Received: 27 November 2022 | Revised: 29 September 2023 | Accepted: 19 October 2023

DOI: 10.1111/1755-0998.13888

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

PoolParty2: An integrated pipeline for analysing 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] | Shawn Narum[1]

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

[2]Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada

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

Correspondence
Stuart Willis, Hagerman Genetics Lab, Columbia River Inter-Tribal Fish Commission, Hagerman, ID, USA.
Email: swillis@critfc.org

Funding information
Bonneville Power Administration, Grant/Award Number: 2008-907-00; National Science Foundation, Grant/Award Number: OIA-1757324

Handling Editor: Sarah Fitzpatrick

Abstract

Whole-genome sequencing data allow survey of variation from across the genome, reducing the constraint of balancing genome sub-sampling with estimating 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 analyse 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 harbouring 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 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 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. Thus, we not only demonstrate the utility of PoolParty2 for genomic studies that combine sequencing data from multiple individuals, but also illustrate how it compliments other bioinformatics resources such as ANGSD.

KEYWORDS
bioinformatics, genome-wide association study, genomics, non-model organism

1 | 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 endeavour 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 et al., 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, demands for statistically robust sample sizes have become more ardent (Schlötterer et al., 2014). As a result, geneticists are still faced with the task of determining the appropriate compromise between the 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 et al., 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 at higher coverage, creating 'reduced representation' libraries wherein sequencing coverage is spread across 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 et al., 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 et al., 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 100 kb, 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).

In contrast, many studies rely largely on methods that utilize allele frequencies rather than individual genotypes to address primary questions. Because of sampling variance, sampling more individuals at low coverage 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 et al., 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 with 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 1–2× 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 analysing allele frequency and genotype probability data is expanding, there remain few user-friendly 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 et al., 2021; Horn et al., 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-genotyping method wherein each individual, regardless of total reads, is allowed to contribute at most two alleles per locus to allele frequencies, 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 et al., 2014). ANGSD requires mapped read alignments produced by other tools as input, while PoolParty takes sequence read files as input, performs sequence cleaning and mapping to a reference genome, and produces numerous assurance reports regarding sequence and mapping quality. PoolParty also facilitates several analyses to identify regions of significant genomic divergence between samples, making PoolParty and ANGSD complementary bioinformatic tools for lcWGS data analyses.

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 that were 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

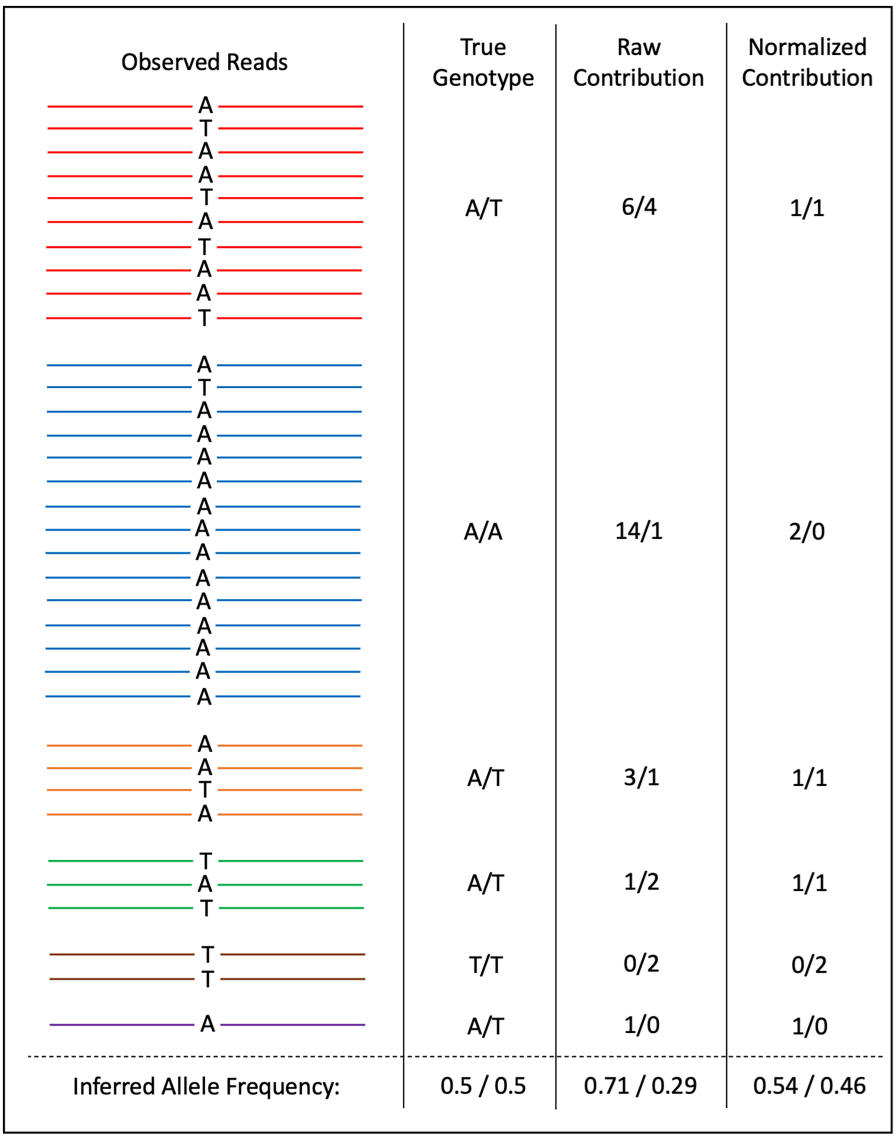

FIGURE 1 Graphical representation of normalization variance in read depth for allele frequency calculation implemented by PoolParty2, provided barcoded individual samples. Here, sampled reads are shown on the left, with each colour representing a different barcoded individual, and A or T representing the nucleotide variant in each read. For each individual, the true genotype, pre-and post-normalization contribution to allele frequencies is shown in the right three columns, respectively, and beneath each, the calculated allele frequency. The threshold for heterozygotes is set as a minimum ratio of 10:1 alternate alleles for a given SNP (e.g. individual #2).

boundaries regions affected by selection. Then, in an empirical example 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.

2 | METHODS

2.1 | 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 (FETs), 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 >1 Tb 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.

2.2 | 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 30 Mb 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 (Figure S1). Sequence data from the simulated chromosomes, generated to reflect Illumina-style paired-end reads including sequencing errors, reflected 8× 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} and higher sequence coverage; (c) low background and 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, from 160 each) and down-sampled the simulated sequence data (from 8×) to reflect the median individual coverage in our empirical data (~0.33×) as well as a common sequencing target for lcWGS studies (1×). 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 8× simulated data. For the 1× 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 0.33× and 1× 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 ≥50 bp), 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 100 kbp windows in 5 kbp sets, as well as applying Fisher's exact tests (FET) for differences in allele frequency between the two populations (via PoPoolations). 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 70th, 80th, 90th, 95th and 99th 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 ($p \leq .001$), 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 three 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 et al., 2019), and we constrained linkage estimation to sites ≤100 kbp from one another. Using these linkage estimates, we calculated mean LD in 100 kbp windows in 5 kbp steps in R, and identified outlier regions as contiguous series of ≥10 windows exceeding 2× the interquartile range (2×IQR) for mean windowed LD.

2.3 | 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 et al., 2000), 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 50 bp 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 1500 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 analysed with PPANALYZE including calculation of F_{ST}, sliding window F_{ST} (100 kbp windows in 5 kbp steps), and FET. Significance values from the Exact tests were used in local score analyses, using three replicate runs with ξ representing the 80th, 90th, 95th and 99th quantiles of significance values (the 70th 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 ≤100 kbp from one another. As above, we calculated mean LD in 100 kbp windows in 5 kbp steps in R, but identified outlier regions as contiguous series of ≥20 windows exceeding 2× the interquartile range (2×IQR) 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.

3 | RESULTS

3.1 | 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 0.33× coverage datasets and at ≥40 reads for the 1× 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 false-positive and false-negative 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 false positives was similar across coverages for PoolParty2 (≤1%), this value changed with coverage for ANGSD (Table 1). In addition, only at the highest significance thresholds did ANGSD recover sites with similar false-positive proportions to PoolParty2, but then in lower absolute numbers. The rates of false-positive and false-negative loci for ANGSD were similar for higher MAF loci, but ANGSD still recovered lower absolute numbers of 'real' loci than PoolParty2 (Table 1).

Across coverage depths, allele frequencies estimated by PPAlign were equivalent or modestly more accurate than those estimated by ANGSD (Figure 2, Figure S2). Allele frequency estimates generally improved with depth for both PoolParty2 and ANGSD, though most

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 (0.33× and 1×). SNPs were filtered at two minor allele frequencies (MAF) and tallied, and the percentage of each that were false positives (loci not simulated) and false negatives (simulated loci ≥MAF not recovered) is reported.

		Poolparty		ANGSD			
		PPalign[a]	PPanalyze[b]	No *p* filter	*p* < .01[c]	*p* < .001[c]	Population filtering[d]
High—0.33×	MAF ≥0.001	70,067	68,292	527,762	94,104	57,027	56,060
	% false positive	0.54	0.54	82.83	30.43	1.26	1.28
	% false negative	56.1	57.2	42.9	58.8	64.5	65.1
	MAF ≥0.05	65,479	63,505	56,729	55,727	52,418	51,451
	% false positive	0.17	0.20	1.07	0.89	0.35	0.35
	% false negative	1.4	4.4	15.4	16.7	21.2	22.7
High—1×	MAF ≥0.001	91,509	86,036	920,208	209,617	127,423	127,415
	% false positive	0.22	0.22	87.95	56.51	32.49	32.49
	% false negative	42.5	45.9	30.1	42.6	45.8	45.8
	MAF ≥0.05	66,733	62,323	57,475	57,473	57,472	57,464
	% false positive	0.02	0.02	0.03	0.03	0.02	0.02
	% false negative	0.0	6.0	13.4	13.4	13.4	13.4
Low—0.33×	MAF ≥0.001	345,223	334,664	864,530	344,606	275,355	271,175
	% false positive	0.23	0.23	48.55	7.41	0.21	0.20
	% false negative	56.4	57.7	43.6	59.6	65.2	65.7
	MAF ≥0.05	3,22,926	3,11,366	2,74,274	2,69,762	2,55,028	2,50,861
	% false positive	0.09	0.09	0.20	0.17	0.11	0.11
	% false negative	1.5	5.0	16.4	17.8	22.2	23.5
Low—1×	MAF ≥0.001	452,827	422,029	1,323,894	411,213	373,603	373,473
	% false positive	0.22	0.21	58.89	1.86	0.39	0.39
	% false negative	42.8	46.6	31.0	48.9	52.9	52.9
	MAF ≥0.05	3,30,525	3,05,110	2,81,546	2,81,527	2,81,343	2,81,216
	% false positive	0.08	0.08	0.09	0.09	0.09	0.09
	% false negative	0.0	6.9	14.1	14.1	14.2	14.2

[a]Min. 20 global reads.

[b]Initial sites specified by PPalign; min. 3 ind + 10 reads each pop.

[c]Min. 20 global reads.

[d]Initial specified by ANGSD *p* < .001; min. 3 ind + 10 reads each pop.

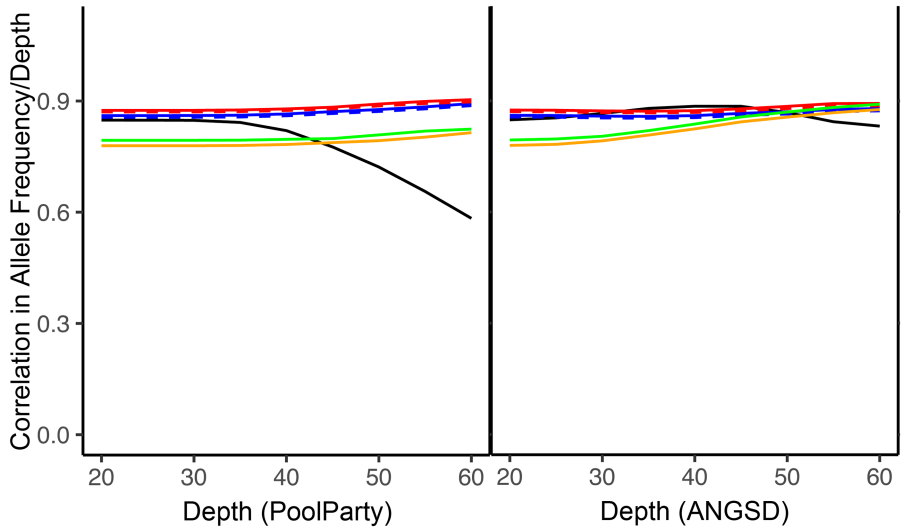

FIGURE 2 Comparing estimated allele frequencies with true allele frequencies across depth cut-offs for data simulated with high background F_{ST} and low coverage (0.33×). Left axis: The red and blue lines show correlation for PoolParty2 estimated frequencies for each population for normalized (solid) 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.

notably for ANGSD when depth was estimated by ANGSD rather than PoolParty2. Indeed, ANGSD and PoolParty2 disagreed considerably about depth (correlation in depth estimates decreased as the threshold for depth increased, but only when depth was reported by PoolParty2), implying that ANGSD's read filter is more stringent than PoolParty2, even with apparently similar parameter values. Moreover, while ANGSD and PoolParty2 both exhibited strong linear correlations for true and estimated allele frequency as well as error magnitudes that remained similar across minor allele frequencies, the marginally higher error rate for ANGSD appeared to be tied to this program's divergent perception of coverage (Figure S3). Indeed, the difference in error between PoolParty2 and ANGSD was more closely tied to the proportion of depth that PoolParty2 reported that ANGSD corroborated (Spearman's rho −0.55) than the total depth ANGSD reported (Spearman's rho −0.33) across sites. Nonetheless, correlation values for estimated and true allele frequencies for both analytical suites ranged from approximately 80%–90% for the lower coverage datasets and 90%–95% for the higher coverage datasets. We note that some diminishment in accuracy was expected due to sampling variance (the 'true' allele frequencies reflect the population frequencies before individual sampling and sequence read simulation), as these determine the truly estimable allele frequencies regardless of the fidelity of each analysis. This is corroborated by the observation that correlations between allele frequencies estimated by PoolParty2 and ANGSD were always higher with each other than with the 'true' 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 (Figures S4–S7). 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 $p \leq .05$, 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.

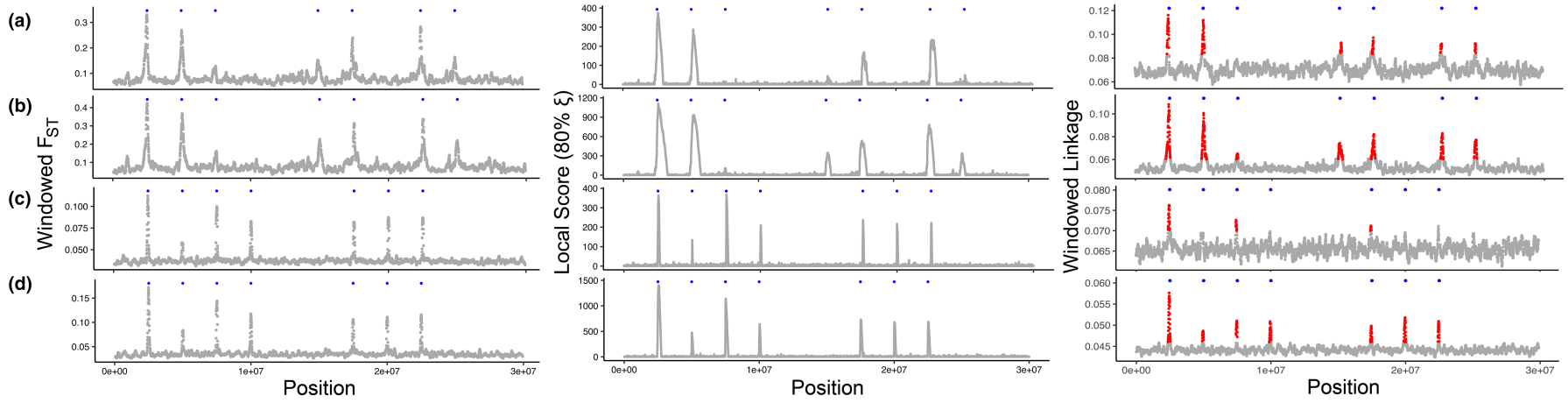

FIGURE 3 Manhattan plots of divergence estimated for variant sites identified by PoolParty2 in simulated data in scenarios with high (a, b) or low (c, d) background F_{ST} and 0.33× (a, c) or 1× coverage (b, d). For each dataset, the divergence (F_{ST}) estimated in 100 kbp sliding windows (left column),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

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.

			Linkage disequilibrium windows		Local score ξ quantiles				
					70	80	90	95	99
Background F_{ST}	Coverage	Outlier (Mbp)	Width	LD mean (max)	Width (observation across replicates)				
High	0.33×	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	1×	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	0.33×	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	1×	2.5	275	0.051 (0.058)	1025 (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)	–	–

3.2 | Application 2: Barcoded individuals in population samples of steelhead

After trimming, mapping, and quality filtering, PPALIGN provided 287–405 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 (Figure S8). The distribution of genomic extent across chromosomes was similar to other lcWGS analyses of the *O. mykiss* genome (e.g. Micheletti et al., 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

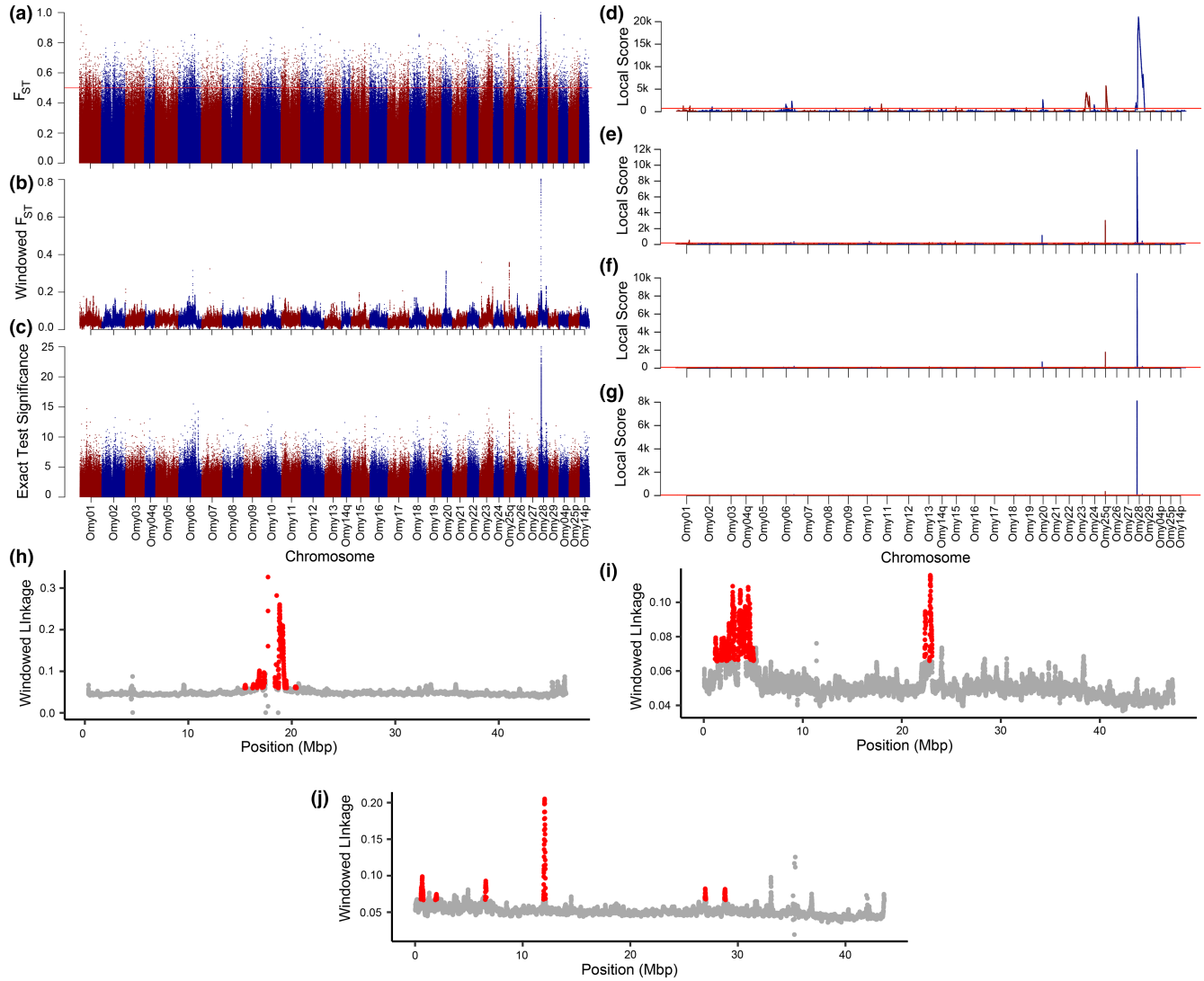

FIGURE 4 Manhattan plots of divergence and linkage among hatchery and natural origin population samples of steelhead. (a) Individual site F_{ST} corrected for sample size (red line indicates F_{ST} of 0.5). (b) Mean F_{ST} in sliding windows of 100 kb across the genome. (c) −log 10 significance (p) values from individual site Fisher's exact tests. (d–g) Local score analysis of significance values from Fisher's exact tests for the 80th, 90th, 95th and 99th quantile of significance values as ξ (red line indicates FDR of 0.05). (h–j) Mean linkage in 100 kbp windows for chromosomes 20, 25 and 28 (red dots indicate series of 20+ windows of mean linkage above 2× the interquartile range).

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 $p \leq .05$ across ξ values and replicates in Local Score analyses (Figure 4, Table 3, Table S1). 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 homologue 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.

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 (Mbp), width (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.

			Start (Mbp)	End (Mbp)	Width (kbp)	LD mean (max)
Omy 20	LS quantile (ξ)	80	16.07	22.98	6907	—
		90	17.34	19.50	2166	—
		95	18.83	19.05	211.8	—
		99	18.91	18.92	6.3	—
	LD outliers		17.70	18.70	1100	0.087 (0.327)
			18.72	19.33	701.9	0.155 (0.261)
Omy 25	LS quantile (ξ)	80	22.31	31.58	9272	—
		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	3608	—
		95	12.02	13.17	1158	—
		99	12.03	12.46	432.6	—
	LD outliers		11.94	12.12	284.8	0.137 (0.206)

4 | DISCUSSION

4.1 | 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 lcWGS data with a high-quality reference genome to perform genome-wide genetic association analysis with discrete 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 and 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.

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 portray, these analyses, with appropriate corroboration, may identify many loci with smaller effects on polygenic traits. These analyses rely on allele frequency data, which are most effectively estimated by sampling larger numbers of individuals, even at lower coverage, and as such are best suited for investigations into discrete phenotypes with high heritability and penetrance or environmental conditions with distinct contrast, such that individuals can be grouped into classes that capture the relevant variation. Loci contributing to polygenic traits with continuous variation or complex interactions, however, will be challenging to discover with analyses that rely mainly on allele frequency. Discovery of these loci will likely depend on identifying combinatorial associations among non-contiguous genotypes from individual-level rather than group-level data, for which genotype likelihoods are better suited. angsd, which requires read alignments such as those produced by PPalign as input, can estimate individual genotype likelihoods or posterior probabilities useful for such analyses. As we demonstrate here, genotype likelihoods also allow the estimation of linkage disequilibrium in specified windows to corroborate outlier regions, even from very low-coverage data, making PoolParty2 and angsd effective complements. It should be noted, however, that genotype likelihoods may only be sufficiently informative for some analyses if coverage per individual is sufficiently high, and the decision as to how to multiplex individuals for a given total number of sequencing reads must be made with respect to the types of analyses that will be needed for the known phenotypic/environmental variation and hypothesized genetic architecture (Lou et al., 2021; Paril et al., 2022).

Using simulated data, we demonstrated that angsd and PoolParty2 both provide precise estimates of allele frequency considering the constraint of individual sequence coverage and variation

with which they were challenged. 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 for calculating F_{ST} with ANGSD, each population must be analysed separately, and if the user desires to utilize the empirical major/minor alleles rather than reference/alternate, a prior run with all data must be made. Beyond these, however, the Local Score analyses (Fariello et al., 2017) and estimates of linkage outliers from NGSLD (Fox et al., 2019) provide powerful and complementary means to identify true outlier regions that can be run in simultaneously. We provide the code demonstrating this for our simulated and empirical data (File S1).

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. Further, an analysis designed to objectively discriminate outlier loci from background rates of variation, Local Score, identified fewer significant regions in this scenario. 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. Moreover, while linkage disequilibrium may reflect regions subject to divergent selection, as our empirical results portray, it may also result from non-selective processes, such as position within the chromosome, admixture or inbreeding, and effectively discriminating high LD regions associated with selection will rely on corroboration with other analyses. 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, that is, 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 artefacts commonly known as batch effects (Leek et al., 2010; Lou & Therkildsen, 2022). These occur when idiosyncrasies of the library and sequencing process introduce artefacts 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 et al., 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 artefacts will be identified as group-wise differences (O'Leary et al., 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 artefacts owing to any technical pair will be identified as consistent, significant differences (Cochran, 1954; Mantel & Haenszel, 1959).

4.2 | Discovery of large effect loci in non-model organisms

We successfully applied our integrated pipeline to a common data arrangement of lcWGS data from steelhead trout (*O. 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., 2000; Carlson & Seamons, 2008; Quinn et al., 2011). Many of these traits are highly variable both among and within phylogeographic lineages and local populations (Busby et al., 2000), are part of the portfolio of life history diversity that assist populations in persisting through natural and anthropogenic environmental challenges (Quinn et al., 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 signalling. This region on chromosome 20 is also known to harbour 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 analysing 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).

CONFLICT OF INTEREST STATEMENT

The authors assert no conflict of interest.

DATA AVAILABILITY STATEMENT

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).

ORCID

Stuart Willis https://orcid.org/0000-0002-2274-1112
Steven Micheletti https://orcid.org/0000-0002-3842-0946
Kimberly R. Andrews https://orcid.org/0000-0003-4721-1924
Shawn Narum https://orcid.org/0000-0002-7470-7485

REFERENCES

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), p1116. https://doi.org/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*, 81–92.

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, 84–88.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. 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), 241–262. https://doi.org/10.1534/genetics.110.117275

Busby, P., Wainwright, T., & Bryant, G. (2000). Status review of steelhead from Washington, Idaho, Oregon, and California. In E. E. Knudsen & D. M. McDonald (Eds.), *Sustainable Fisheries Management* (pp. 173–175). CRC Press.

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. https://doi.org/10.1111/j.1752-4571.2008.00025.x

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*(4), 417–451. https://doi.org/10.2307/3001616

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Fariello, M. I., Boitard, S., Mercier, S., Robelin, D., Faraut, T., Arnould, C., Recoquillay, J., Bouchez, O., Salin, G., Dehais, P., Gourichon, D., Leroux, S., Pitel, F., Leterrier, C., & SanCristobal, M. (2017). Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: The local score approach. *Molecular Ecology*, *26*(14), 3700–3714. https://doi.org/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), 3855–3856. https://doi.org/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), 207–218. https://doi.org/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), 205–220. https://doi.org/10.1534/genetics.113.152462

Hess, J. E., Collins, E. E., Harmon, S. A., Horn, R. L., Koch, I. J., Stephenson, J., Willis, S., & Narum, S. R. (2021). Genetic assessment of Columbia river stocks: 1/1/2020–12/31/2020 Annual Report, 2008–907-00.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., & Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *The American Naturalist*, *188*, 379–397. https://doi.org/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), 1011–1021. https://doi.org/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), 356. https://doi.org/10.1186/s12859-014-0356-4

Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundström, G., Rubin, C. J., Gilbert, E. R., Berglund, J., Wetterbom, A., Laikre, L., Webster, M. T., Grabherr, M., Ryman, N., & 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), 19345–19350. https://doi.org/10.1073/pnas.1216128109

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*, 733–739. https://doi.org/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*, *24*(5), 1031–1046. https://doi.org/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), 5966–5993. https://doi.org/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), 1678–1692. https://doi.org/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), 142–152. https://doi.org/10.1111/1755-0998.12635

Lyu, G., Feng, C., Zhu, S., Ren, S., Dang, W., Irwin, D. M., Wang, Z., & Zhang, S. (2021). Whole genome sequencing reveals signatures for artificial selection for different sizes in Japanese primitive dog breeds. *Frontiers in Genetics*, *12*, 671686. https://doi.org/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), 719–748. https://doi.org/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. *BMC Evolutionary Biology*, *18*(1), 140. https://doi.org/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*, *18*(4), 825–837. https://doi.org/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*, *27*(16), 3193–3206. https://doi.org/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), 137–152. https://doi.org/10.1111/1755-0998.13458

Pearse, D. E., Barson, N. J., Nome, T., Gao, G., Campbell, M. A., Abadía-Cardoso, A., Anderson, E. C., Rundio, D. E., Williams, T. H., Naish, K. A., Moen, T., Liu, S., Kent, M., Moser, M., Minkley, D. R., Rondeau, E. B., Brieuc, M. S. O., Sandve, S. R., Miller, M. R., … Lien, S. (2019). Sex-dependent dominance maintains migration supergene in rainbow trout. *Nature Ecology and Evolution*, *3*(12), 1731–1742. https://doi.org/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*, *73*, 1015–1030. https://doi.org/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*, *140*(1), 45–51. https://doi.org/10.1080/00028487.2010.550244

Ren, S., Lyu, G., Irwin, D. M., Liu, X., Feng, C., Luo, R., Zhang, J., Sun, Y., Shang, S., Zhang, S., & Wang, Z. (2021). Pooled sequencing analysis of geese (*Anser cygnoides*) reveals genomic variations associated with feather color. *Frontiers in Genetics*, *12*, 650013. https://doi.org/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), 749–763. https://doi.org/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), 194–208. https://doi.org/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*, *29*, 673–680. https://doi.org/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. https://doi.org/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), 435–447. https://doi.org/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), 0041901. https://doi.org/10.1371/journal.pone.0041901

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Willis, S., Micheletti, S., Andrews, K. R., & Narum, S. (2023). PoolParty2: An integrated pipeline for analysing pooled or indexed low-coverage whole-genome sequencing data to discover the genetic basis of diversity. *Molecular Ecology Resources*, *00*, 1–15. https://doi.org/10.1111/1755-0998.13888