POOLPARTY2: An integrated pipeline for analyzing pooled or indexed low coverage whole genome sequencing data to discover the genetic basis of diversity

Stuart Willis[1], Steven Micheletti[1], Kimberly Andrews[2], and Shawn Narum[1]

[1]Columbia River Inter-Tribal Fish Commission
[2]University of Idaho

June 6, 2023

Abstract

Whole genome sequencing data allow survey of variation from across the genome, reducing the constraint of balancing genome sub-sampling with recombination rates and linkage between sampled markers and target loci. As sequencing costs decrease, low coverage whole genome sequencing of pooled or indexed-individual samples is commonly utilized to identify loci associated with phenotypes or environmental axes in non-model organisms. There are, however, relatively few publicly available bioinformatic pipelines designed explicitly to analyze these types of data, and fewer still that process the raw sequencing data, provide useful metrics of quality control, and then execute analyses. Here, we present an updated version of a bioinformatics pipeline called POOLPARTY2 that can effectively handle either pooled or indexed DNA samples and includes new features to improve computational efficiency. Using simulated data, we demonstrate the ability of our pipeline to recover segregating variants, estimate their allele frequencies accurately, and identify genomic regions harboring loci under selection. Based on the simulated data set, we benchmark the efficacy of our pipeline with another bioinformatic suite, ANGSD, and illustrate the compatibility and complementarity of these suites by using ANGSD to generate genotype likelihoods as input for identifying linkage outlier regions using alignment files and variants provided by POOLPARTY2. Finally, we apply our updated pipeline to an empirical dataset of low coverage whole genomic data from uncurated population samples of Columbia River steelhead trout (Oncorhynchus mykiss), results from which demonstrate the genomic impacts of decades of artificial selection in a prominent hatchery stock.

PoolParty2: An integrated pipeline for analyzing pooled or indexed low coverage whole genome sequencing data to discover the genetic basis of diversity

Stuart Willis*[1], Steven Micheletti[2], Kimberly R. Andrews[3], and Shawn Narum[1]

[1]Hagerman Genetics Lab, Columbia River Inter-Tribal Fish Commission, Hagerman, ID, USA

[2]Dept. Zoology, University of British Columbia, Vancouver, BC, Canada

[3]Institute for Interdisciplinary Data Sciences, University of Idaho, Moscow, ID, USA

[4]Dept. Fishery Science, Columbia River Inter-Tribal Fish Commission, Portland, OR, USA

*corresponding author's email: swillis@critfc.org

Abstract

Whole genome sequencing data allow survey of variation from across the genome, reducing the constraint of balancing genome sub-sampling with recombination rates and linkage between sampled markers and target loci. As sequencing costs decrease, low coverage whole genome sequencing of pooled or indexed-individual

samples is commonly utilized to identify loci associated with phenotypes or environmental axes in non-model organisms. There are, however, relatively few publicly available bioinformatic pipelines designed explicitly to analyze these types of data, and fewer still that process the raw sequencing data, provide useful metrics of quality control, and then execute analyses. Here, we present an updated version of a bioinformatics pipeline called PoolParty2 that can effectively handle either pooled or indexed DNA samples and includes new features to improve computational efficiency. Using simulated data, we demonstrate the ability of our pipeline to recover segregating variants, estimate their allele frequencies accurately, and identify genomic regions harboring loci under selection. Based on the simulated data set, we benchmark the efficacy of our pipeline with another bioinformatic suite, angsd, and illustrate the compatibility and complementarity of these suites by using angsd to generate genotype likelihoods as input for identifying linkage outlier regions using alignment files and variants provided by PoolParty2. Finally, we apply our updated pipeline to an empirical dataset of low coverage whole genomic data from uncurated population samples of Columbia River steelhead trout (*Oncorhynchus mykiss*), results from which demonstrate the genomic impacts of decades of artificial selection in a prominent hatchery stock.

Introduction

A primary goal of molecular ecology is to understand the genetic basis of diversity such as targets of divergent selection or loci underlying heritable life history variations or ecotypes. Critical to this endeavor is the ability to survey the genome to discover genetic variants associated with phenotypic differences or environmental axes (Günther & Coop, 2013; Hoban et al., 2016; Paril, Balding, & Fournier-Level, 2022). Massively parallel or 'next-generation' sequencing has dramatically decreased the cost of surveying genetic variation across statistically meaningful numbers of individuals and has made these kinds of investigations accessible for researchers working with limited budgets on non-model organisms. However, despite the rapid decrease in per-base sequencing costs, sequencing the complete genome of each surveyed individual at high coverage is often not practical, in part because as the cost of sequencing has decreased but demands for statistically robust sample sizes have become more ardent (Schlötterer, Tobler, Kofler, & Nolte, 2014). As a result, geneticists are still faced with the task of determining the appropriate compromise between number of reads devoted to surveying each individual, which in many cases determines the extent of the genome that can be observed, and the number of individuals surveyed (Lou, Jacobs, Wilder, & Therkildsen, 2021).

This compromise has been addressed in a number of ways depending on the goals of the individual project. Many researchers have opted to survey only a fraction of the genome, creating 'reduced representation' libraries wherein sequencing coverage is spread across fewer, reproducible subsets of loci (e.g. Baird et al., 2008). With careful tuning of library preparation and sequencing methods, the coverage at each locus may be sufficient to confidently infer genotypes across nearly all individuals at hundreds to tens of thousands of variable loci (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016; Puritz et al., 2014). For analyses where individual genotypes are important but high genomic density of loci is not critical, such as determining relatedness or migration among recently diverged populations, these reduced representation techniques can produce data cost effectively for hundreds of individuals (e.g. Willis, Hollenbeck, Puritz, & Portnoy, 2022). However, while these techniques provide data on many more loci than what was historically accessible, only a fraction of the genome is ultimately surveyed, meaning that for species for which linkage blocks are typically less than 100Kb, many linkage blocks may not be surveyed. As a result, except in cases of regions of high linkage disequilibrium such as inversions or strong selective sweeps, investigations that only survey a few thousand linkage groups may fail to identify loci strongly associated with selection or heritable phenotypic variation (Lowry et al., 2017; Tiffin & Ross-Ibarra, 2014).

On the other hand, for many association and genome scan methods the input is allele frequencies rather than individual genotypes. And notably, because of sampling variance, sampling more individuals at low coverage actually provides more accurate estimates of phenotype or population allele frequencies than sequencing fewer individuals at high coverage (Futschik & Schlötterer, 2010; Günther & Coop, 2013; Schlötterer et al., 2014; Zhu, Bergland, González, & Petrov, 2012). Many analyses, including those that compare allele frequencies between phenotypic variants or populations situated along an environmental gradient and depend on high

density sampling across linkage groups to discover the regions of highest divergence, may thus be performed more effectively by low coverage whole genome sequencing (lcWGS) (Lamichhaney et al., 2012; Lou et al., 2021; Schlötterer et al., 2014; Therkildsen & Palumbi, 2017). Moreover, there have been a proliferation of analyses that are able to account for uncertainty in the genotype of each individual (likelihoods), even with data sequenced with <1x coverage per individual (Lou et al., 2021). This low coverage sequencing approach provides compromise among the portion of the genome surveyed, accurate allele frequency estimates, and in many cases analyses that require individual genotype data (Lou et al., 2021; Therkildsen & Palumbi, 2017). However, while lcWGS data may be highly appropriate for these types of investigations and the toolkit for analyzing allele frequency and genotype probability data is expanding, there remain few pipelines specifically designed to take unmapped lcWGS reads and convey the data through quality control and bioinformatic analyses.

To address that need, an integrated, modular bioinformatic pipeline, PoolParty, was developed that facilitates the use of lcWGS data to search for genomic regions showing strong divergence between samples with discrete phenotypic differences or other group-wise characteristics (Micheletti & Narum, 2018). This pipeline has been applied to detect genome-wide genetic association across multiple species (e.g. Aguirre-Ramirez, Velasco-Cuervo, & Toro-Perea, 2021; Horn, Kamphaus, Murdoch, & Narum, 2020; Lyu et al., 2021; Ren et al., 2021). Although most published applications have utilized data from libraries of pooled DNA, the pipeline can also utilize data from individuals sequenced in multiplex using indexed or barcoded libraries, which allows a normalization procedure that corrects for uneven contribution to group allele frequencies across individuals. This normalization is a pseudo-genotying method wherein each individual, regardless of total reads, is allowed to contribute only one or two alleles per locus to the count from which allele frequencies are calculated, depending on that individual's depth of coverage and the ratio of major and minor alleles (>10:1 is considered a homozygote; Figure 1). PoolParty shares this goal of managing uneven contribution among individuals when estimating allele frequency with another bioinformatic suite designed for use with lcWGS data, angsd, which also generates individual genotype likelihood or posterior probabilities from lcWGS data (Korneliussen, Albrechtsen, & Nielsen, 2014). However, PoolParty takes sequence read files as input, performs sequence cleaning and mapping to a reference genome, produces numerous assurance reports regarding sequence and mapping quality, and facilitates several analyses to identify regions of significant genomic divergence between samples, while angsd requires mapped read alignments produced by other tools as input. Moreover, as demonstrated below, the alignment files created by PoolParty are compatible input to angsd, making these complementary bioinformatic tools for lcWGS data analysis.

To demonstrate various utilities and upgrades of the PoolParty2 pipeline and compatibility with angsd, we apply it to two lcWGS datasets, one simulated and one empirical, that reflect the type of questions to which PoolParty2 may be routinely applied. We utilize data simulated to reflect different demographic contexts and degrees of sequence coverage to show the relative strengths, accuracy, and complementarity of PoolParty2 and angsd to identify segregating loci, estimate their allele frequencies, and identify outlier loci and the boundaries regions affected by selection. Then, using barcoded lcWGS data from natural and hatchery populations of steelhead trout (anadromous *Oncorhynchus mykiss*), we demonstrate the potential of integrated application of these bioinformatic suites to identify regions under selection in landscape-level population samples.

Methods

The bioinformatic pipeline : PoolParty2

PoolParty2 is an updated suite of scripts written in the BASH and R computer languages that create and manipulate text files, including sequence read files, and call freely distributed programs to efficiently operate on the data as needed. After installation of dependencies in a Linux computing environment, for which we provide explicit instructions on our Github page (https://github.com/stuartwillis/poolparty) and most of which are available using the Conda package and environment management system (Anaconda Software Distribution), users need only provide sequence read files and haploid genome assembly, a text file listing sequence read files with their group or population affiliation, and tailored configuration files for each of

the three modules as appropriate. We distribute two tutorials that with the scripts that help ensure that dependencies are accessible and illustrate the main features of the pipeline. We additionally provide example code to assist users in conveying output from the PoolParty2 modules into angsd and associated utilities.

The three main modules of the pipeline focus on distinct aspects of the bioinformatics process. The PPalign module calls dependency packages (i.e., BWA mem) for quality trimming, mapping and filtering, and SNP calling functions to create read alignments to the genome assembly, identify genetic variants and their frequencies, and produce input files for the other modules. The PPstats module utilizes output from the first module and reports a number of useful statistics about the sample groups, such as genomic extent at candidate depths or coverage variation among chromosomes, and allows the user to confirm that sufficient and similar coverage has been achieved across samples. The PPanalyze module utilizes and subsets allele summary data from the first module and performs user-specified analyses, such as principal components, sliding window F_{ST}, and Fisher's Exact Tests, to resolve population structure and identify regions of significant genetic divergence between groups. Additional modules are provided that run further statistical tests that utilize replicate sample pairs (Cochran, 1954; Mantel & Haenszel, 1959), which take into account background variance and linkage (local score; Fariello et al., 2017), or account for population structure (Lewontin and Krakauer test with kinship, or FLK; Bonhomme et al., 2010), as well as one for plotting results from these analyses.

Computational requirements for running the pipeline will depend on the size of the dataset and user-specified configuration, and may range from a handful of threads and tens of Gb of RAM to dozens of processors and >1Tb of RAM. Runs for each module usually last a few hours but could take several days for large datasets with limited processing and RAM resources. In the tutorials we describe strategies for piecemeal runs of the different modules as data are generated and assembled to coordinate and combine data subsets, confirm quality early in the process, and reduce the overall bioinformatic processing time.

Application 1:

lcWGS data simulated from distinct demographic backgrounds and coverage

We employed simulated data from Lou et al. (2021) to demonstrate the ability of these bioinformatic suites to utilize lcWGS data to accurately estimate allele frequencies and identify outlier regions at various coverages and sample sizes. Details of simulated data are included in Lou et al. (2021), but briefly, nucleotide sequence data including mutation and recombination on a single 30Mb chromosome were simulated for two populations exchanging genes under two demographic scenarios: a lower effective population size and lower rate of gene exchange that produced a higher background F_{ST} (hereafter, the 'high background F_{ST}' scenario) and a higher gene exchange and effective population size that produced a lower background F_{ST}('low background F_{ST}' scenario). In both scenarios, several sites under selection were introduced and allowed to evolve under divergent selection in each population, ultimately resulting in seven outlier regions (Supplemental Figure 1). Sequence data from the simulated chromosomes, generated to reflect Illumina-style paired-end reads including sequencing errors, reflected 8x coverage for each of several hundred individuals from each population to enable down-sampling at various coverage levels. See Lou et al. (2021) for additional details about simulated data.

While extensive scenarios were tested in Lou et al. (2021)with these simulated data, we selected four scenarios to benchmark PoolParty2 against angsd including: a) high background F_{ST} & low sequence coverage; b) high background F_{ST} & higher sequence coverage; c) low background & low sequence coverage; d) low background F_{ST} and higher sequence coverage. From the simulated sequence data, we randomly selected a number of individuals from each population to reflect sample sizes that are common for empirical datasets in the literature and the dataset included in this study (70 and 63 from each population) and down-sampled the simulated sequence data (8x) to reflect the median individual coverage in our empirical data (˜0.33x) as well as a common sequencing target for lcWGS studies (1x). As Lou et al. (2021) examined the effects of depth and sample size on allele frequency estimation accuracy across a greater range of values, it was not our intent to duplicate their efforts except to compare the abilities of the two bioinformatic suites at these coverage depths. However, to additionally challenge these suites with the range of variability in coverage

among individuals commonly reflected in empirical data, we fit several simple mathematical distributions to the sums of individual coverage in our empirical data for variant positions following an initial set of global filters (global depth of 10 and minor allele frequency, MAF, of 0.005). A logistic distribution exhibited the best fit based on information theoretic criteria (results not shown), and using parameters for this distribution, we down-sampled the 8x simulated data. For the 1x coverage dataset, the empirical scale parameter for this distribution was used, but the location parameter was set to 1, to produce higher coverage with similar variation. Individuals utilized at different coverages of the same scenario were not identical. This resulted in four simulated datasets (low and high background F_{ST} at 0.33x and 1x coverage).

For each of the four datasets, sequence data were processed with the PPalign module of PoolParty2, including quality trimming (sliding window PHRED [?] 20, retained length [?] 50bp), read mapping (mapQ [?] 20), variant scoring and filtering by global parameters (snpQ [?] 20, global depth of 10 reads, MAF of 0.001), and estimation of raw and normalized allele frequencies. Unless otherwise indicated, normalized allele frequencies were utilized in all analyses. SNP variants and their normalized frequencies identified by PPalign were used as input to PPanalyze, which applied additional filters for sites observed in fewer than 10 reads per population. PPanalyze also calculated $F_{ST}$for individual sites and sliding windows of 100Kbp windows in 5Kbp sets, as well as applying Fisher's Exact tests (FET) for differences in allele frequency between the two populations. Significance (p) values for the latter test were used as input for the Local Score test (Fariello et al., 2017), which 'smooths' the background variation in significance to identify contiguous outlier regions depending on a user-specified smoothing parameter (ξ). In addition, this test determines the significance of putative outlier regions through accounting for linkage (autocorrelation in p -values) in a chromosome-specific manner. However, because each run samples different SNPs to calculate autocorrelation and determine significance, individual runs may fail to identify outlier regions with borderline significance. Lower rates of smoothing (smaller ξ) generally retain more power, but are less precise in determining the boundaries of outlier regions, and moreover, because the landscape of background significance changes with various ξ values, and thus the thresholds for significance in this test, power and smoothing do not have a directly inverse linear relationship. For these reasons, multiple runs with application of different values of ξ are useful. We therefore ran the local score test three times for increasing values of ξ representing the 70^{th}, 80^{th}, 90^{th}, 95^{th}, and 99^{th} quantiles of significance values from the Exact tests of each dataset, and tallied how often each simulated outlier region was recovered as well as the width estimated for each region.

Using the filtered BAM files produced by PPalign as input, we applied angsd to all four simulated datasets in two manners. First, we ran angsd undirected by any variant identification from PoolParty2, relying on angsd's filters to identify and screen variant sites. We applied filters to several runs of each dataset (similar to PPalign: mapQ [?] 20, snpQ [?] 20, global depth [?] 20, MAF [?] 0.001) but with varying significance thresholds for variant discovery: none, 0.01, and 0.001. We ran this configuration to compare angsd's ability to detect true variants and discover outlier regions to PoolParty2. Subsequently, specifying for angsd to consider only variants it discovered with the most stringent significance threshold (p [?] 0.001), we ran angsd on data from each simulated population separately in order to generate site frequency spectra and calculate F_{ST} using angsd's realSFS utility. We only retained sites observed in more than 10 reads and 3 individuals in each population, and then directed ANGSD to run association tests on these sites (-doAssoc 1, 2, and 5) to identify outliers, specifying population membership of each individual. Second, we directed angsd to consider only sites discovered by PPalign and subsequently retained by filtering with PPanalyze ("PoolParty2 to angsd"), and we utilized allele frequencies estimated from these runs to compare the ability of PoolParty2 and angsd to recover the known allele frequencies accurately. Subsequently, genotype likelihoods inferred by angsd from these runs were used as input for estimation of linkage using ngsLD (Fox, Wright, Fumagalli, & Vieira, 2019), and we constrained linkage estimation to sites [?] 100Kbp from one another. Using these linkage estimates, we calculated mean LD in 100Kbp windows in 5Kbp steps in R, and identified outlier regions as contiguous series of [?]10 windows exceeding 2x the interquartile range (2xIQR) for mean windowed LD.

Application 2: Barcoded individuals in population samples of steelhead

Hatcheries have an important but controversial role in supplementing dwindling fish stocks in the Columbia River basin (Busby, Wainwright, & Bryant, 1999), including, in a few cases, selection for particular traits in hatchery stocks that differ from the stocks into which they are outplanted or stray (disperse to non-natal areas). One of the most abundant and widely outplanted hatchery stocks of steelhead trout in the Columbia Basin comes from Skamania Hatchery (Washougal, WA). The Skamania stock has a long history of deliberate selection for earlier spawning and larger fish (Ayerst, 1976), which has resulted in the evolution of fish that migrate notably earlier than conspecifics and almost exclusively after two or more years ocean duration (Hess et al., 2021). Without choosing individuals with known phenotypes, but rather undirectedly sampling individuals from the Skamania hatchery stock as well as individuals from two nearby natural origin stocks (Lewis River and Eagle Creek-Willamette River) in the same steelhead lineage (Coastal), we tested if genomic regions previously associated with these traits or others would appear strongly differentiated in the Skamania stock.

Library preparation followed the individual barcoding protocol from Horn et al. (2020) and sequencing was done separately for each population on the Illumina NextSeq 550 with 150-bp paired-end reads. The number of individuals per pool ranged from 60 to 78. Data were processed with PoolParty2, including discarding of reads if trimmed below 50bp from sliding windows with a minimum mean PHRED quality of 20, and filtering SNPs if they were below a PHRED quality of 20, three or fewer bases from an insertion-deletion position, observed in fewer than 10 reads in each sample pool or more than 1,500 globally, if the number of individuals surveyed per population was fewer than three of if the global minor allele frequency was less than 0.005. The allele frequency data were normalized in PPalign to mediate non-uniform read contribution among individuals. Using the PPstats module, we assessed data coverage distributions, proportion of the genome covered at specified depths, and evenness of coverage across chromosomes. Normalized allele frequencies were filtered and analyzed with PPanalyze including calculation of F_{ST}, sliding window F_{ST} (100Kbp windows in 5Kbp steps), and Fisher's Exact test (FET). Significance values from the Exact tests were used in local score analyses, using three replicate runs with ξ representing the 80^{th}, 90^{th}, 95^{th}, and 99^{th} quantiles of significance values (the 70^{th} quantile did not produce a mean local score distribution below zero). Filtered read alignment files (BAMs) created by PPalign were used as input for angsd, which was directed to consider the variants filtered by PPanalyze, and from which we utilized the genotype likelihoods provided by angsd as input to estimate linkage with ngsLD for three chromosomes with the most significant and consistent outlier regions in the Local Score results, considering only sites [?] 100Kbp from one another. As above, we calculated mean LD in 100Kbp windows in 5Kbp steps in R, but identified outlier regions as contiguous series of [?]20 windows exceeding 2x the interquartile range (2xIQR) for mean windowed LD. When multiple contiguous outlier window series were present in the range identified by the lowest Local Score quantile, we report all those series.

Results

Application 1: lcWGS data simulated from distinct demographic backgrounds and sequencing coverage

Down-sampling, trimming, and mapping of simulated data results in 100% of the simulated chromosome being covered at [?]10 reads per population in the 0.33x coverage datasets and at [?]40 reads for the 1x coverage datasets. The two demographic scenarios simulated by Lou et al. (2021) resulted in notably different numbers of segregating variants: 158,746 variants with MAF [?] 0.001 (of 245,412 total variants) in the high background F_{ST} scenario and 789,423 variants with MAF [?] 0.001 (1,209,625 total) in the low background F_{ST} scenario. PoolParty2 and angsd differed considerably in the dynamics of variant discovery, particularly identification of 'real', simulated variants. Both PoolParty2 and angsd recovered a larger number of sites in the datasets with greater coverage (all of which must have MAF [?] 0.001), but while the proportion of sites that were 'real' was similar across coverages for PoolParty2 ([?] 99.5%), this value changed with coverage for angsd (Table 1). In addition, only at the highest significance thresholds did angsd recover sites with similar 'real' proportions to PoolParty2, but then in lower numbers.

Across coverage depths, allele frequencies estimated by PPalign were always more accurate than those estimated by angsd, although only modestly (Figure 2, Supplemental Figure 2). Allele frequency estimates

generally improved with depth for both PoolParty2 and ANGSD, though more notably for angsd when depth was estimated by angsd rather than PoolParty2, and despite the large decrease in sites passing increasing depth thresholds. Indeed, angsd and PoolParty2 disagreed considerably about depth (correlation in depth estimates decreased) as the threshold for depth increased, implying that angsd's read filter is more stringent than that applied by PoolParty2 even with apparently similar parameter values. Nonetheless, correlation values for estimated to true allele frequencies ranged from approximately 80% to 90% for the lower coverage datasets and 90% to 95% for the higher coverage datasets, with PoolParty2 slightly higher than angsd in each case. It should also be noted that some diminishment in accuracy was expected due to sampling variance (which individuals and reads were sampled for each dataset), which determines the truly estimable allele frequencies regardless of the efficacy of each analysis, implying that each analysis is slightly closer to accurate than the reported values imply. This is reflected in the observation that correlations between allele frequencies estimated by PoolParty2 and angsd were always higher with each other than with 'real' allele frequencies in each case (87-98%; data not shown).

Both analytical suites were able to provide results which allowed visual identification of most if not all of the simulated outlier regions, particularly in the sliding window F_{ST}, Local Score, and linkage outlier results (Figure 3). Results provide by PoolParty2 and angsd for F_{ST}, sliding window F_{ST}, and FET (PoolParty2) or frequency test (angsd -doAssoc 1) were roughly equivalent (Supplemental Figures 3-6). The score and hybrid latent-score tests from angsd (-doAssoc 2 and 5) failed to produce any significant results. Both PoolParty2 and angsd had more difficulty in providing results that unambiguously identified outlier regions for the high background F_{ST} scenario at lower coverage, although even at higher coverage, outliers were less obvious than in either of the low background F_{ST} scenario datasets. The analyses that were designed to provide less ambiguous identification of outlier regions, Local Score and linkage outliers identified above twice the IQR, also exhibited efficacy moderated by demography and coverage (Table 2). In the case of Local Score, while peaks corresponding to the outlier regions were clearly visible in plots of smoothed FET significance, it was more difficult to determine significance thresholds that effectively identified the outlier regions with high background F_{ST}, though this test did not appear to be constrained significantly by coverage for the low background F_{ST}scenario. Moreover, in the high background F_{ST} scenario, there was no obvious inverse relationship between smoothing value (ξ) and power, as some replicates with higher ξ values identified more outlier regions at p [?] 0.05, though an inverse relationship was apparent in the low background F_{ST} scenario. Moreover, the precision of identified regions narrowed with increasing ξ values, as expected. In contrast, our identification of outliers using linkage was more constrained by coverage, with lower efficacy in lower coverage datasets regardless of background F_{ST}. Notably, the width of the region affected by hitchhiking was smaller with lower coverage in both scenarios, with similarly smaller outlier regions estimated by the Local Score analyses across ξ values at lower coverage in the low background F_{ST} scenario. Importantly, none of the analyses that considered broad range divergence or significance (windowed FST, Local Score, windowed linkage) identified any false positive outlier regions.

Application 2: Barcoded individuals in population samples of steelhead

After trimming, mapping, and quality filtering, PPalign provided 287 to 405 million mapped reads per sample, which allowed between 67.2 and 70.8% sampling of the genome at the minimum number of reads per sample (10), as revealed by PPstats (Supplemental Figure 7). The distribution of genomic extent across chromosomes was similar to other lcWGS analyses of the *O. mykiss* genome (e.g. Micheletti, Hess, Zendt, & Narum, 2018), indicating this pattern is a function of the library preparation technique used for all these samples or idiosyncrasy of this genome. Sequencing of indexed individuals allowed us to estimate that the mean coverage per individual ranged from 0 to 2.3 (median 0.23), to confirm that it was similar across populations (median 0.23, 0.26, 0.23 and standard deviation 0.39, 0.28, and 0.28 for Willamette River, Lewis River, and Skamania Hatchery, respectively), and to reduce bias in allele frequency estimates introduced by sampling variance across samples (normalize). After population-specific filters, PPanalyze examined 22,934,298 variants (22,832,805 [99.5%] in the chromosome scaffolds) with a suite of analyses. Density plots revealed that variants were sampled from across the genome, with a handful of areas of notable density. A principal components analysis made with loci with a maximum difference in allele frequencies below 0.9

(thus excluding the most divergent outlier loci), while unremarkable, confirmed that the primary axis, which explained ˜86% of the variance in the data, did*not* segregate the Skamania hatchery sample from the natural origin samples, implying that outlier regions related to the main contrast (hatchery vs. natural) would not be confounded by background population structure. Raw PPanalyze output revealed many small regions of strong genomic divergence, while 51 separate regions were identified as significant at p [?] 0.05 across ξ values and replicates in Local Score analyses (Figure 4, Table 3, Supplemental Table 1). The two most significant (highest local score) regions were the region of chr. 28 containing the genes *GREB1L* and*ROCK1* and the region of chr. 25 containing the gene *SIX6* , which have been previously found associated with migration timing and age at maturity in steelhead and other salmonids, respectively (e.g. Willis et al., 2020). There were also many additional regions whose potential association with migration phenology, age at maturity, or domestication (adaptation to hatchery production) could be explored further. For example, a region of chromosome 20 that was consistently recovered in the Local Score analyses contained two protein coding genes: ATP-citrate lysase (synthase), or ACLY, and, dnaJ homolog subfamily C member 7, or DNAJC7. ACLY is a ubiquitous cytosolic enzyme positioned at the intersection of nutrients catabolism and cholesterol and fatty acid biosynthesis, and DNAJC7 is a member of the heat shock protein 40 family and acts as co-chaperone regulating the molecular chaperones HSP70 and HSP90 in folding of steroid receptors, such as the glucocorticoid receptor and the progesterone receptor. Notably, identification of linkage outliers for these three chromosomes identified the same regions, but in the case of chromosomes 25 and 28, also identified other regions that Local Score did not, presumably because, while they exhibit strong linkage across all samples, these regions were not consistently divergent between the hatchery and natural origin samples.

Discussion

Bioinformatic suites for low coverage whole genome sequencing data

PoolParty2 is a bioinformatic pipeline that was designed for the utilization of pooled or individually barcoded low coverage whole genome sequencing (lcWGS) data with a high-quality reference genome to perform genome-wide genetic association analysis with binary phenotypic traits or environmental variables. Since its original presentation, it has seen numerous updates, though most of these will be invisible to the user since their effect is to provide greater efficiency or stability. For example, the scripts that implement the Local Score analysis of Fariello et al. (2017) have been improved to make them more robust to variations in SNP and allele frequency count and distribution and provide more accurate reports with respect to outlier region margins. The major workhorse of the pipeline, the PPalign module, has received the most updates since the pipeline's first presentation. Notably, this module is now able to call variants in parallel (across multiple computational threads) by dividing chromosomes or scaffolds into groups based on a user-specified number of threads, which substantially reduces processing time for large datasets. Another notable update includes the user's ability to limit the number of individuals to normalize simultaneously, which reduces RAM memory requirements in memory-limited or shared computing environments, with the proviso that this also causes the duration of processing time for normalization to increase linearly.

These improvements to the PPalign module of PoolParty2 are important since, while there are myriad analyses available for the discovery of large effect loci, few bioinformatic suites are designed to convey lcWGS data from sequencing output, via quality assurance analyses, through to analytical results. angsd, another useful bioinformatic suite designed for barcoded but not pooled lcWGS data, requires read alignments as input, and provides additional functionalities that PoolParty2 does not, making these bioinformatic suites strong complements. In particular, angsd can estimate individual genotype likelihoods or posterior probabilities. For analyses where individual phenotype data are important, such as random forest analyses (e.g. Hess, Zendt, Matala, & Narum, 2016), genotype likelihoods would be necessary. Moreover, as we demonstrate here, these genotype likelihoods allow the estimation of linkage disequilibrium in specified windows, and which can be used to identify outlier regions under putative selection, though, as our empirical results portray, most effectively in combination with analyses that consider divergence among focal sets of samples. The analyses available through the PPanalyze module of PoolParty2 have most often been applied for the discovery of one or a few genomic regions with large, independent or simple-interaction effects on traits (e.g.

direct epistasis), though, as results from our empirical data also portray, these analyses, with appropriate corroboration, may identify many loci with smaller effects on polygenic traits. Loci contributing to polygenic traits with continuous variation and low penetrance, however, will be challenging to discover with analyses that rely mainly on allele frequency, and their discovery will likely depend on identifying combinatorial association of non-contiguous genotypes, for which genotype likelihoods may be better suited. It should be noted, however, that genotype likelihoods may only be sufficiently informative for some analyses if coverage is high enough for each individual, and the decision of how to multiplex individuals or how deeply to sequence, made early in the data preparation phase, must be done with respect to the types of analyses that will be needed (Lou et al., 2021; Paril et al., 2022).

Using simulated data, we demonstrated that angsd and PoolParty2 both provide remarkably precise estimates of allele frequency considering the degree of individual coverage constraint and variation with which we challenged them, with PoolParty2 exhibiting a moderate edge in identifying true sites with fewer artifacts and accurately estimating their allele frequencies. Both bioinformatic suites also produce very low rates of false positive outliers, at least when data are examined on a site-aggregate (windowed or serial) basis. A number of the per-site divergence analyses are roughly equivalent between bioinformatic suites, including estimates of F_{ST}, sliding window F_{ST}, and basic tests of allele frequency proportions, though we note that to estimate site frequency spectra in order to calculate F_{ST} with angsd, each population must be analyzed separately, and if the user desires to utilize the empirical major/minor alleles rather than reference/alternate or ancestral/derived, a prior run with all data must be made. Beyond these, however, the Local Score analyses (Fariello et al., 2017) and estimate of linkage outliers from ngsLD (Fox et al., 2019) provide powerful and complementary means to identify true outlier regions. Moreover, these analyses can be run in parallel by supplying filtered BAM files and a list of variant sites to angsd to generate genotype likelihoods for linkage estimation with ngsLD while simultaneously running PPanalyze and Local Score analyses. We provide the code to demonstrate how we executed this for our simulated and empirical data (Supplemental File 1).

However, results from the analysis of simulated data also indicate that, while both angsd and PoolParty2 are effective at identifying segregating variants, estimating allele frequencies, and identifying outlier loci using lcWGS data, the particular evolutionary context and genomic environment of putatively adaptive loci may constrain the efficacy of any bioinformatic suite at a given depth of coverage. For example, we observed that while both analytical suites provided results that correctly portrayed the presence and location of outlier loci in each of the simulated datasets, peaks in the higher background divergence scenarios were more difficult to discern visually and an analysis designed to objectively discriminate outlier loci from background rates of variation, Local Score, identified fewer significant regions. While outlier identification using linkage was more effective in high background divergence scenarios, and coordination of this approach with the other analyses increases flexibility of the combined toolkit, this analysis was less effective with lower coverage regardless of evolutionary context. Although we did not investigate it directly here because of the nature of the simulated data, one strategy to increase power in challenging scenarios is through the use of paired sample replicates, i.e., samples of the same phenotype or across the same environmental axes from multiple populations (e.g. Lotterhos & Whitlock, 2015). Analyses that emphasize repeated differences across these replicates, such as the Cochran–Mantel–Haenszel test available in PoolParty2, increase the power to identify regions associated with the focal contrasts while minimizing the influence of background variation (Cochran, 1954; Mantel & Haenszel, 1959). Our recommendation based on these observations is for researchers to consider existing information about the demographic and genetic context of the traits and populations under study, and to carefully arrange experimental design to maximize power for any given scenario (Lou et al., 2021).

One underappreciated phenomenon in lcWGS are technical artifacts commonly known as batch effects (Leek et al., 2010; Lou & Therkildsen, 2022). These occur when idiosyncrasies of the library and sequencing process introduce artifacts for samples processed together that are later misinterpreted as true biological differences between groups under analysis, and may have a number of contributing causes in lcWGS data (Lou & Therkildsen, 2022). While batch effects are certainly not exclusive to lcWGS, barcoded and especially pooled lcWGS data that are prepared group-wise may be subject to them. Importantly, both PoolParty2 and angsd have a number of utilities included that help eliminate some batch effects, including read trimming, mapping

and SNP quality filtering, minor allele frequency thresholds, and read and individual observation minima. Moreover, with any bioinformatic suite it is wise that users conduct tests to determine if results are sensitive to various trimming and filtering parameters. The alignment filtering utilities provided with angsd appear to be more stringent, even with the same parameters, than what is currently applied in PoolParty2, as evidenced by diminished correlation in allele frequencies for sites with higher depth as estimated by PoolParty2 but not for angsd. Additional variant filters are accessible in angsd, including those that examine strand balance, and though not built-in, these same filters can be applied to the variants discovered by PoolParty2 through VCF filtering tools such as bcftools or vcflib (Danecek et al., 2021; Garrison, Kronenberg, Dawson, Pedersen, & Prins, 2021), as described in our tutorials. Aside from these filters, one important means of controlling batch effects is to distribute barcoded samples from different analytical groups (e.g., populations) among sequencing runs, and ideally library preparation groups, since this reduces the chance that technical artifacts will be identified as group-wise differences (O'Leary, Puritz, Willis, Hollenbeck, & Portnoy, 2018). In addition, another strategy to minimize the influence of batch effects and other false positives is through the use of paired sample replicates as described above: analyses that utilize paired replicates reduce the chance that artifacts owing to any technical pair will be identified as consistent, significant differences (Cochran, 1954; Mantel & Haenszel, 1959).

Discovery of large effect loci in non-model organisms

We successfully apply our integrated pipeline to a common data arrangment of low coverage whole genome sequencing (lcWGS) data from steelhead trout (*Oncorhynchus mykiss*) to identify loci strongly divergent among populations with distinct histories and putatively of strong effect on traits under selection in these groups. This species exhibits an impressive array of life history diversity even among salmonids, including notable and important variation in migration phenology (run timing), age at maturity (age at first reproduction), propensity for residency or anadromy, precocial sexual maturation, and iteroparity (repeat spawning) (Busby et al., 1999; Carlson & Seamons, 2008; Quinn, Seamons, Vollestad, & Duffy, 2011). Many of these traits are highly variable both among and within phylogeographic lineages and local populations (Busby et al., 1999), are part of the portfolio of life history diversity that assist populations in persisting through natural and anthropogenic environmental challenges (Quinn, McGinnity, & Reed, 2016), and are among the outward physical traits used by managers to estimate the contribution of distinct stocks to mixed-stock fisheries (Hess et al., 2021).

The diversity of life history ensembles among regional and local stocks of steelhead reflects a multifaceted history of selection, and upon these effects anthropogenic forces have applied new challenges. One of the most direct human impacts on steelhead are use of hatcheries to mitigate for the loss of fish from dams, overfishing, habitat degradation, and other human actions. While hatcheries have begun to shift towards inclusion of natural origin (NOR) returns in broodstock recruitment to avoid 'domestication' selection, these are still the minority, and indeed some operations have taken the opposite approach, actively choosing hatchery returns with characteristics desirable for hatchery managers or fishers, with many consequences to genomic variation. While the typical application of our pipeline for discovering loci under selection relies upon groups with known phenotypic differences, discrete phenotypes with heritable bases may be overall uncommon in non-model organisms, at least relative to opportunities for examining populations arrayed across replicate environmental axes (e.g. Lotterhos & Whitlock, 2015). To reflect this, we chose to examine genomic divergence in a hatchery stock with a notable history of hatchery selection without *a priori* designation other than group membership, as would be the case for most landscape-level genome scans. However, in this case the Skamania Hatchery stock is well known for its phenotypic ensemble, its history of hatchery origin recruitment and artificial selection, and the extent to which this stock has been outplanted or strayed into numerous Columbia Basin tributaries. Not surprisingly, we observed strong divergence in two regions known to be associated with early migration timing and age at maturity (ocean duration) on chromosomes 28 and 25, respectively. However, we also saw significant divergences in regions that contained several dozen additional genes, including a prominent region on chromosome 20 containing two genes important in metabolism and cellular signaling. This region on chromosome 20 is also known to harbor a structural inversion in this species (Pearse et al., 2019), which may have been involved in artificial

selection of the hatchery stock. Importantly, these regions were emphasized as outliers both by the analyses we applied that considered only global linkage patterns as well as the analyses that considered contrast-specific divergence while controlling for linkage. However, the linkage-only analysis identified additional regions that did not appear to exhibit strong divergence across our contrast of interest (hatchery vs. natural origin), demonstrating the importance of corroborating outliers identified from any single analysis. While we decline to hypothesize how these additional genes may be involved in domestication, these results demonstrate the utility of applying our pipeline to landscape-level samples and the efficacy and complementarity of the PoolParty and angsd pipelines for analyzing lcWGS data.

Acknowledgements

We thank colleagues at the Columbia River Inter-Tribal Fish Commission (CRITFC) in Portland, OR and Hagerman, ID for facilitating the current dataset and the associate editor and two anonymous reviewers for comments on a previous version of this manuscript. We also greatly appreciate the assistance by Nicolas R. Lou and Nina Overgaard Therkildsen for providing and explaining the simulated sequence data. Samples were collected by CRITFC staff or graciously provided by the Oregon and Washington Departments of Fisheries and Wildlife. Funding for this project was contributed by Bonneville Power Administration (Grant no. 2008-907-00), by the NSF Idaho EPSCoR Program, and by the National Science Foundation (award no. OIA-1757324).

Data availability

Data referenced in this manuscript are available from NBCI Short Read Archive (PRJNA854899). The pipeline used for bioinformatic processing of these data, PoolParty, is available from Github (https://github.com/stuartwillis/poolparty).

References Cited

Aguirre-Ramirez, E., Velasco-Cuervo, S., & Toro-Perea, N. (2021). Genomic traces of the fruit fly Anastrepha obliqua associated with its polyphagous nature. *Insects* , *12* (12). doi: 10.3390/insects12121116

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecol & evol genomics.pdf. *Nature Reviews Genetics* , *17* .

Ayerst, J. D. (1976). The Role of Hatcheries in Rebuilding Stcelhead Runs of the Columbia River System. *American Fisheries Society Special Publication: Proceedings of a Symposium Held in Vancouver, Washington, March 5-6, 1976* , 84–88.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* ,*3* (e3376).

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics* ,*186* (1). doi: 10.1534/genetics.110.117275

Busby, P., Wainwright, T., & Bryant, G. (1999). Status Review of Steelhead from Washington, Idaho, Oregon, and California.*Sustainable Fisheries Management* . doi: 10.1201/9781439822678.ch11

Carlson, S. M., & Seamons, T. R. (2008). SYNTHESIS: A review of quantitative genetic components of fitness in salmonids: implications for adaptation to future change. *Evolutionary Applications* ,*1* (2), 222–238. doi: 10.1111/j.1752-4571.2008.00025.x

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ 2 Tests. *Biometrics* , *10* (4). doi: 10.2307/3001616

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* , *10* (2). doi: 10.1093/gigascience/giab008

Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C., ... SanCristobal, M. (2017). Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: The local score approach. *Molecular Ecology* ,*26* (14). doi: 10.1111/mec.14141

Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). NgsLD: Evaluating linkage disequilibrium using genotype likelihoods.*Bioinformatics* , *35* (19). doi: 10.1093/bioinformatics/btz200

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* , *186* (1). doi: 10.1534/genetics.110.114397

Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2021). Vcflib and tools for processing the VCF variant call format. *BioRxiv* .

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* , *195* (1). doi: 10.1534/genetics.113.152462

Hess, J. E., Collins, E. E., Harmon, S. A., Horn, R. L., Koch, I. J., Stephenson, J., ... Narum, S. R. (2021). *GENETIC ASSESSMENT OF COLUMBIA RIVER STOCKS: 1/1/2020 - 12/31/2020 Annual Report, 2008-907-00* .

Hess, J. E., Zendt, J. S., Matala, A. R., & Narum, S. R. (2016). Genetic basis of adult migration timing in anadromous steelhead discovered through multivariate association testing. *Proceedings of the Royal Society B: Biological Sciences* , *283* (1830). doi: 10.1098/rspb.2015.3064

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., ... Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *American Naturalist* . doi: 10.1086/688018

Horn, R. L., Kamphaus, C., Murdoch, K., & Narum, S. R. (2020). Detecting genomic variation underlying phenotypic characteristics of reintroduced Coho salmon (Oncorhynchus kisutch). *Conservation Genetics* , *21* (6). doi: 10.1007/s10592-020-01307-0

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* ,*15* (1). doi: 10.1186/s12859-014-0356-4

Lamichhaney, S., Barrio, A. M., Rafati, N., Sundström, G., Rubin, C. J., Gilbert, E. R., ... Andersson, L. (2012). Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proceedings of the National Academy of Sciences of the United States of America* , *109* (47). doi: 10.1073/pnas.1216128109

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data.*Nature Reviews Genetics* , Vol. 11. doi: 10.1038/nrg2825

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology* . doi: 10.1111/mec.13100

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology* , *30* (23). doi: 10.1111/mec.16077

Lou, R. N., & Therkildsen, N. O. (2022). Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. *Molecular Ecology Resources* ,*22* (5). doi: 10.1111/1755-0998.13559

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources* , *17* (2). doi: 10.1111/1755-0998.12635

Lyu, G., Feng, C., Zhu, S., Ren, S., Dang, W., Irwin, D. M., ... Zhang, S. (2021). Whole Genome Sequencing Reveals Signatures for Artificial Selection for Different Sizes in Japanese Primitive Dog Breeds. *Frontiers in Genetics* , *12* . doi: 10.3389/fgene.2021.671686

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* , *22* (4). doi: 10.1093/jnci/22.4.719

Micheletti, S. J., Hess, J. E., Zendt, J. S., & Narum, S. R. (2018). Selection at a genomic region of major effect is responsible for evolution of complex life histories in anadromous steelhead 06 Biological Sciences 0604 Genetics. *BMC Evolutionary Biology* ,*18* (1). doi: 10.1186/s12862-018-1255-5

Micheletti, S. J., & Narum, S. R. (2018). Utility of pooled sequencing for association mapping in nonmodel organisms. *Molecular Ecology Resources* . doi: 10.1111/1755-0998.12784

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists.*Molecular Ecology* , *0* (ja). doi: 10.1111/mec.14792

Paril, J. F., Balding, D. J., & Fournier-Level, A. (2022). Optimizing sampling design and sequencing strategy for the genomic analysis of quantitative traits in natural populations. *Molecular Ecology Resources* , *22* (1). doi: 10.1111/1755-0998.13458

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., ... Lien, S. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature Ecology and Evolution* , *3* (12). doi: 10.1038/s41559-019-1044-6

Puritz, J. B., Matz, M. V, Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology* , *23* (24), 5937–5942.

Quinn, T. P., McGinnity, P., & Reed, T. E. (2016). The paradox of "premature migration" by adult anadromous salmonid fishes: Patterns and hypotheses. *Canadian Journal of Fisheries and Aquatic Sciences* , Vol. 73, pp. 1015–1030. doi: 10.1139/cjfas-2015-0345

Quinn, T. P., Seamons, T. R., Vollestad, L. A., & Duffy, E. (2011). Effects of growth and reproductive history on the egg size-fecundity trade-off in steelhead. *Transactions of the American Fisheries Society* . doi: 10.1080/00028487.2010.550244

Ren, S., Lyu, G., Irwin, D. M., Liu, X., Feng, C., Luo, R., ... Wang, Z. (2021). Pooled Sequencing Analysis of Geese (Anser cygnoides) Reveals Genomic Variations Associated With Feather Color.*Frontiers in Genetics* , *12* . doi: 10.3389/fgene.2021.650013

Schlötterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals-mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics* , *15* (11). doi: 10.1038/nrg3803

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species.*Molecular Ecology Resources* , *17* (2). doi: 10.1111/1755-0998.12593

Tiffin, P., & Ross-Ibarra, J. (2014). Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology and Evolution* , Vol. 29. doi: 10.1016/j.tree.2014.10.004

Willis, S. C., Hess, J. E., Fryer, J. K., Whiteaker, J. M., Brun, C., Gerstenberger, R., & Narum, S. R. (2020). Steelhead (Oncorhynchus mykiss) lineages and sexes show variable patterns of association of adult migration timing and age-at-maturity traits with two genomic regions. *Evolutionary Applications* , *13* (10), 2836–2856. doi: 10.1111/eva.13088

Willis, S. C., Hollenbeck, C. M., Puritz, J. B., & Portnoy, D. S. (2022). Genetic recruitment patterns are patchy and spatiotemporally unpredictable in a deep-water snapper (Lutjanus vivanus) sampled in fished and protected areas of western Puerto Rico. *Conservation Genetics* , *23* (3). doi: 10.1007/s10592-021-01426-2

Zhu, Y., Bergland, A. O., González, J., & Petrov, D. A. (2012). Empirical validation of pooled whole genome population re-sequencing in Drosophila melanogaster. *PLoS ONE* , *7* (7). doi: 10.1371/journal.pone.0041901

Tables

Table 1. Results of SNP calling and filtering by PoolParty2 and angsd with different modules and filtering thresholds in two different demographic scenarios (high and low background F_{ST}) and coverage levels (0.33x and 1x). The true number of simulated sites for the high background F_{ST} scenarios with minor allele frequency >=0.001 was 158,746 SNPs and for low background F_{ST} was 789,423 SNPs.

Table 2. Outlier regions identified from simulated data using windowed mean linkage and regions identified with Local Score analysis of significance from Fisher's Exact tests. For each outlier region identified, the width in Kbp is reported. For windowed linkage analysis, the mean and maximum windowed mean of linkage is listed, while for Local Score analyses across quantiles of significance as smoothing parameter values, ξ, the number of replicates out of three in which a region was significant are listed.

Table 3. Outlier regions identified in select chromosomes from empirical steelhead trout data using windowed mean linkage and regions identified with Local Score analysis of significance from Fisher's Exact tests. For each outlier region identified, the beginning and end of the regions (in Mbp), width (in Kbp), and, for regions identified with windowed linkage analysis, the mean and maximum windowed mean of linkage, are reported. Local score results are reported across quantiles of significance as smoothing parameter values, ξ.

Supplemental Table 1. Outlier regions identified in across all chromosomes from empirical steelhead trout data using Local Score analysis of significance from Fisher's Exact tests. For each outlier region identified, the beginning and end, width, local score of the peak, and number of replicates out of 3 in which the region was significant, are reported. Regions are reported across quantiles of significance as smoothing parameter values, ξ.

Figures

Figure 1. Graphical representation of the normalization process of allele frequency estimates for variance in read depth possible with the incorporation on barcoded individual samples.

Figure 2. Comparing estimated allele frequencies with true allele frequencies across depth cutoffs for data simulated with high background F_{ST} and low coverage (0.33x). Left axis: This red and blue lines show correlation for PoolParty2 estimated frequencies for each population for normalized (solid) and non-normalized (dashed) data. Orange and green lines show correlation with angsd estimated frequencies. Black line shows correlation in depths estimated by angsd and PoolParty2. Right axis: thick purple line shows number of sites passing depth thresholds. Left and right panels shows depths reported by PoolParty2 and angsd, respectively.

Figure 3. Manhattan plots of divergence estimated for variant sites identified by PoolParty2 in simulated data in scenarios with high (A,B) or low (C,D) background F_{ST} and 0.33x (A,C) or 1x coverage (B,D). For each dataset, the divergence (F_{ST}) estimated in 100Kbp sliding windows (left column), Local Score estimated with ξ reflecting the 80^{th} quantile of Fisher's Exact Test significance values (middle column), and mean linkage in 100Kbp windows (right column) are shown. Red dots indicate series of 10+ windows of mean linkage above 2x the interquartile range. Blue dots in each panel reflect the positions (but not values) of loci simulated under selection.

Figure 4. Manhattan plots of divergence and linkage among hatchery and natural origin population samples of steelhead. A) Individual site F_{ST} corrected for sample size (red line indicates F_{ST} of 0.5) B) Mean F_{ST}in sliding windows of 100Kb across the genome. C) -Log10 significance (p) values from individual site Fisher's Exact tests D-G) Local score analysis of significance values from Fisher's Exact tests for the 80^{th}, 90^{th}, 95^{th}, and 99^{th} quantile of significance values as ξ (red line indicates FDR of 0.05); H-J) Mean linkage in 100Kbp windows for chromosomes 20, 25, and 28 (red dots indicate series of 20+ windows of mean linkage above 2x the interquartile range).

Supplemental Figure 1. True F_{ST} for loci simulated under demographic scenarios producing A) high background F_{ST}(low N_E, low migration) or B) low background F_{ST} (high N_E, high migration). Black dots indicated the F_{ST} of positions simulated under selection.

Supplemental Figure 2. Comparing estimated allele frequencies with true allele frequencies across depth cutoffs for data simulated with high background F_{ST} and low coverage (0.33x). Left axis: This red and blue lines show correlation for PoolParty2 estimated frequencies for each population for normalized (solid) and non-normalized (dashed) data. Orange and green lines show correlation with angsd estimated frequencies. Black line shows correlation in depths estimated by angsd and PoolParty2. Right axis: thick blue line shows number of sites passing depth thresholds. Left and right panels shows depths reported by PoolParty2 and angsd, respectively. A) high background F_{ST}, 0.33x coverage; B) high background F_{ST}, 1x coverage; C) low background F_{ST}, 0.33x coverage; D) low background F_{ST}, 1x coverage

Supplemental Figure 3. Manhattan plot of results from analysis of 0.33x coverage data simulated with high background F_{ST}. A) F_{ST} estimated by PoolParty; B) 100Kbp sliding windows of F_{ST} estimated by PoolParty2; C) -Log10 significance (p) values from Fisher's Exact test of allele frequencies between populations calculated by PoolParty2; D) F_{ST} estimated by angsd; E) 100Kbp sliding windows of F_{ST} estimated by angsd; F) Likelihood ratio of frequency differences between population calculated by angsd; G) Mean linkeage in 100Kbp windows estimated for sites identified by PoolParty2, genotype likelihood estimated by angsd, and linkage calculated with ngsLD, with series of 10+ windows with mean linkage above 2x the interquartile range in red; H-L) Local score values for ξ values reflecting the 70^{th}, 80^{th}, 90^{th}, 95^{th}, and 99^{th} quantile of significance values from Fisher's Exact test. Blue dots in each panel reflect the positions (but not values) of loci simulated under selection.

Supplemental Figure 4. Manhattan plot of results from analysis of 1x coverage data simulated with high background F_{ST}. A) F_{ST} estimated by PoolParty2; B) 100Kbp sliding windows of F_{ST} estimated by PoolParty2; C) -Log10 significance (p) values from Fisher's Exact test of allele frequencies between populations calculated by PoolParty2; D) F_{ST} estimated by angsd; E) 100Kbp sliding windows of F_{ST} estimated by angsd; F) Likelihood ratio of frequency differences between population calculated by angsd; G) Mean linkeage in 100Kbp windows estimated for sites identified by PoolParty2, genotype likelihood estimated by angsd, and linkage calculated with ngsLD, with series of 10+ windows with mean linkage above 2x the interquartile range in red; H-L) Local score values for ξ values reflecting the 70^{th}, 80^{th}, 90^{th}, 95^{th}, and 99^{th} quantile of significance values from Fisher's Exact test. Blue dots in each panel reflect the positions (but not values) of loci simulated under selection.

Supplemental Figure 5. Manhattan plot of results from analysis of 0.33x coverage data simulated with low background F_{ST}. A) F_{ST} estimated by PoolParty2; B) 100Kbp sliding windows of F_{ST} estimated by PoolParty2; C) -Log10 significance (p) values from Fisher's Exact test of allele frequencies between populations calculated by PoolParty2; D) F_{ST} estimated by angsd; E) 100Kbp sliding windows of F_{ST} estimated by angsd; F) Likelihood ratio of frequency differences between population calculated by angsd; G) Mean linkeage in 100Kbp windows estimated for sites identified by PoolParty2, genotype likelihood estimated by angsd, and linkage calculated with ngsLD, with series of 10+ windows with mean linkage above 2x the interquartile range in red; H-L) Local score values for ξ values reflecting the 70^{th}, 80^{th}, 90^{th}, 95^{th}, and 99^{th} quantile of significance values from Fisher's Exact test. Blue dots in each panel reflect the positions (but not values) of loci simulated under selection.

Supplemental Figure 6. Manhattan plot of results from analysis of 1x coverage data simulated with low background F_{ST}. A) F_{ST} estimated by PoolParty; B) 100Kbp sliding windows of F_{ST} estimated by PoolParty2; C) -Log10 significance (p) values from Fisher's Exact test of allele frequencies between populations calculated by PoolParty2; D) F_{ST} estimated by angsd; E) 100Kbp sliding windows of F_{ST} estimated by angsd; F) Likelihood ratio of frequency differences between population calculated by angsd; G) Mean linkeage in 100Kbp windows estimated for sites identified by PoolParty2, genotype likelihood estimated by angsd, and linkage calculated with ngsLD, with series of 10+ windows with mean linkage above 2x the interquartile range in red; H-L) Local score values for ξ values reflecting the 70^{th}, 80^{th}, 90^{th}, 95^{th}, and 99^{th} quantile of

significance values from Fisher's Exact test. Blue dots in each panel reflect the positions (but not values) of loci simulated under selection.

Supplemental Figure 7. Quality control analyses for steelhead from Skamania hatchery or two natural origin populations, Willamette River (Eagle Creek) or Lewis River. A) Proportion of the genome covered at various minimum coverage limits. B) Mean coverage across sites for each population sample (summed across individuals) C) Principal components analysis of allele frequencies for variants with a maximum allele frequency difference [?] 0.9. D) Mean proportion across population samples of each chromosome. E) Number of SNPs per 10 Kb across the genome (SNP density).

		high - 0.33x		high - 1x		low - 0.33x		low - 1x	
		total sites	% real	total sites	% real	total sites	% real	total sites	% real
Poolparty	PPalign[a]	70,067	99.5%	91,509	99.8%	345,223	99.8%	452,827	99.8%
	PPanalyze[b]	68,292	99.5%	86,036	99.8%	334,664	99.8%	422,029	99.8%
ANGSD[c]	$p < 0.0$	527,762	17.2%	920,208	12.0%	864,530	51.4%	1,323,894	41.1%
	$p < 0.01$	94,104	69.6%	209,617	43.5%	344,606	92.6%	411,213	98.1%
	$p < 0.001$	57,027	98.7%	127,423	67.5%	275,355	99.8%	373,603	99.6%
	10 reads + 3 inds/pop[d]	56,060	98.7%	127,415	67.5%	271,175	99.8%	373,473	99.6%
Poolparty to ANGSD	20 global reads[e]	67,233	99.5%	86,011	99.8%	328,654	99.8%	421,737	99.8%
	10 reads + 3 inds/pop[e]	66,168	99.5%	85,987	99.8%	323,806	99.8%	421,568	99.8%

a min. 20 global reads, MAF 0.001
b sites specified by PPalign; min. 3 ind+10 reads/pop
c min. 20 global reads, MAF 0.001, min. 6 global individuals
d filtering of sites specified by ANGSD p<0.001
e no additional p-value or MAF filter

Background F_{ST}	coverage	outlier (Mbp)	Linkage Disequilibrium Windows		Local Score ξ Quantiles				
					70%	80%	90%	95%	99%
			width	LD mean (max)	width (observation across replicates)				
high	0.33x	2.5	245	0.103 (0.116)	--	--	--	93 (1)	--
		5	275	0.097 (0.112)	--	--	--	--	--
		7.5	--	--	--	--	--	--	--
		15	195	0.088 (0.093)	--	--	--	--	--
		17.5	203	0.090 (0.097)	--	--	--	--	--
		22.5	194	0.089 (0.092)	--	648 (3)	249 (2)	74 (1)	--
		25	169	0.089 (0.092)	--	--	--	--	--
high	1x	2.5	420	0.078 (0.108)	--	--	--	134 (1)	52 (1)
		5	490	0.073 (0.101)	--	--	--	--	--
		7.5	224	0.063 (0.065)	--	--	--	--	--
		15	434	0.067 (0.074)	--	--	--	--	--
		17.5	408	0.069 (0.082)	--	--	--	--	--
		22.5	314	0.071 (0.083)	--	--	--	--	--
		25	305	0.070 (0.077)	--	--	--	--	--
low	0.33x	2.5	220	0.073 (0.076)	547 (3)	196 (3)	81 (3)	22 (1)	--
		5	--	--	--	61 (3)	--	5 (1)	--
		7.5	180	0.071 (0.073)	558 (3)	165 (3)	69 (3)	32 (3)	13 (3)
		10	--	--	472 (3)	134 (3)	--	--	--
		17.5	160	0.071 (0.071)	450 (3)	118 (3)	48 (2)	29 (1)	--
		22.5	--	--	380 (3)	99 (3)	46 (3)	25 (1)	--
		25	--	--	360 (3)	99 (3)	54 (1)	27 (1)	--
low	1x	2.5	275	0.051 (0.058)	1,025 (3)	323 (3)	134 (3)	66 (3)	10 (3)
		5	200	0.048 (0.049)	779 (3)	--	39 (2)	--	--
		7.5	195	0.050 (0.051)	752 (3)	243 (3)	101 (3)	55 (3)	31 (3)
		10	225	0.049 (0.051)	579 (3)	175 (3)	63 (2)	--	--
		17.5	215	0.048 (0.050)	960 (3)	219 (3)	71 (3)	--	--
		22.5	250	0.049 (0.052)	745 (3)	158 (3)	73 (3)	35 (3)	20 (3)
		25	200	0.049 (0.051)	771 (3)	193 (3)	80 (3)	--	--

			start (Mbp)	end (Mbp)	width (Kbp)	LD mean (max)
Omy 20	LS Quantile (ξ)	80%	16.07	22.98	6,907	--
		90%	17.34	19.50	2,166	--
		95%	18.83	19.05	211.8	--
		99%	18.91	18.92	6.3	--
	LD Outliers		17.70	18.70	1,100	0.087 (0.327)
			18.72	19.33	701.9	0.155 (0.261)
Omy 25	LS Quantile (ξ)	80%	22.31	31.58	9,272	--
		90%	22.87	23.66	797.1	--
		95%	22.89	23.21	317.7	--
		99%	22.92	22.98	51.2	--
	LD Outliers		22.31	22.44	224.9	0.086 (0.095)
			22.80	23.05	354.9	0.091 (0.116)
Omy 28	LS Quantile (ξ)	80%	1.69	42.93	41,243	--
		90%	11.90	15.50	3,608	--
		95%	12.02	13.17	1,158	--
		99%	12.03	12.46	432.6	--
	LD Outliers		11.94	12.12	284.8	0.137 (0.206)