite-based admixture estimates. We selected individuals across the range of admixture proportions previously estimated within each population.

RAD sequencing

We prepared RAD sequencing libraries for 97 samples from the five WCT populations described previously, following the protocol of Etter *et al.* (2011b). The RAD protocol produces libraries of genomic fragments bounded on one end by a restriction enzyme cut site (therefore common across individuals), with the other end randomly sheared. Typically, fragments in RAD libraries are size selected simply to optimize the efficiency of the Illumina sequencing process. Here, we used the restriction enzyme SbfI and 6-nucleotide barcoded adaptors differing from each other by at least three nucleotides to identify individuals. We modified the standard protocol to target DNA fragments of 330–400 bp during gel size selection, so that the size of genomic DNA inserts targeted the range 200–270 bp, to produce overlapping paired-end reads for a large proportion of sequenced fragments (Fig. 2). We sequenced the RAD libraries in portions of two lanes (grouped with other RAD sequencing experiments) on an Illumina HiSeq sequencer at the University of Oregon, producing 153-bp paired-end reads.

We processed the sequence data and grouped the read pairs from all individuals into RAD loci using several modules from the STACKS software package, version 0.998 (Catchen *et al.* 2011). First, using the STACKS program process_radtags.pl, we sorted read pairs by barcode, filtered for read quality and removed any pairs in which the forward read did not contain both a

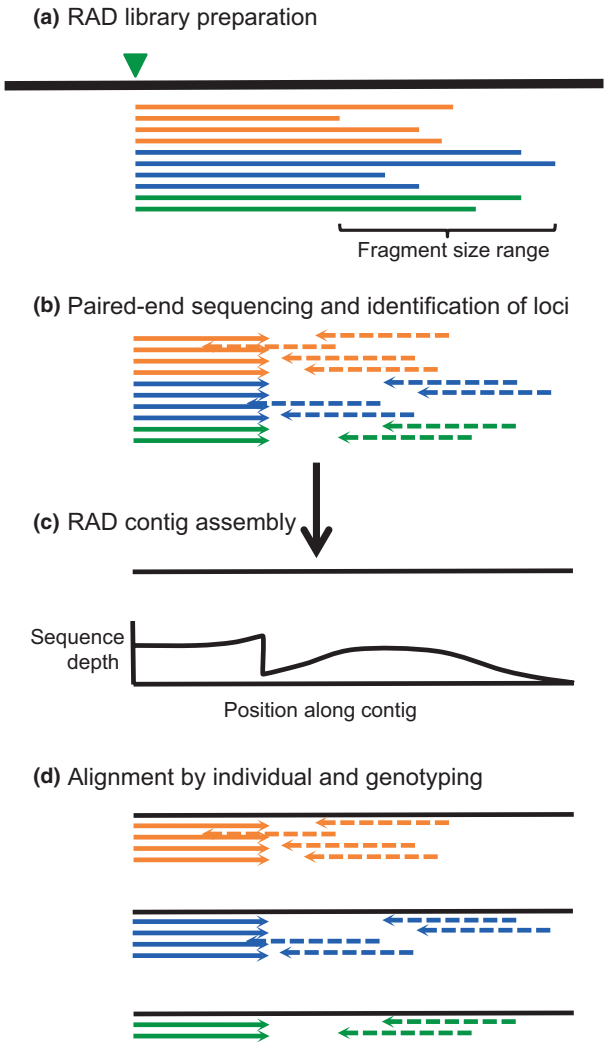

Fig. 2 Schematic diagram of overlapping paired-end restriction-site-associated DNA (RAD) sequencing. (a) RAD libraries are prepared according to Etter *et al.* (2011a,b), with the exception that a smaller size range of fragments are selected to obtain overlapping reads. The green triangle indicates the restriction enzyme cut site, and fragments from only one side of the cut site are shown for three individuals (represented by different colours). (b) Libraries are sequenced by Illumina with paired-end reads. Loci are identified with STACKS software, using only the forward reads (solid lines) to cluster reads by locus. (c) Both the forward and reverse reads from each locus are pooled across a set of individuals and assembled into a RAD contig. The depth of sequencing coverage across overlapping paired-end RAD contigs has a unique signature. (d) Reads from each individual are separately aligned against the reference contig set and diploid single nucleotide polymorphism genotypes are called statistically. The length of genotyped sequence data may vary across individuals, and in some cases genotype data may have a gap where paired ends did not overlap.

correct barcode and the remaining six bases of the SbfI recognition sequence. We then removed read pairs that represented PCR duplicates using the STACKS program clone_filter. The random shearing step in RAD sequencing produces staggered paired-end reads as described previously, so that any set of read pairs that are identical across both the forward and reverse reads are probably PCR duplicates of a single original genomic DNA fragment (Davey *et al.* 2011). Because genotyping depends on using read counts of alternative alleles in a statistical sampling model, PCR duplicates can be misleading because they do not represent independent samples from the genomic pool of DNA.

We identified RAD loci by applying ustacks to the forward reads across all individuals. We enabled the Deleveraging and Removal algorithms to filter out highly repetitive, likely paralogous loci, and we used a maximum nucleotide distance between stacks of 4 to achieve a balance between filtering paralogs and maintaining true alleles at a single locus approximately consistent with the expected number of RAD loci (Hohenlohe *et al.* 2011; Miller *et al.* 2011). We created a catalog of RAD tag loci using cstacks and matched individuals against the catalog using sstacks. We populated and indexed a MYSQL database of loci using load_radtags.pl and index_radtags.pl and then exported the data using export_sql.pl. Finally, we grouped the forward and reverse reads from each individual corresponding to each RAD locus using sort_read_pairs.pl.

Contig assembly

We pooled many individuals for contig assembly to increase sequence coverage of read pairs at each RAD locus (Fig. 2b). However, we also wanted to limit levels of polymorphism that could complicate assembly. Therefore, we pooled data from 60 individuals from the three populations with the lowest level of admixture as estimated from previous microsatellite data (Boyer *et al.* 2008): Lower Hay, Nicola and Tepee. We grouped the forward and reverse reads from all individuals in these populations into a separate file for each RAD locus, using the STACKS program sort_read_pairs.pl. We assembled the reads in each file separately to produce a set of RAD contigs (Fig. 2b), using both VELVET (Zerbino & Birney 2008) and CAP3 (Huang & Madan 1999) assembly software. Because CAP3 performed better (see Results), all further analyses mentioned below used the CAP3 assemblies. Because of our pooling strategy, the consensus sequences in this reference set of RAD contigs represent primarily WCT with minimal RBT admixture.

Genotyping and admixture estimates

We aligned the filtered read pairs for each individual from all five populations against the reference set of RAD contigs (Fig. 2c). (Three individuals with very

low coverage were dropped: one each from Meadow, Nicola and Tepee, leaving a total sample size of 94 individuals.) We used the alignment software BOWTIE (Langmead *et al.* 2009), allowing up to three nucleotide mismatches in the first 30 bp of each read and up to 15 mismatches over the total read. These parameters represent a compromise aimed at producing valid alignments to the reference, while minimizing bias against divergent RBT haplotypes. We chose them after aligning and genotyping a subset of the data across a wide range of parameter values, but we found that alignment parameters created only marginal differences in overall genotype calls (not shown). We retained only those read pairs that aligned uniquely to the reference contig set and that aligned in the expected orientation (i.e. the forward read aligns at position 0 of the contig, matching the position of the restriction enzyme cut site and the reverse read aligns in the opposite direction along the same contig within a distance up to 750 bp).

We assigned diploid genotypes to each nucleotide position for each individual using the maximum-likelihood method of Hohenlohe *et al.* (2010), modified by bounds on the per-nucleotide sequencing error rate of $0.0001 < \varepsilon < 0.0025$ and a significance level of $\alpha = 0.05$ (custom software available at http://webpages.uidaho.edu/hohenlohe/software.html). These limits on ε have the effect of being more likely to call a heterozygous genotype. While in *de novo* genotyping, these bounds on ε would increase the frequency of false alleles; here, we are genotyping against previously identified WCT and RBT alleles. This strategy and the relatively high significance threshold are also justified because of the quality filtering and removal of PCR duplicates described previously, which increases confidence that each read represents a true independent sample of genomic sequence.

We used previously identified species-diagnostic SNP loci to assess introgression from RBT into these WCT populations. From the RAD sequencing data in WCT and RBT published by Hohenlohe *et al.* (2011), we extracted all RAD loci in which there was either one SNP fixed between species and no other polymorphism in the 54-bp sequence (2923 loci), two fixed SNPs and no other polymorphism (643 loci), or one fixed SNP and one additional SNP polymorphic within either species (1348 loci), for a total of 4914 diagnostic SNPs (see Amish *et al.* 2012 for validation of some of these SNP markers). We aligned both the WCT and RBT alleles of these 54-bp sequences against the new reference set of RAD contigs, using BOWTIE (Langmead *et al.* 2009) and allowing up to two nucleotide mismatches. We retained only those diagnostic loci that aligned uniquely with up to two mismatches (for both the RBT and WCT alleles) to the reference contig set.

We then genotyped all individuals from the five admixed WCT populations in the current study as WCT, RBT or heterozygous at each of these loci for which genotype calls were made previously (any genotype calls that did not match previously identified alleles at these SNPs were treated as missing data). As a final filtering step for paralogous loci, we removed loci for which these genotypes exhibited observed heterozygosity >0.5 and $F_{IS} < -0.5$ (Hohenlohe *et al.* 2011). Using all such diagnostic SNPs for which at least half of the individuals (47 or more) were genotyped, we estimated proportion of admixture at the locus, individual and population levels as the frequency of RBT alleles across diagnostic loci.

We applied the heterogeneity test of Long (1991) to test for super invasive alleles. This analysis tests whether the variance in admixture across loci exceeds that expected from random sampling as well as genetic drift across loci (other tests for admixture outliers do not account for drift and may suffer from a high false-positive rate, so our approach is a conservative test; Fitzpatrick *et al.* 2009). Because this method cannot handle allele frequencies of 0.0, we used Bayesian estimates of allele frequencies with an uninformative prior (Fitzpatrick *et al.* 2009). We adjusted for differences in sample size of genotypes across loci, which affect the expected variance in allele frequency estimates, in equation 6 of Long (1991). For each locus in each population, we calculated a *P*-value for the deviation from expected admixture and adjusted for false discovery rate at a level of $\alpha = 0.05$ within each population (Benjamini & Hochberg 1995). We identified candidate super invasive alleles as those with significantly elevated admixture proportions in two or more populations.

Results

RAD sequencing and contig assembly

After filtering for read quality and presence of a correct barcode and SbfI recognition site, we generated 63 061 577 RAD sequence read pairs across 94 individuals in five admixed WCT populations. Of these, 22% represented PCR duplicates and were removed, leaving 49 248 922 unique read pairs. We identified a total of 222 830 putative RAD loci in STACKS using the forward reads of each pair across all individuals. Only 82 721 of these loci represented eight or more read pairs across all individuals.

We pooled the read pairs corresponding to these 82 721 loci for individuals from three populations with the lowest previously estimated admixture proportions (Lower Hay, Nicola and Tepee). We conducted separate assemblies at each locus using both VELVET (Zerbino

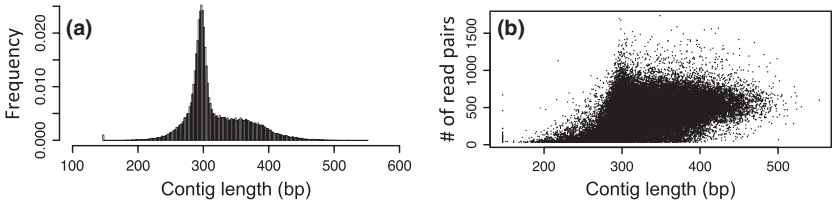

Fig. 3 (a) Frequency histogram of consensus sequence lengths across 77 141 contigs assembled by CAP3 from overlapping paired-end restriction-site-associated DNA (RAD) sequencing in admixed westslope cutthroat trout populations. (b) Relationship between sequencing depth at each locus (number of sequence pairs from 60 pooled individuals) and RAD contig length.

& Birney 2008) and CAP3 (Huang & Madan 1999). In VELVET, we used fixed *k*-mer lengths of 25, 35, 45 and 55 bp as well as optimizing the *k*-mer length across these values independently at each locus. All of these assemblies failed to connect overlapping paired-end reads at many loci, and the maximum contig length per locus was only ~100–300 bp (Fig. S1, Supporting information). Thus, in many cases, the contigs assembled were smaller than the read length of 147 bp (after trimming the barcode) for the forward reads (Fig. S1, Supporting information), meaning that sequences were broken into *k*-mers and unable to be reassembled. This difficulty in paired-end assembly of RAD data has been observed elsewhere (Davey *et al*. 2012), although that study had better success than we did in optimizing assembly parameters per locus. The general problem may be due to the unique signature of sequence coverage expected across contigs for overlapping paired-end RAD data (Fig. 2c; Etter *et al*. 2011a; Fig. 1 of Davey *et al*. 2012).

In contrast, the simpler algorithm of CAP3 performed much better. While more computationally intensive, it is still feasible on a desktop computer because the locus identification from STACKS significantly reduces the complexity of each individual assembly. Of the 82 721 loci, 72 124 (87.2%) assembled into single contigs, all but one containing both the overlapping forward and reverse reads. An additional 5017 loci assembled into two or more contigs, of which only the largest contig was anchored at the expected restriction enzyme recognition site. Of these, all but 151 contained both the forward and reverse reads. We combined these to produce our final reference set of RAD contigs, which contained 77 141 contigs from 82 721 loci (93.3%). Fragment size selection to produce overlapping paired-end reads was remarkably successful, so that over 93% of loci produced contigs spanning the forward and reverse reads. Contig lengths ranged from 147 to 519 bp with most between 250 and 450 bp (Fig. 3a), suggesting that longer fragments were carried through the gel-based size selection step. The mean number of read pairs contributing to each contig was 379.3. Contig length was positively related to the number of sequence pairs contributing to each assembly (Fig. 3b), so our strategy of pooling individuals to increase coverage at this consensus assembly step appears sound.

Genotyping and admixture

We aligned 54-bp RAD sequences for 4914 previously identified SNP loci (Hohenlohe *et al*. 2011; Amish *et al*. 2012) against the reference RAD contig set. Of these, 3456 (70.4%) aligned uniquely to a single contig in the reference set with up to two mismatches for both the RBT and WCT alleles. In addition, 392 (8.0%) aligned to multiple contigs with relatively few mismatches. These multiple contigs appear to represent genomic regions with duplicate sequence beyond the 54 bp length of the previously identified RAD sequence. Figure S2 (Supporting information) shows one such example in which sequences diverge relatively rapidly beyond the first 54 bp, illustrating how longer reads and overlapping paired-end RAD sequencing may provide powerful tools for distinguishing paralogous sequence from polymorphism at homologous loci.

We genotyped each individual at all nucleotide positions aligned to the reference contig set using the maximum-likelihood statistical approach described previously. Of the 3456 uniquely aligned diagnostic SNP loci, 3182 had diploid genotype calls for at least half the individuals sampled. Two of these were probably paralogous loci, with elevated observed heterozygosity (0.95 and 0.80) and reduced F_{IS} (−0.90 and −0.61, respectively), and these were removed from further analysis. The remaining 3180 loci had observed heterozygosity <0.45 and $F_{IS} > -0.23$, suggesting a clear break between them and the two presumptive paralogous loci. We translated genotypes for the final list of 3180 loci into homozygous WCT, heterozygous or homozygous RBT and assessed proportion of admixture as simply the frequency of RBT alleles.

For all of the individuals genotyped here, we also had individual-level estimates of admixture proportion based on seven species-diagnostic microsatellite loci (Boyer *et al*. 2008). Our SNP-based estimates were highly correlated with previous microsatellite-based estimates overall and within each population (Table 1),

Table 1 Correlation between previous microsatellite and current single nucleotide polymorphism (SNP)-based estimates of individual-level admixture proportions, and super invasive alleles exhibiting significantly elevated introgression with a false discovery rate corrected *P*-value

	SNP-microsatellite correlation			
Population	*r*	*P*-value	# super invasive alleles	FDR *P*-value threshold
Meadow	0.879	$<10^{-5}$	2	3.4×10^{-5}
Nicola	0.837	$<10^{-5}$	5	8.0×10^{-5}
Dutch	0.711	0.0009	1	2.0×10^{-3}
Lower Hay	0.960	$<10^{-10}$	5	4.9×10^{-4}
Tepee	0.844	$<10^{-4}$	4	4.2×10^{-4}
All 5 populations	0.805	$<10^{-15}$	2	7.8×10^{-5}

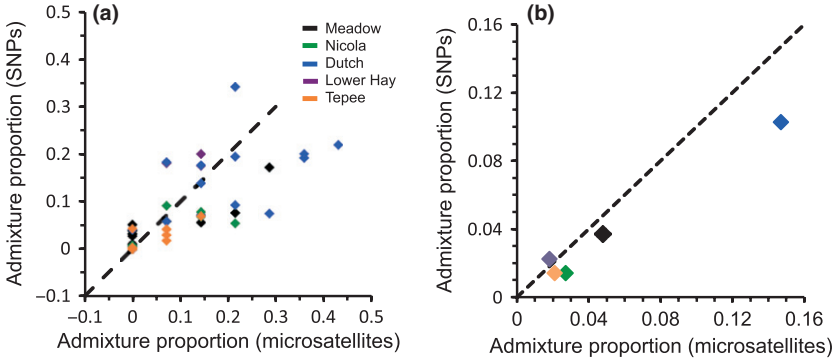

Fig. 4 (a) Individual-level admixture proportions estimated from seven diagnostic microsatellite loci (Boyer *et al.* 2008) *vs.* current estimates from 3180 single nucleotide polymorphism loci across 94 westslope cutthroat trout individuals from five populations. Note that many of the points, particularly those with admixture proportions near 0.0, lie on top of each other. (b) Population-level admixture proportions estimated from the same two data sets, calculated using only the individuals genotyped by both Boyer *et al.* (2008) and the current study.

although they tended to be slightly lower (Fig. 4a). We detected evidence of introgression in all 53 individuals for which no RBT alleles had been observed at the microsatellite loci. In these individuals, RBT alleles were detected at 1–235 loci, leading to individual admixture proportions ranging from 0.0013 to 0.0439 that were undetected in the microsatellite data. Average population-level admixture proportions are also consistent with microsatellite-based estimates (Pearson $r = 0.99$; $P = 0.0013$; Fig. 4b), in which Dutch and Meadow exhibited higher levels of admixture than the other three populations, although SNP-based estimates were lower than microsatellite estimates for four of the five populations.

Comparing admixture proportions across SNP loci reveals a positively skewed distribution within each population and overall, with many loci showing little or no admixture and a small set of outlier loci (Fig. 5). Of the 3180 diagnostic SNP loci genotyped, 634 showed no RBT alleles in any of the five populations. However, 94 loci exhibited admixture levels of 0.1 or greater across all five populations combined, up to a maximum of 0.542 (Fig. 5f). These are candidate super invasive alleles: RBT alleles that may have spread rapidly or have higher probabilities of persistence in WCT populations. Within each population, loci exhibited significantly elevated admixture proportions using the heterogeneity test of Long (1991), corrected for false discovery rate (Table 1). Three loci were significantly invasive in two or more populations, one of which was significant across all five populations (Fig. 5).

We conducted a translated nucleotide BLAST search using the RAD contig sequence for each of these three super invasive alleles. Two of them aligned closely to annotated genes whose function is consistent with selection in hybridized WCT populations. The locus significantly admixed in all five populations (RAD locus 118904) aligned significantly to the vertebrate gene latent transforming growth factor beta-binding protein 2 (LTBP2), with the most significant hit in *Bos taurus* (*E*-value = 10^{-7}). The second locus, significantly admixed in the Nicola and Tepee populations (RAD locus 117399), aligned to the vertebrate gene furry homolog-like (FRYL), with the most significant hit in zebrafish (*Danio rerio*, *E*-value = 10^{-9}). It is worth noting that the BLAST alignments to these two annotated gene

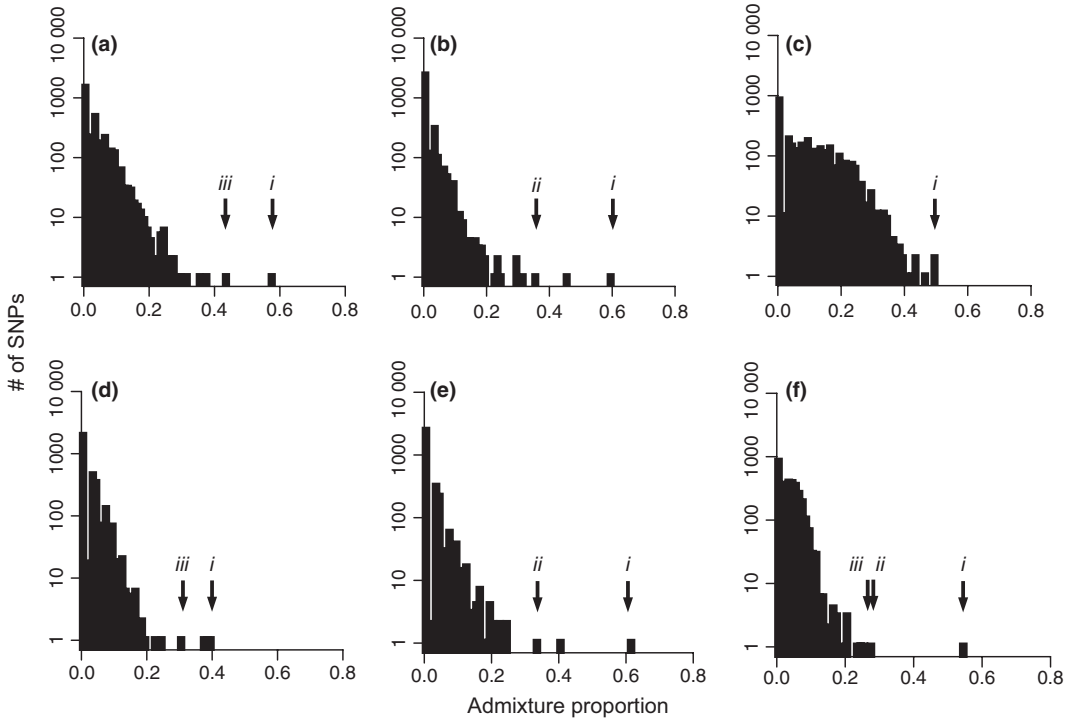

Fig. 5 Frequency histograms of admixture proportion across 3180 diagnostic single nucleotide polymorphism loci. (a) Meadow. (b) Nicola. (c) Dutch. (d) Lower Hay. (e) Tepee. (f) All five populations combined. Arrows indicate super invasive alleles—loci with significantly elevated admixture proportion ($\alpha = 0.05$, corrected for false discovery rate) independently in two or more populations. (i) restriction-site-associated DNA (RAD) locus 118904. (ii) RAD locus 117399. (iii) RAD locus 82847.

sequences began at nucleotide positions 191 and 210, respectively, of the RAD contigs so that the identification of these candidate genes would not have been possible solely with single-end RAD sequence data.

Discussion

Genomic tools hold remarkable promise for conservation and management of many taxa. The ability to rapidly identify and genotype large numbers of genetic markers allows improved estimates of demographic parameters (gene flow, effective population size, population-level admixture), as well as identification of outlier loci (locally adapted genes, invasive alleles). Overlapping paired-end RAD sequencing offers advantages for rapid development of large numbers of candidate SNPs that can be used in high-throughput genotyping assays, particularly in the case of large or repetitive genomes.

In a specific application of this technique, here we assessed genomic patterns of introgression and were able to detect individuals with very low levels of admixture, precisely estimate individual- and population-level admixture and detect candidate super invasive alleles driven to high frequency by selection. Below, we discuss some general aspects of the sequencing technique for conservation genomics and lessons from its application to the genomics of hybridization.

Overlapping paired-end RAD for conservation genomics

By assembling contigs of 400 bp or more adjacent to RAD loci, overlapping paired-end RAD provides sufficient flanking sequence for SNP assay design simultaneous with SNP discovery. The ability to generate sufficient flanking sequence has previously been a limitation of RAD sequencing for converting rapid SNP discovery to a set of high-throughput assays (Ogden 2011; Amish *et al*. 2012). Our approach can rapidly provide a multitude of candidate SNP markers for high-throughput assay development. Here, we only analysed a few thousand diagnostic markers that had been previously identified. In general, the majority of contigs of 300–400 bp or longer would be expected to contain SNPs relevant for most population genomic or conservation applications.

Assembling RAD contigs provides more continuous genomic sequence data for discriminating paralogous loci. This is a particular challenge in salmonids because of their ancestral genome duplication, which created homeologous duplicate sequence across the genome

(Allendorf & Danzmann 1997; Everett *et al.* 2011; Seeb *et al.* 2011b). Here we found examples of loci sharing very similar sequence over ~50 bp, so that they were grouped together in previous analysis, but diverged beyond that length. As a result, we were able to further screen the candidate diagnostic SNP loci we had previously identified (Hohenlohe *et al.* 2011; Amish *et al.* 2012) by removing the 8% that aligned to multiple RAD contigs. Ongoing validation of the reduced set will determine the success rate of these refined candidate markers.

Our approach to RAD contig assembly produced a single contig with high average read depth for most of our RAD loci. Nonetheless, the assembly and validation of RAD contigs can be challenging (Davey *et al.* 2012). Assemblies using the de Bruijn graph technique of VELVET (Zerbino & Birney 2008) produced consistently shorter contigs than a simpler (but more computationally intensive) assembly algorithm in CAP3 (Huang & Madan 1999) (compare Fig. 3 and Fig. S1, Supporting information). This contrasts with the results of Etter *et al.* (2011a,b), who had better success with VELVET in assembling the reverse reads from nonoverlapping paired-end RAD. Willing *et al.* (2011) used nonoverlapping paired-end RAD in guppies and assembled the reverse reads for 91.3% of loci into a single contig with generally lower sequence coverage than used at the assembly step here. That study used the assembler LOCAS, specifically designed by one of the authors for low-coverage data. Davey *et al.* (2012) had poor results with LOCASOPT and VELVET in assembling paired-end RAD data from *Heliconius* butterflies, but better results using the computationally intensive VELVETOPTIMISER. In our trout data set, over 87% of loci produced a single contig of both forward and reverse reads with CAP3, and many of the remainder could be filtered out as paralogs.

Techniques like overlapping paired-end RAD sequencing may allow new analytical power. Compared with other markers like microsatellites, SNPs can be limiting in that they typically exhibit only two alleles in natural populations. More power to understand population genetic processes would come from using multi-allelic haplotypes instead of SNPs in analyses of high-throughput sequence data (Gompert & Buerkle 2011; Buerkle *et al.* 2011). Because of the relatively long contigs that can be generated (Etter *et al.* 2011a,b; Willing *et al.* 2011) and because haplotype phase is known across read pairs and thus can be inferred along the length of RAD contigs, paired-end RAD offers the possibility of using haplotype- rather than SNP-based analyses. Genealogical relationships among multiple haplotypes are very useful for inferring demographic and evolutionary history (Sunnucks 2000; Beaumont & Rannala 2004).

Assessing genome-wide patterns of introgression

Here we provide one of the first genomewide assessments of human-mediated introgressive hybridization in salmonid fishes (see also Lamaze *et al.* 2012). Our results confirm previous patterns of hybridization between introduced RBT and native WCT in the North Fork Flathead system (Boyer *et al.* 2008; Muhlfeld *et al.* 2009c). Population-level admixture estimates were generally consistent for diagnostic microsatellites and RAD-based SNP loci, suggesting that thousands of diagnostic loci are generally unnecessary for approximate estimates of population-level admixture. However, one estimate did differ: the estimate for Dutch Creek was over 40% higher using the microsatellite data (Fig. 4b). This may be explained by selection against RBT alleles in chromosomal regions near RAD loci and/or sampling error from using only seven diagnostic microsatellite loci, especially for populations with low levels of introgression. Given the variation in introgression we observed here among SNP loci, the genomic location of those microsatellite loci could also be a major source of variation.

Overestimation of admixture (by using only a handful of neutral loci) could cause populations to not be protected under conservation laws, such as the U.S. Endangered Species Act (ESA). For Lahontan cutthroat trout, listed under the ESA, 10% RBT admixture is the threshold for a population to be protected as if it were nonhybridized (pure native) Lahontan. Based on sampling theory for neutral loci, it is likely that 50–100 diagnostic loci would improve accuracy to levels approaching that of thousands of RAD loci, if those diagnostic loci are widely distributed across the genome (Amish *et al.* 2012).

At the individual level, overlapping paired-end RAD sequencing allowed detection of very low levels of RBT introgression. Here, we detected RBT alleles in all 94 samples analysed, over half of which did not exhibit RBT alleles at seven microsatellite loci (Boyer *et al.* 2008). Some of the assumed RBT-diagnostic alleles could actually exist in nonhybridized WCT populations. Additional RAD sequencing of pure-native populations (e.g. isolated above barriers in the Flathead River) could help identify assumed diagnostic RBT alleles that might exist in WCT (e.g. due to maintenance of ancestral polymorphism).

Genomewide marker coverage is an important advance for conservation and management because it allows powerful screening of individuals to prevent inadvertent release of hybridized individuals into populations (e.g. during assisted migration, broodstock development, translocation and reintroduction) and identification of markers for rapid screening for early detection of hybridization. From a landscape genetics perspective, the ability

to precisely estimate admixture would allow fine spatial mapping of hybridization and introgression patterns. This approach may be useful in monitoring and preventing the spread of invasive species and their alleles in many plant and animal species facing hybridization threats in nature (Schwartz *et al.* 2007).

Dense coverage of markers across the genome allows for detection of candidate super invasive alleles—alleles of an invasive taxon that rise to much higher frequency (level of introgression) than the genomic background, analogous to outlier loci in genome scans for selection (Luikart *et al.* 2003). Here we detected several candidate super invasive alleles as evidenced by the distributions of admixture proportions among SNPs (in all populations) containing a long tail of outlier loci. Several of these loci were consistent as outliers across populations. Further study is needed to confirm that these are indeed RBT alleles that have introgressed into these WCT populations. The haplotype information provided by longer overlapping paired-end RAD (e.g. using 250-bp reads as provided by Illumina MiSeq technology) may facilitate that analysis. Further study would also be needed to identify the phenotypic and fitness consequences of these invasive alleles.

BLAST searching revealed close sequence matches for two candidate invasive alleles to vertebrate genes (LTBP-2 and FRYL). Super invasive alleles may be under positive selection and increase fitness in hybridized populations. Alternatively, they may spread by having phenotypic effects on dispersal or through segregation distortion, despite reducing overall fitness from outbreeding depression (Shine 2011). The LTBP family of proteins interacts with TGF-beta and has a wide range of developmental and physiological functions, including effects on fertility (Morén *et al.* 1994; Öklü & Hesketh 2000; Kosova *et al.* 2012), although the specific relationship between LTBP-2 and TGF-beta is unclear (Hirani *et al.* 2007). In RBT, the related protein LTBP-3 and other related proteins have been implicated in early ovarian development and early embryonic development (Andersson & Eggen 2006; Lankford & Weber 2010; Gahr *et al.* 2012), suggesting the hypothesis that the RBT allele at this locus positively affects fecundity in admixed individuals. It is exciting that future research and additional studies like this one will help understand mechanisms driving super invasive alleles and genomewide introgression in natural populations.

Acknowledgements

P.A.H. and M.D.D. received support from U.S. National Institutes of Health/NCRR grant P20RR16448 (L. Forney, PI). FWA and G.L. were partially supported by the U.S. National Science Foundation grants DEB-0742181. G.L. also received support from Montana Fish Wildlife and Parks and NSF grant DEB-1067613. This work was partially supported by BPA contract #199101993. We thank J.J. Giersch for providing the study area map, Robb Leary for helpful comments on the potential of some assumed diagnostic alleles for RBT to be present at low frequency in nonhybridized WCT, and Montana Fish Wildlife and Parks and Glacier National Park for support and help in sampling. Any use of trade, product or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This research was conducted in accordance with the Animal Welfare Act and its subsequent amendments.

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

Amish SJ, Hohenlohe PA, Painter S *et al.* (2012) RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Molecular Ecology Resources*, **12**, 653–660.

Andersson ML, Eggen RI (2006) Transcription of the fish latent TGFbeta-binding protein gene is controlled by estrogen receptor alpha. *Toxicology in vitro*, **20**, 417–425.

Angeloni F, Wagemaker N, Vergeer P, Ougorg J (2011) Genomic toolboxes for conservation biologists. *Evolutionary Applications*, **5**, 130–143.

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

Baxter SW, Davey JW, Johnston JS *et al.* (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. *PLoS One*, **6**, e19315.

Beaumont M, Rannala B (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics*, **5**, 251–261.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Boyer MC, Muhlfeld CC, Allendorf FW (2008) Rainbow trout (*Oncorhynchus mykiss*) invasion and the spread of hybridization with native westslope cutthroat trout (*Oncorhynchus clarkia lewisii*). *Canadian Journal of Fisheries and Aquatic Sciences*, **65**, 658–669.

Buerkle CA, Gompert Z, Parchman TL (2011) The $n = 1$ constraint in population genomics. *Molecular Ecology*, **20**, 1575–1581.

Campbell NR, Amish SJ, Pritchard VL *et al.* (2012) Development and evaluation of 200 novel SNP assays for population genetic studies of westslope cutthroat trout and genetic identification of related taxa. *Molecular Ecology Resources*, **12**, 942–949.

Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3 Genes Genomes Genetics*, **1**, 171–182.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2012) Special features of RAD sequencing data: implications for genotyping. *Molecular Ecology*, **22**, 3151–3164.

Etter PD, Johnson EA (2012) RAD paired-end sequencing for local *de novo* assembly and SNP discovery in non-model organisms. In: *Data Production and Analysis in Population Genomics: Methods and Protocols* (eds Pompanon F & Bonin A), pp. 135–151. Humana Press, New York, New York.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011a) Local *de novo* assembly of RAD paired-end contigs using short sequencing reads. *PLoS One*, **6**, e18561.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011b) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds Orgogozo V, Rockman MV), pp. 157–178. Humana Press, New York, New York.

Everett MV, Grau ED, Seeb JE (2011) Short reads and non-model species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources*, **11**, 93–108.

Finger AJ, Stephens MR, Clipperton NW, May B (2009) Six diagnostic single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trout. *Molecular Ecology Resources*, **9**, 759–763.

Fitzpatrick BM, Johnson JR, Kump DK, Shaffer HB, Smith JJ, Voss SR (2009) Rapid fixation of non-native alleles revealed by genome-wide SNP analysis of hybrid tiger salamanders. *BMC Evolutionary Biology*, **9**, 176.

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.

Gahr SA, Weber GM, Rexroad CE III (2012) Identification and expression of Smads associated with TGF-β/activin/nodal signaling pathways in the rainbow trout (*Oncorhynchus mykiss*). *Fish Physiology and Biochemistry*, **38**, 1233–1244.

Gompert Z, Buerkle CA (2009) A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology*, **18**, 1207–1224.

Gompert Z, Buerkle CA (2011) A hierarchical Bayesian model for next-generation population genomics. *Genetics*, **187**, 903–917.

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

Harwood AS, Phillips RB (2011) A suite of twelve single nucleotide polymorphism markers for detecting introgression between cutthroat and rainbow trout. *Molecular Ecology Resources*, **11**, 382–385.

Hirani R, Hanssen E, Gibson MA (2007) LTBP-2 specifically interacts with the amino-terminal region of fibrillin-2 and competes with LTBP-1 for binding to this microfibrillar protein. *Matrix Biology*, **26**, 213–223.

Hitt NP, Frissell CA, Muhlfeld CC, Allendorf FW (2003) Spread of hybridization between native westslope cutthroat trout, *Oncorhynchus clarki lewisi*, and nonnative rainbow trout, *Oncorhynchus mykiss*. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 1440–1451.

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.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.

Kalinowski ST, Novak BJ, Drinan DP, Jennings Rde M, Vu NV (2011) Diagnostic single nucleotide polymorphisms for identifying westslope cutthroat trout (*Oncorhynchus clarki lewisi*), Yellowstone cutthroat trout (*Oncorhynchus clarki bouvieri*) and rainbow trout (*Oncorhynchus mykiss*). *Molecular Ecology Resources*, **11**, 389–393.

Kelly BP, Whiteley A, Tallmon D (2010) The Arctic melting pot. *Nature*, **468**, 891.

Kosova G, Scott NM, Niederberger C, Prins GS, Ober C (2012) Genome-wide association study identifies candidate genes for male fertility traits in humans. *The American Journal of Human Genetics*, **90**, 950–961.

Lamaze FC, Sauvage C, Marie A, Garant D, Bernatchez L (2012) Dynamics of introgressive hybridization assessed by SNP population genomics of coding genes in stocked brook charr (*Salvelinus fontinalis*). *Molecular Ecology*, **21**, 2877–2895.

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.

Lankford SE, Weber GM (2010) Temporal mRNA expression of transforming growth factor-beta superfamily members and inhibitors in the developing rainbow trout ovary. *General and Comparative Endocrinology*, **166**, 250–258.

Long JC (1991) The genetic structure of admixed populations. *Genetics*, **127**, 417–428.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

McGlauflin MT, Smith MJ, Wang JT *et al.* (2010) High-resolution melting analysis for the discovery of novel single-nucleotide polymorphisms in rainbow and cutthroat trout for species identification. *Transactions of the American Fisheries Society*, **139**, 676–684.

Miller MR, Brunelli JP, Wheeler PA *et al.* (2011) A conserved haplotype controls parallel adaptation in geographically distant salmonid populations. *Molecular Ecology*, **21**, 237–249.

Miller JM, Poissant J, Hogg JT, Coltman DW (2012) Genomic consequences of genetic rescue in an insular population of bighorn sheep (*Ovis canadensis*). *Molecular Ecology*, **21**, 1583–1596.

Morén A, Olofsson A, Stenman G *et al.* (1994) Identification and characterization of LTBP-2, a novel latent transforming growth factor-β-binding protein. *Journal of Biological Chemistry*, **269**, 32469–32478.

Morin PA, Luikart G, Wayne RK, SNP workshop group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution*, **19**, 208–216.

Muhlfeld CC, Kalinowski ST, McMahon TE *et al.* (2009a) Hybridization rapidly reduces fitness of a native trout in the wild. *Biology Letters*, **5**, 328–331.

Muhlfeld CC, McMahon TE, Belcer D, Kershner JL (2009b) Spatial and temporal spawning dynamics of native westslope cutthroat trout, *Oncorhynchus clarkii lewisi*, introduced

rainbow trout, *Oncorhynchus mykiss*, and their hybrids. *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 1153–1168.

Muhlfeld CC, McMahon TE, Boyer MC, Gresswell RE (2009c) Local habitat, watershed, and biotic factors influencing the spread of hybridizations between native westslope cutthroat trout and introduced rainbow trout. *Transactions of the American Fisheries Society*, **138**, 1036–1051.

Narum SR, Campbell NR, Kozfkay CC, Meyer KA (2010) Adaptation of redband trout in desert and montane environments. *Molecular Ecology*, **19**, 4622–4637.

Ogden R (2011) Unlocking the potential of genomic technologies for wildlife forensics. *Molecular Ecology Resources*, **11**, 109–116.

Öklü R, Hesketh R (2000) The latent transforming growth factor β binding protein (LTBP) family. *Biochemical Journal*, **352**, 601–610.

Perry GML, Danzmann RG, Ferguson MM, Gibson JP (2001) Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (*Oncorhynchus mykiss*). *Heredity*, **86**, 333–341.

Pritchard VL, Abadia-Cardoso A, Garza JC (2012) Discovery and characterization of a large number of diagnostic markers to discriminate Onchorhynchus mykiss and O. clarkii. *Molecular Ecology Resources*, **12**, 918–931.

Schwartz MK, Luikart G, Waples RS (2007) Genetic monitoring as a promising tool for conservation and management. *Trends in Ecology and Evolution*, **22**, 25–33.

Seeb JE, Pascal CE, Ramakrishnan R, Seeb LW (2009) SNP genotyping by the 5′-nuclease reaction: advances in high throughput genotyping with non-model organisms. In: *Methods in Molecular Biology, Single Nucleotide Polymorphisms* (ed. Komar A), 2nd edn, pp. 277–292. Humana Press, New York, New York.

Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011a) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11**, 1–8.

Seeb LW, Templin WD, Sato S *et al.* (2011b) Single nucleotide polymorphisms across a species' range implications for conservation studies of Pacific salmon. *Molecular Ecology Resources*, **11**, 195–217.

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.

Shine R (2011) Invasive species as drivers of evolutionary change: cane toads in tropical Australia. *Evolutionary Applications*, **5**, 107–116.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology and Evolution*, **15**, 199–203.

Teeter KC, Thibodeau LM, Gompert Z, Buerkle CA, Nachman MW, Tucker PK (2010) The variable genomic architecture of isolation between hybridizing species of house mice. *Evolution*, **64**, 472–485.

Twyford AD, Ennos RA (2012) Next-generation hybridization and introgression. *Heredity*, **108**, 179–189.

Willing E-M, Hoffmann M, Klein JD, Weigel D, Dreyer C (2011) Paired-end RAD-seq for de-novo assembly and marker design without available reference. *Bioinformatics*, **27**, 2187–2193.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

P.A.H., S.J.A., F.W.A. and G.L. conceived and designed experiments. S.J.A., M.R.M., M.C.B. and C.C.M. contributed samples and supporting data. N.K.-H. and E.A.J. constructed RAD libraries and conducted sequencing. P.A.H. and M.D.D. analyzed the data. P.A.H. and G.L. wrote the paper with contributions from all co-authors.

Data accessibility

Raw sequence data, RAD contig sequences, and genotype data: Dryad doi:10.5061/dryad.32b88.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Frequency histograms of contig length across all loci assembled from overlapping paired-end restriction-site-associated DNA sequencing, using VELVET.

Fig. S2 A representative example of discrimination of duplicate sequence with longer reads and paired-end restriction-site-associated DNA sequencing.