Migration traits are presumed to be complex and to involve interaction among multiple genes. We used both univariate analyses and a multivariate random forest (RF) machine learning algorithm to conduct association mapping of 15,239 single nucleotide polymorphisms (SNPs) for adult migration-timing phenotype in steelhead (Oncorhynchus mykiss). Our study focused on a model natural population of steelhead that exhibits two distinct migration-timing life histories with high levels of admixture in nature. Neutral divergence was limited between fish exhibiting summer- and winter-run migration owing to high levels of interbreeding, but a univariate mixed linear model found three SNPs from a major effect gene to be significantly associated with migration timing ($p < 0.000005$) that explained 46% of trait variation. Alignment to the annotated Salmo salar genome provided evidence that all three SNPs localize within a 46 kb region overlapping GREB1-like (an oestrogen target gene) on chromosome Ssa03. Additionally, multivariate analyses with RF identified that these three SNPs plus 15 additional SNPs explained up to 60% of trait variation. These candidate SNPs may provide the ability to predict adult migration timing of steelhead to facilitate conservation management of this species, and this study demonstrates the benefit of multivariate analyses for association studies.

1. Introduction

Fitness trade-offs may serve to differentiate migrating individuals and cause them to exhibit particular strategies that vary by timing [1], speed [2], distance [3] and destination of migration [4]. The attributes of migration behaviour are probably fine-tuned within populations to confer optimal fitness for a particular habitat. Genetic mechanisms can operate and drive variation in migratory behaviour, and the study of this architecture has begun to enhance our understanding of a variety of different migratory insects, birds and fishes [5]. A group of anadromous fishes, salmonids (Family Salmonidae), represent an opportunity to expand this understanding given their life-history diversity of migratory traits [6] and the availability of resources for study of these species owing to their high cultural and economic value. Specifically, the bioinformatic resources for salmonids now include whole-genome sequences of Salmo salar [7] and Oncorhynchus mykiss [8]. These resources increase the likelihood that candidate loci identified by genome-wide association studies (GWAS) of salmonids will be annotated and thereby provide insight into homologous functioning of genetic architecture of migratory behaviour.

In this study, we test for genome-wide association of a migratory trait in steelhead (O. mykiss), which is the anadromous form of this salmonid species. Steelhead exhibit a high degree of life-history diversity throughout their geographical range, including two divergent freshwater migration-timing life histories of adults known as winter- and summer-run. Both run types spend similar periods of time rearing in freshwater and the ocean (1–4 years for...
each habitat). However, summer-run steelhead return to their natal freshwater streams between May and October when they are relatively sexually immature, and winter-run steelhead return to freshwater between December and May as sexually mature fish [9]. Both summer- and winter-run steelhead hold in freshwater until they spawn together each year on the ascending limb of flows during annual spring run-off [10]. The alternative life-history strategies exhibited by the summer- and winter-run steelhead of the Columbia River Basin are probably associated with fitness trade-offs that may explain the differences in their respective geographical distributions [11]. For example, summer-run steelhead displays an optimal migration timing for passage over partial barriers such as waterfalls that are relatively easy to navigate during summer flows. However, the fitness cost of this strategy may be relatively high pre-spawning mortality owing to the protracted freshwater residence associated with this early migration timing [12].

Genetic analysis of native, sympatric summer and winter runs indicate that these two migratory forms are more similar within drainages than either type is to other populations of the same migratory type [13,14]. However, the degree of gene flow among the two migratory types can vary depending on location. For the Columbia River, only summer-run steelhead are distributed upstream of the major historic barrier of Celilo Falls (near the eastern edge of the Cascade Crest; electronic supplementary material, figure S1). However, downstream of this point both runs have historically coexisted. In some coastal tributary populations where hatchery stocking has occurred, separation in their spawn timing and habitat preference has generally reinforced some degree of reproductive isolation [15–17]. This isolation can cause an accumulation of divergence throughout the genome, which would present added challenge to uncovering the genetic basis of the trait of interest because any alleles that vary between subpopulations can lead to spurious associations with the trait [18]. However, our study focused on steelhead from a tributary of the Columbia River, the Klickitat River, which is located in an intermediate geographical location between coastal and inland lineages where neutral genetic divergence is very low for native anadromous steelhead [19] and both migratory life-history types occur in sympathy. Therefore, this represents a model system for association testing of migratory life histories with little confounding neutral structure.

Traits underlying migration behaviour that appear bimodal can be polygenetic, involving interaction among multiple genes of minor effect [20], and so we included both univariate and multivariate models for association tests of 15 239 SNPs with adult migration timing in steelhead. This combination of approaches takes advantage of testing framework of univariate analysis to evaluate independent associations of SNPs with our trait of interest, along with benefits of multivariate analyses with random forest (RF) such as the ability to capture the complexity of epistatic interactions that are typically involved among genes underlying quantitative traits [21]. Further, we examined whether information (the degree of association each SNP has with the trait) from the univariate analyses could be useful as a prior to help narrow the search for candidate SNPs in the RF approach and increase efficiency for identifying a set of SNPs with greatest prediction power for the trait of interest.

2. Material and methods

(a) Fish collection, sample pre-screening and DNA analysis

Adult steelhead were trapped throughout each year of a consecutive 3-year period (2007–2009) in the Klickitat River at the Lyle Falls fish ladder (Rkm 2.4; electronic supplementary material, figure S1) for non-lethal collection of biological data, and then released back into the stream. Biological data included migration date, fork length, sex and a clip of tissue from the caudal fin. DNA was extracted and genotypes from 180 TaqMan SNP assays were analysed initially to filter out non-target samples (non-local or hatchery-origin steelhead) via STRUCTURE v. 2.3.4 analysis [22] (electronic supplementary material, figure S2 and supplemental methods S1). This set of 180 SNPs has been highly vetted for Hardy–Weinberg equilibrium across the steelhead range in the Columbia River Basin and is a standard marker set for characterizing population structure in this region [23]. Only fish determined to be natural-origin steelhead from the Klickitat River were retained for further inclusion in this study (electronic supplementary material, figure S3). This included a total of 320 fish over three migration years (n = 83 for 2007; n = 93 for 2008; n = 144 for 2009).

The trait we tested was migration timing in units of ordinal day as the phenotype, which we ordered to reflect the biological sequence of annual steelhead migrations (figure 1). For both summer- and winter-run life-history types, spawning coincides temporally in April of each year for all steelhead that returned to the Klickitat River during all previous seasons of the migration year, as well as those that returned during January to April of the current year. Therefore, we reordered April 30 as ordinal day 365 and May 1 as ordinal day 1 to represent the end of one spawning cycle and the beginning of the next migration and spawning cycle, respectively (figure 1). The summer-run migration typically lasts from May to October of the same calendar year. The winter-run migration typically begins around December and continues through April of the following calendar year. Thus, there are two periods where the two migratory types are expected to overlap; primarily in the transition from summer-run to winter-run, which is thought to occur in the autumn months (September through November); and secondarily at the end of the spawning season in April, where late winter-run fish and early summer-run fish are migrating.

DNA from each of the 320 steelhead and a set of 10 doubled haploid O. mykiss individuals (included to identify paralogous sequence variants, PSVs) was quantified in 96-well assay plates using the Quant-IT dsDNA pico-green assay kit (Life Technologies, Grand Island, NY, USA) and a Perkin Elmer Victor 5  plate reader. RAD-seq libraries were prepared for Illumina sequencing with a protocol that has been previously published [24] and subsequently modified [25]. Prior to sequencing, RAD libraries were quantified by q-PCR and Illumina library quantification standards (Kappa Biosystems Inc, Woburn, MA, USA) on an ABI 7900HT Sequence Detection System (Life Technologies). Libraries were sequenced with single-end 100 bp reads on an Illumina HiSeq 1500 (Illumina Inc., San Diego, CA, USA).

(b) Bioinformatics

Bioinformatics for de novo SNP discovery and genotyping were completed with STACKS [26], and subsequently filtered to remove markers and individuals that failed quality thresholds (electronic supplementary material, supplemental methods S2). The final dataset included 15 239 SNPs with greater than 3% MAF that were successfully genotyped across more than 80% of 237 individuals. The decrease in sample size from the initial 320 individuals was largely due to tissue degradation. These
failed tissues appeared evenly distributed across run times and were unlikely to cause sample bias. Sequence for each RAD tag was used to align to the *O. mykiss* and the *S. salar* reference genomes using BOWTIE2 [27]. The positions from the alignment to the *O. mykiss* genome were used to order the loci by chromosome number. Many of the tags were found to occur within an unknown chromosome and so were simply ordered by their scaffold position number after the markers on the last chromosome (i.e. after sex chromosome). Coding regions of the *O. mykiss* genome were identified within 5 kb of each SNP locus, and for comparison, coding regions of the annotated *S. salar* genome were also examined within 5 kb of the putative candidate SNPs identified in this study. To identify gene functions and annotations, coding sequences were then queried against the NCBI nucleotide sequence database using the software program BLAST2GO [28]. We used program search parameters Blastx (BLAST program), nt (Blast database), retrieve 10 Blast hits and Blast expectation = 10−5 default value. A Fisher’s exact test in the FatiGo [29] package was used to remove linear dependency as recommended in the TASSEL manual [30].

(c) Univariate genome-wide association studies
We performed univariate analyses using a general linear model (GLM) and a mixed linear model (MLM) with TASSEL v. 5.1.0 [30]. The GLM is a fixed effects linear model that is used in TASSEL to identify significant associations between phenotypes and genotypes. TASSEL takes population structure into account by using membership in underlying populations as covariates in the model. The MLM is similar to GLM but includes both fixed (i.e. gender, year, population structure and genetic marker) and random effects (i.e. relationships among individuals), and can thus account for both population structure and kinship to reduce false positive associations [31]. Equations for the GLM and MLM are described in the TASSEL manual [30]. A kinship matrix using the ‘scaled IBS’ method [32] based on all 15 239 SNPs was generated in TASSEL to represent cryptic familial relationships. The MLM was implemented using default options (i.e. ‘P3D’ [33] parameter option and the ‘optimum compression’ option). The GLM effectively represents a ‘maximum compression’ option, and thus provides contrast to the MLM. Permutation tests (1000) were used to calculate p-values to determine significant associations of SNPs with traits. Owing to the potential to identify false associations when many SNPs, traits and individuals are included, a Bonferroni correction was applied to a level 0.05 (i.e. corrected $\alpha = 3.28 \times 10^{-5}$) to reduce false positives and stringently control for Type I errors. The association tests using a GLM were performed with covariates of gender, year and population structure. For population structure, coefficients of ancestry for nine STRUCtURE clusters (only 9 of the $K = 10$ clusters from the STRUCtURE ‘filter’ analysis were used to remove linear dependency as recommended in the TASSEL manual). This dataset was also analysed using an MLM, and the kinship matrix was included as an additional covariate. Some individuals (4%) were missing gender data, which was imputed in TASSEL by using default parameter settings to average the nearest three neighbours.

(d) Random forest
We performed the RF multivariate analysis on our dataset by using two different approaches (henceforth referred to as ‘RF-rank’ and ‘MLM-rank’), and within each approach we employed ‘coarse-sweep’ and ‘fine-sweep’ analyses to identify the number of SNPs required to explain the most trait variation. For the first approach (‘RF-rank’), our ‘coarse-sweep’ involved an iterative process (electronic supplementary material, supplemental methods S3) to build a set of predictive models for the migration trait (dependent variable) based on subsets of the 15 239 SNP loci (independent variables). The RF model’s predictive accuracy is estimated by using a portion (approx. 2/3) of the data that are randomly selected with replacement to build a predictive model, which is then used to predict the phenotypes of the remainder of the data [34]. The amount of trait variance explained by the predicted values relative to the total amount of variation in the observed phenotypes yields the percentage of trait variation explained, a measure of the model’s predictive accuracy. Results from this ‘coarse-sweep’ analysis showed that the maximum phenotypic variance could be explained with the top 25 loci. Using this group of loci, we then ranked their importance using a backward purging analysis [34], which was

![Figure 1. Mean monthly counts of adult steelhead passing Lyle Falls during 2007–2014. Ordinal ‘day’ was re-ordered to begin with the summer-run period and end after the spawning period. The bi-modal distribution represents the peaks of the summer- and winter-run steelhead migration periods (horizontal bars) and includes a ‘transitional’ period of overlap in distributions (hatched bars). Sample sizes (total $n = 237$) analysed from the three time periods are indicated.](http://rspb.royalsocietypublishing.org/10.1098/rspb.2015.1564)
automated with R scripts [35]. In this backward purging (i.e. ‘fine-sweep’ analysis), the least important loci are removed one by one, starting with a greater number than the initial optimum number of loci (here, the best 150 loci rather than the top 25; electronic supplementary material, supplemental methods S3). This ‘fine-sweep’ analysis was necessary to identify candidate SNPs with more precision than could be accomplished with the ‘coarse-sweep’ analysis, and this analysis helps to minimize the potential negative interactions across loci that may otherwise prevent identification of the best candidate SNPs [34,35].

The migration trait data in this case were the residuals of the day of passage trait from the previous MLM univariate analysis, which minimized bias from population structure and kinship. For comparison, we also used the uncorrected trait values (i.e. re-ordered ordinal day of migration timing) to estimate the percentage of total variation that each set of SNPs could explain. Genotypes were converted to 0 (aa), 0.5 (ab) and 1 (bb), and missing data were imputed using the nearest three individuals which was generated using TASSEL. The RF analyses were performed using the RANDOMFOREST package [36] in R.

For the second approach (‘MLM-rank’), we used the results of the univariate MLM analysis in order to rank all the loci based on their p-value, lowest to highest. As a ‘coarse-sweep’ analysis, we used subsets of top-ranked loci (i.e. top 3, 5, 10, 25, 50, 75, 100, 200, 300, 400, 500, 750, 1500, 3000 and 7500 SNPs) to determine the approximate number of loci necessary to explain maximal phenotypic variation. Subsequently, we conducted a ‘fine-sweep’ backward purging analysis using a starting point of the 150 top-ranked SNPs. We felt this ‘MLM-rank’ approach would make a useful comparison to the alternative ‘RF-rank’ approach, which relied solely on importance values estimated by RF analysis to rank loci and was more time-consuming due to the non-automated, iterative steps. We evaluated the two approaches for their efficiency (i.e. their ability to identify the smallest number of loci required to explain the most amount of trait variation).

(e) Ability of top candidate markers to distinguish alternative run types

The top candidate markers identified by the ‘MLM-rank’ approach were analysed in STRUCTURE to demonstrate their effectiveness for distinguishing individuals according to their migration-timing. Individuals were categorized into three groups divided by the periods May–August, September–November and December–April to capture a summer, transitional and winter-run group, respectively (figure 1). Correlation between ordinal day-of-passage and probability of ancestry to a cluster (output from STRUCTURE) was estimated using a Mantel test performed with 9999 permutations in PASS@GE v. 2 [37]. We calculated pairwise FST among the three migration-timing groups using ARLEQUIN v. 3.5.1.2 [38] with 10000 permutations, and significance was determined at \( a = 0.05 \) level corrected for multiple comparisons using B–Y method FDR [39] (corrected \( a = 0.01092 \)). For comparison, these same analyses were repeated using a set of 180 non-candidate SNPs that are part of a standard panel for genetic stock identification and population structure analyses in the Columbia River Basin [23].

3. Results

(a) Univariate genome-wide association studies

The univariate analyses performed in TASSEL indicated that MLM (figure 2) was a better fit than GLM (electronic supplementary material, figure S4) due to the inclusion of the kinship matrix as a covariate in the MLM. Three SNPs were significantly associated with migration timing in the MLM, specifically SNPs 47080_53, 52458_16 and 54772_22 (Bonferroni’s adjusted \( a \leq 3.28 \times 10^{-16} \)). These three SNPs were highly evident in the QQ plot (figure 2a) above the 1:1 line and outside the significance threshold, with the remainder of markers not statistically significant and largely fitting expected p-values. We used the residual trait values from TASSEL for subsequent multivariate analyses because of the relatively good fit of this model, and its ability to minimize bias from both population structure and cryptic relatedness.

(b) Random forest multivariate analysis

The initial ‘coarse-sweep’ using the RF-rank approach (with no prior information from MLM) demonstrated that approximately 25–50 SNPs were able to explain the largest amount of residual trait variation (approx. 44%; figure 3a). The subsequent ‘fine-sweep’ backward purge analysis corroborated this approximate number of SNPs; however, the analysis more precisely identified a set of 44 SNPs to explain 46% of residual trait variation (figure 3b) or 60% uncorrected trait variation. The coarse-sweep and fine-sweep analyses were highly concordant, because 84% of the top 44 SNPs identified in the fine sweep were included in the top 50 SNPs identified in the coarse-sweep. Further, these overlapping SNPs included the three candidate SNPs identified by the univariate analysis, and two of these candidate SNPs were ranked first and second by RF importance values.

The MLM-rank approach using a coarse-sweep analysis in RF demonstrated that this approach was less effective than RF-rank. With the MLM-rank approach, many more SNPs (a minimum of 100 SNPs) were necessary to explain the highest residual trait variation. Additionally, the highest residual trait variation explained with the MLM-rank approach (approx. 27%; figure 3a) was much lower than that with the RF-rank approach (27% versus 44% of variation, respectively). The same top three SNPs were identified, but were only able to explain 7% residual variation (46% uncorrected variation). Overall, the RF-rank coarse-sweep approach was more efficient because it identified fewer SNPs to explain a greater amount of variation.

Despite relatively low efficiency of the MLM-rank approach in the coarse-sweep analysis, backward purging with fine-sweep analysis using the top-ranked 150 SNPs performed more comparably to the RF-rank approach. A relatively small number of SNPs (18) were identified with the MLM-rank fine-sweep analysis that could explain 44% of the residual trait variation (figure 3b) or 60% uncorrected trait variation. This level of explained residual trait variation with 18 SNPs surpassed the limit of residual trait variation (approx. 27%) that could be explained from the coarse-sweep (figure 3a), and was similar to the level achieved with the RF-rank fine-sweep analysis (46% residual variation with 44 SNPs), but with many fewer markers. Among the 18 top SNPs, 78% of them were found in common among the 44 top SNPs identified by the RF-rank approach. Further, the same three candidate SNPs identified by the MLM univariate analysis were included among the top candidate SNPs identified by both RF-rank and MLM-rank approaches.

(c) Candidate loci for migration timing

In total, 8224 of our 15 239 loci were found to be within 5 kb of at least one coding DNA sequence from the O. mykiss reference genome. Using a single coding sequence per locus, there
were 8154 loci that returned at least one sequence description, gene ontology term or InterProScan result in the BLAST2GO analysis (electronic supplementary material, table S1). The gene ontology enrichment analysis using a Fisher’s exact test identified no statistically significant enrichments for either the 18 top MLM-rank loci or the top 44 RF-rank loci after applying FDR [40] for multiple testing. Lack of significant enrichments may be due to limited annotation of candidate SNPs (e.g. among the 18 top MLM-rank loci there were only eight SNPs determined to have a sequence description, and of those, 3 were identified with GO names; electronic supplementary material, table S1). Coding DNA sequence from the Atlantic salmon genome corroborated some (\(n = 12\)) of the \(O.\ mykiss\) genome annotation of candidate loci and identified eight additional homologous genes nearby these candidate loci (electronic supplementary material, table S1); however, these additions still yielded no statistically significant gene ontology enrichments.

Based on the \(O.\ mykiss\) genomic positions of the 18 top MLM-rank loci and the top 44 RF-rank loci, nearly half of the SNPs were on an unknown chromosome, and the remaining SNPs were distributed across multiple chromosomes and did not appear to be concentrated on any particular linkage group (electronic supplementary material, table S1). However, alignment of these candidate loci to the \(S.\ salar\) genome provided more complete chromosome positional information (94\% of loci were positioned on chromosomes), and notably showed that a single chromosome (ssa03) included all three candidate SNPs identified by the MLM univariate analysis and two additional candidate SNPs (77429_7 and 34348_17). Among the three candidate SNPs identified by the MLM univariate analysis, we determined that one SNP (47080_53) occurs in \(O.\ mykiss\) chromosome 28 (figure 2b), and localizes to the gene GDF11, which was recently characterized in \(O.\ mykiss\) [41]. The other two SNPs, 52458_16 and 54772_22, were found to occur less than 3 kb apart in an unknown \(O.\ mykiss\) chromosome (i.e. ChrUn positions 366307580 and 366310211, respectively). The region spanned by these two markers contains three CDS of a predicted gene in the GREB1/GREB1-like gene family (electronic supplementary material, figure S5). The alignment to the annotated \(S.\ salar\) genome corroborated the close proximity of SNPs 52458_16
and 54772_22 within GREB1-like, and, further, provided evidence that all three SNPs localize within a 46 kb region on chromosome Ssa03 (electronic supplementary material, figure S5). However, based on the annotated S. salar genome, there was no evidence that SNP 47080_53 is in fact localized within a gene homologous to the O. mykiss GDF11; therefore, its close proximity to the GREB1-like gene is perhaps a more notable and substantiated characteristic.

The STRUCTURE analyses with the top 18 candidate SNPs showed support for two clusters with steep increase in mean LnP(K) between K = 1 and K = 2. The mean LnP(K) plateaued for K > 2 (electronic supplementary material, figure S6a). The proportion of ancestry to these clusters (figure 4a) was correlated with migration timing (r = 0.458, figure 4b), and the Mantel test was significant (p = 0.0001, B-Y FDR corrected α = 0.0083). By contrast, based on the standard 180 SNP marker set used for general population genetic applications, mean LnP(K) plateaued between K = 1 and 2, after which point Pr(K) declined rapidly (electronic supplementary material, figure S6b). Delta K [42] supported K = 2 for both marker sets, but the correlation of proportion of ancestry (figure 4c) versus migration timing was lower (r = 0.037) for the 180 SNP marker set (Mantel test, p = 0.0056; figure 4d). Pairwise FST values were high and significant (B-Y FDR corrected α = 0.05) [39] among migration trait categories based on the 18 candidate loci, and relatively lower (70× lower) and mostly non-significant for the 180 SNPs (electronic supplementary material, table S2).

4. Discussion

Our study provides evidence for the genetic basis of differentiation in migration timing exhibited by anadromous steelhead. However, we cannot conclude that these alternative life-history strategies are entirely under genetic control. Instead, this example may be similar to the polygenic class of threshold traits, or alternative migratory tactics exhibited generally across salmonid fishes [43]. Specifically, individual variation in behaviour and physiology of alternative migratory tactics (e.g. residency versus anadromy) may be controlled by developmental thresholds; and a genetic basis may be controlling the variation in proximal traits activating the development of alternative migratory tactics [43,44]. Furthermore, phenotypic plasticity may contribute to variation in this trait as has been observed in other organisms [45–47]. Our study also supports that this migration trait is polygenic, and the combination of univariate GWAS and multivariate RF analyses that we employed was helpful for uncovering multiple SNPs and their associated genomic regions that may be involved in this trait and are distributed across multiple chromosomes. Our approach to use information from the univariate analyses as
Figure 4. STRUCTURE plot of individual ancestry to two clusters for (a) 18 candidate SNPs and (c) 180 non-candidate SNPs. Individuals were ordered by their migration timing into the Klickitat River and grouped by the summer-run (May–Aug), transitional (Sep–Nov) and winter-run (Dec–Apr) periods. Plots of the Mantel test correlation between individual ancestry and day of migration (units are ordinal day as reordered summer–winter) are shown for (b) 18 candidate SNPs and (d) 180 non-candidate SNPs.

(a) Genetic architecture of migration timing in steelhead

Concentrated genetic architecture (i.e. few quantitative trait loci, QTL, of large effect) has been predicted to evolve under a set of conditions that include, among other factors, higher rates of gene flow between diverging populations compared with conditions leading to more diffuse genetic architecture (i.e. many QTL of small effect) [48]. Recent empirical studies demonstrate that even seemingly complex phenotypes can be controlled by concentrated architecture [49,50]. Life-history strategies for migration timing in salmonids appear to be less extreme than these examples of traits with concentrated genetic architecture, because many loci of both large and small effects have been implicated across chromosomes [35,51]. Our study of migration timing in steelhead lends support to this notion because this trait involves at least one chromosome region of relatively large effect as well as loci across multiple chromosomes with relatively minor effects. The region of large effect was identified by three SNPs that were clearly distinguished by their highly significant association with this trait using GWAS analyses and were found to localize within a 46 kb region of a single chromosome (ss03) based on the *S. salar* genome alignment. Placement of these three SNPs together on a single *O. mykiss* chromosome is also likely given that chromosome ss03 of *S. salar* shares synteny with *O. mykiss* chromosome 28 [52], and one of these three SNPs (47080_53) that could be assigned to an *O. mykiss* chromosome was found on chromosome 28.

These three SNPs were found to localize within a region that overlaps at least one gene: an oestrogen receptor cofactor, GREB1. Homologous functions of the GREB1 gene that have been characterized in other species appear to support a putative major role in this salmonid migration trait. For example, GREB1 mediates the interaction of oestrogen receptors with other proteins and plays an important role in enhancing the ability of oestrogen to act on cells to induce transcription [53]. This link to oestrogen receptors is relevant to upstream migration timing in steelhead and other salmon, because these species migrate upstream when they are sexually maturing. Further, oestrogen (steroid hormone) levels in the blood increase when salmon are maturing and affect multiple developmental pathways in sexual maturation, including oogenesis, vitellogenesis and testicular development [54].

However, the three top SNPs were found to explain a relatively small portion of the residual variation (approx. 7%) of the migration-timing trait (i.e. the variation remaining after kinship, population structure, gender and year effects were accounted for in the univariate MLM). A multivariate approach using the RF machine learning algorithm demonstrated a much higher proportion of residual trait variation (44%) could be explained using a minimum of 18 minor candidate SNPs. We also demonstrated that the variation at these 18 candidate SNPs can distinguish the individuals in this dataset by their migration timing in contrast with a set of 180 non-candidate SNPs that are a standard marker set for population genetic applications in the Columbia River Basin [23]. Therefore, this evidence suggests multiple loci of both major and minor effects may be involved in this migration trait.

In addition to GREB1, the set of 18 candidate SNPs included five other genes identified by our Blast2Go query: transmembrane protein 230-like gene (TMEM230-like), raptase-activating-like protein (I0GAP3), angiopeptin-related protein 2-like (ANGPTL2), glutamate receptor (NMDA 2C-like) and an unnamed protein product. However, GO-enrichment analysis did not identify a significant pattern of genes involved in any particular function. The enrichment analysis was limited by small sample sizes of the number of
genes that could be identified and by the relatively low-density genomic scan we could accomplish with 15 000 SNPs. There are a number of assumptions required when linking a SNP locus with a gene based on homology, including a subjective decision on how nearby a gene should be to consider it linked to a SNP [55]. We chose a conservatively narrow distance (5 kb) threshold within which to identify genes nearby candidate SNPs, and note concordance of gene annotation information from the *S. salar* and *O. mykiss* genomes for all five of the 18 candidate SNPs that could be compared. It may be beneficial to reassess the *O. mykiss* genomic positions and gene ontology of these candidate loci in the near future following continued improvements to the genomic resources for this species [56].

(b) Utility for a combined univariate and multivariate approach in genome-wide association studies

We demonstrated high concordance between results from univariate analyses and RF by noting that of the three SNPs (54772_22, 47080_53 and 52458_16) that were significant in the univariate MLM analysis, two SNPs held the first and second rank (54772_22 and 47080_53, respectively) in the RF analysis (RF-rank backward purging). The third SNP (52458_16) was ranked 15th by this RF analysis, which is probably due to the fact that it is in near perfect linkage disequilibrium with the first-rank SNP. The level of concordance of rankings between the univariate MLM analysis and RF analysis decreased as a larger set of top-ranked markers were compared between methods (i.e. 29% of markers were overlapping in the 150 top-ranked by p-values from univariate analysis and 150 top-ranked by RF importance values).

Importantly, we show that ranking of loci based on results from a univariate analysis may be helpful for narrowing the RF search. Using a starting point of a set of 150 top-ranked SNPs based on p-value (lowest to highest) from the univariate analysis, the RF backward purging steps were able to identify a set of relatively few SNPs (18) required to reach nearly maximal explained residual trait variation (44%). Therefore, when results from a univariate GWAS are used to help inform a search within RF, this step may enable one to identify an optimal set of candidate SNPs (based on their explanatory power for a trait of interest) but with fewer steps than if one were to use RF exclusively. More investigation is warranted with this relatively new application of RF for GWAS; however, RF as a stand-alone analysis may have difficulty finding the optimum solution when so many possible (albeit suboptimal) solutions are probably present in the data given the many thousands of SNP loci that are used in a typical GWAS. Further, univariate analyses can help to overcome one limitation of RF (i.e. a lack of statistical properties for evaluation of SNP variables found to have high predictive power for a trait of interest [21]). The advantage of RF for identifying sets of candidate SNPs that can operate synergistically to provide maximal predictive power for a trait of interest will probably continue to drive its adoption as an analytical tool for GWAS.

(c) Conservation implications

The candidate genetic markers identified in this study may help predict the adult migration timing of individual steelhead throughout the entire course of their life cycle, which would benefit long-term conservation management of this protected species. Ability to identify the genetic propensity for migratory traits in steelhead would be useful for a multitude of applications including characterizing differences associated with these adult alternative migration tactics that pertain to pre-adult life stages (e.g. juvenile migration and size-at-age), and categorizing adults on spawning grounds into migration categories. Currently, steelhead are categorized into summer- or winter-run based on the timing when they enter streams near the mouth of their natal tributary. However, steelhead may overwinter in freshwater areas outside of their natal tributary, which complicates their classification as summer- or winter-run, and therefore underscores the need for a method of genetic classification. Further research is warranted to characterize the extent to which this genetic mechanism for this migration-timing trait applies across the geographical distribution of the species.

**Ethics.** Trapping and sampling was authorized by permit from the National Marine Fisheries Service under the Endangered Species Act.

**Data accessibility.** The data used in this study are available in Dryad: http://dx.doi.org/10.5061/dryad.62qhn.

**Authors’ contributions.** J.E.H. conceived and designed the study, analysed the data and drafted the manuscript; A.R.M. conducted bioinformatics; S.R.N. coordinated the data collection; A.R.M. conducted bioinformatics; S.R.N. coordinated the study and helped draft the manuscript. All authors gave final approval for publication.

**Competing interests.** We have no competing interests.

**Funding.** This study was funded by Bonneville Power Administration.

**Acknowledgements.** We thank Marine Brieuc, Benjamin Hecht and Nate Campbell for their help with analyses. We also thank two anonymous reviewers and Fuwen Wei for suggestions to improve the original manuscript.

**References**


