Received: 13 October 2020 | Revised: 18 March 2021 | Accepted: 27 April 2021

DOI: 10.1111/1755-0998.13474

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Haploid, diploid, and pooled exome capture recapitulate features of biology and paralogy in two non-model tree species

Brandon M. Lind[1] | Mengmeng Lu[2] | Dragana Obreht Vidakovic[1] | Pooja Singh[2] | Tom R. Booker[1,3] | Sally N. Aitken[1] | Sam Yeaman[2]

[1]Department of Forest and Conservation Sciences, Centre for Forest Conservation Genetics, University of British Columbia, Vancouver, BC, Canada

[2]Department of Biological Sciences, University of Calgary, Calgary, AB, Canada

[3]Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada

Correspondence
Brandon M. Lind, University of British Columbia, 2424 Main Mall, 3027 Forest Science Centre, Vancouver, BC, Canada.
Email: lind.brandon.m@gmail.com

Funding information
Genome Canada 241REF; Genome British Columbia; Genome Alberta; Génome Québec; British Columbia Ministry of Forests, Lands and Natural Resources; Alberta Innovates Bio Solutions; Vernon Seed Orchard Company; Forest Genetics Council of British Columbia; University of Alberta; University of British Columbia, Faculty of Forestry Universite Laval; Compute Canada; Mosaic Forest Management; Western Forest Products; University of Toronto; Swiss Federal Research Institute; University of Calgary; Unites States Forestry Service; TimberWest; Canadian Forest Service

Abstract

Despite their suitability for studying evolution, many conifer species have large and repetitive giga-genomes (16–31 Gbp) that create hurdles to producing high coverage SNP data sets that capture diversity from across the entirety of the genome. Due in part to multiple ancient whole genome duplication events, gene family expansion and subsequent evolution within *Pinaceae*, false diversity from the misalignment of paralog copies creates further challenges in accurately and reproducibly inferring evolutionary history from sequence data. Here, we leverage the cost-saving benefits of pool-seq and exome-capture to discover SNPs in two conifer species, Douglas-fir (*Pseudotsuga menziesii* var. *menziesii* (Mirb.) Franco, *Pinaceae*) and jack pine (*Pinus banksiana* Lamb., *Pinaceae*). We show, using minimal baseline filtering, that allele frequencies estimated from pooled individuals show a strong, positive correlation with those estimated by sequencing the same population as individuals ($r > .948$), on par with such comparisons made in model organisms. Further, we highlight the utility of haploid megagametophyte tissue for identifying sites that are probably due to misaligned paralogs. Together with additional minor filtering, we show that it is possible to remove many of the loci with large frequency estimate discrepancies between individual and pooled sequencing approaches, improving the correlation further ($r > .973$). Our work addresses bioinformatic challenges in non-model organisms with large and complex genomes, highlights the use of megagametophyte tissue for the identification of paralogous artefacts, and suggests the combination of pool-seq and exome capture to be robust for further evolutionary hypothesis testing in these systems.

KEYWORDS
exome-capture, non-model, paralogy, *Pinaceae*, pool-seq

Lind and Lu contributed equally.

1 | INTRODUCTION

Quantifying the spatial structure of neutral and adaptive genetic variation within ecologically and economically important tree species and their close relatives is fundamental to forecasting and managing their response to changing selection pressures from pests, pathogens, and climate (Aitken et al., 2008; Alberto et al., 2013; Holliday et al., 2017; Janes & Hamilton, 2017; Sniezko & Winn, 2017). Prerequisite to this information is the ability to produce high quality and cost-effective data from which to generate reliable inference. While the life history of many tree species offers some ideal circumstances for studying adaptive evolution at the genetic level (Neale & Kremer, 2011; Neale & Savolainen, 2004), two ancient whole-genome duplication events in the progenitors of the *Pinaceae* lineages (Li et al., 2015), transposable element dynamics (Morse et al., 2009; Voronova et al., 2017), tandemly arrayed genes (Pavy et al., 2017), subsequent gene duplication (Casola & Koralewski, 2018; Krutovsky et al., 2004) and gene family expansion (e.g., Liu et al., 2016) have led to giga-genomes (>16 Gb in size) recalcitrant to chromosome-level genome assembly under current sequencing and computational constraints (Neale, Martínez-García, et al., 2017; but see Scott et al., 2020). For example, analysis of *Pinus taeda* L. (*Pinaceae*) has yielded estimates that upwards of 82% of its 22 Gb genome is repetitive, and 75% of the repetitive sequence is due to retrotransposons (Nystedt et al., 2013; Wegrzyn et al., 2014). It is also thought to be rich in pseudogenes (Kovach et al., 2010).

Such large genome sizes have hampered production of dense SNP data sets across a large number of individuals (Lind et al., 2018). Most recent sequencing efforts in conifers have either used some form of reduced representation sequencing such as restriction-site associated DNA sequencing (i.e., RADseq; reviewed in Andrews et al., 2016 and Parchman et al., 2018), which relies upon relatively few genomic resources, or targeted capture (e.g., Lu et al., 2016; Suren et al., 2016), which requires significant genomic and budgetary resources including the design of capture arrays (but see Puritz & Lotterhos, 2018). To capture population-level polymorphism information while minimizing cost, sequencing pooled individuals (i.e., pool-seq approaches) has emerged as a cost-effective alternative to sequencing individuals (Gautier et al., 2013; Schlötterer et al., 2014). Further, pool-seq can be combined with targeted capture approaches to both reduce cost and sample specific areas of the genome that are *a priori* considered functionally relevant (e.g., Rellstab et al., 2019).

The pooling of biological samples has been commonplace for decades (Dorfman, 1943), owing to the cost-efficiency of analysing multiple samples together. Such methods have expanded to other purposes, such as the estimation of allele frequencies of nucleotide polymorphisms in next-generation sequence data (i.e., pool-seq). Pool-seq approaches use read counts across pooled individuals to estimate allele frequencies, generally for a single population, with individuals pooled with equimolar contributions. A number of studies have empirically evaluated the congruence between individual and pool-seq allele frequency estimates across various taxa (e.g., Fracassetti et al., 2015; Futschik & Schlötterer, 2010; Rellstab et al., 2013, 2019). Such studies have led to broad agreement on the accuracy of pool-seq when following best practices for the organism and study design. Of exceptional significance for the estimation of allele frequency from read count data is the proper alignment of reads to the reference. Misalignments, which may be particularly important for exome capture data from members of *Pinaceae*, can be due to reads from paralog gene copies in the data mapping to the incorrect copy in the reference, or from paralog copies being collapsed into a single sequence in the reference assembly where copies in the data map to this single sequence. Such misalignments can be exacerbated by assembly errors in the reference, particularly for organisms with repetitive genomes. These misalignments will skew allele ratios and bias allele frequency estimates downstream. In particular for non-model species with histories of whole genome duplication or gene family expansion, steps must be taken to categorize misalignments so that there are not substantial allele frequency biases in downstream data sets. Indeed, methods by which to detect such loci have received considerable attention (see Table 1 in McKinney et al., 2017). Among these, one such method uses haploid samples and the presence of heterozygote genotype calls to identify potential paralogous artefacts (Limborg et al., 2016), since haploid samples can only be monoallelic. Another uses read ratio depths among heterozygote individuals from individual sequence data to identify deviations expected from duplicated loci (McKinney et al., 2016). While multicellular haploid tissue is not present in vertebrates, such tissue is readily accessible from gametophytic life stages in many plant species, and in particular from the

TABLE 1 Description of datasets used to call SNPs for both Douglas-fir and jack pine. indSeq[a] and poolSeq data sets for a given species share the same individuals from a single population. The megaSeq data set consists of haploid megagametophyte tissue from a single individual not included in the indSeq or poolSeq data sets

Data set	Ploidy per sample (number of samples)	SNP Caller	Purpose
indSeq	2 (20)	`GATK4`	Validate poolSeq allele frequency estimates; calculate read ratio statistics to validate candidate paralog misalignments
poolSeq	2 (20)	`VarScan`	Compare with indSeq SNP set to determine filtering strategy
megaSeq	1 (1)	`VarScan`	Identify heterozygous sites as candidates for false SNPs due to misalignment of diverged/duplicated paralogs

Note: [a]Note that we use camelCase to denote our data sets, and reserve hyphens (e.g., pool-seq) to denote methodologies.

maternally-derived megagametophyte tissue that can be excised from the seeds of conifer species.

Here we harness the multicellular haploid megagametophyte of conifers to aid in mapping and analysing pool-seq data from diploid individuals. We use this pooled exome capture approach for two conifers: coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii* (Mirb.) Franco, *Pinaceae*) and jack pine (*Pinus banksiana* Lamb., *Pinaceae*), to evaluate the utility of pool-seq approaches in these systems. We use sequence data from haploid samples to identify misalignments from paralogous sites, and use individual sequence data to validate both the allele frequency estimates of the same individuals in pools and the candidate regions affected by paralog misalignments detected with haploid data (Table 1). We then use this information to quantify their effects on the congruence between individual and pool-seq allele frequency estimates. Together, these data sets provide a path forward for filtering pool-seq data of this kind, particularly for studies of non-model organisms using a diverged, and potentially fragmented, reference genome. Our methods further highlight a cost-effective means to empirically isolate potentially misaligned paralogs in species with accessible haploid tissue, which to date has not been widely used for such purposes in conifers.

2 | MATERIALS AND METHODS

2.1 | Focal species and population sampling

Coastal Douglas-fir (*Pseudotsuga menziesii* var. *menziesii*) is a temperate species occupying primarily coastal habitat along the west coast of North America from California to British Columbia as well as inland habitat in the Cascade and Klamath ranges of Washington, Oregon, and California. It is important to the ecology and economical value of many of these forests. Jack pine (*Pinus banksiana*) has a vast distribution across the Canadian boreal forest, stretching from Atlantic Canada into western Alberta and Northwest Territories, and is important to the ecology of many of these systems and to the forest industry in some regions.

For both Douglas-fir and jack pine, we sampled 20 individuals for use in individual and pooled sequencing sets from operational reforestation seedlots created from open-pollinated seeds from tens or hundreds of seed parents from a single provenance (see Appendix S1: Section 1.1). We used a single jack pine seed to extract megagametophyte haploid tissue. For Douglas-fir haploid data, we downloaded paired-end fastq files from a previously sequenced Douglas-fir megagametophyte taken from a single individual (NCBI SRA accession SAMN0333061, Neale, McGuire, et al., 2017) to match our sequencing effort for jack pine haploid tissue (Appendix S1: Section 1.2).

2.2 | Exome capture probe design

The capture probes were designed based on the genes identified using RNA sequencing (RNA-seq) data for Douglas-fir and jack pine. De novo transcriptome assembly was performed for each species using RNA-seq reads. For jack pine, RNA-seq reads were sequenced from a frozen sample of young needles taken from a recently flushed bud of a single tree grown in a growth chamber with a simulated climate corresponding to a mean annual temperature of 6°C (Appendix S1: Section 1.3). For Douglas-fir, RNA-seq reads were obtained from two sources: one source was the read sets deposited in NCBI SRA, including SRX1851630 (Little et al., 2016), SRX1286745 (Hess et al., 2016), SRX1341335 (Cronn et al., 2017a), and SRX136240 (Cronn et al., 2017b). The other source was the reads sequenced from two needle samples infected by the fungal pathogen *Phaeocryptopus gaeumannii*, which causes Swiss needle cast disease in Douglas-fir (Appendix S1: Section 1.3).

The raw reads were processed by the software FASTX TOOLKIT (v0.0.13, http://hannonlab.cshl.edu/fastx_toolkit), including clipping the adaptors (`-l 25`), filtering the artefacts, and keeping the reads with a minimum quality score of 20. The filtered reads were used to perform de novo transcriptome assembly using the software TRINITY v2.4.0 (`--bowtie2`, Grabherr et al., 2011). Among the assembled transcripts, only the longest isoforms with a length of at least 300 bp for each gene were retained, which were then used as reference to align the reads using the software RSEM (v1.3.0 Li & Dewey, 2011). From the expression quantification of transcripts, transcripts with aligned reads and transcript per million (TPM) ≥1 were retained. The completeness of the filtered transcripts was examined using the 1375 orthologues in the Benchmarking Universal Single Copy Orthologues (BUSCO: v3.0), set of embryophyta_odb10 (`--evalue 1e-10`, Simão et al., 2015).

To avoid probes spanning exon-intron boundaries, exons were targeted to design probes. Using the software GMAP (v2017-06-20, Wu & Watanabe, 2005), the filtered transcripts from Douglas-fir were aligned to the convarietal reference (*P. menziesii var. menziesii* (coastal Douglas-fir; v1.0, Neale, McGuire, et al., 2017). The jack pine transcripts were aligned to the congeneric loblolly pine (*Pinus taeda*) reference genome (v.1.01, Wegrzyn et al., 2014) as there is no available jack pine reference genome, and both loblolly and jack pine belong to *Pinus* subgenus *Pinus*, the hard pines. Exon sequences with a length of at least 100 bp were submitted to Roche NimbleGen for Custom SeqCap EZ probe design.

To evaluate the capture efficiency of the probes, the captured sequences were aligned to reference genomes and numbers of reads on-target, near-target (≤500 bp from target regions), and off-target regions were counted using "intersect" function in the software BEDTOOLS v2.28.0 (`-f 0.75`, Quinlan & Hall, 2010). The depth of captured sequences was counted using "depth" function in the software SAMTOOLS v1.3 (`-q 30 -Q 20`, Li et al., 2009). The cumulative depth was calculated and plotted using R (R Core Team, 2018).

2.3 | DNA extraction, library preparation, and sequencing

In total, three data sets were created for each of the two species (Table 1)—note that we use camelCase (e.g., poolSeq) to denote our

data sets, and reserve hyphens (e.g., pool-seq) to denote methodologies. These data sets included individual sequencing of 20 diploid individuals from a single population (hereafter indSeq), the same individuals pooled together with equimolar contributions prior to sequencing (hereafter poolSeq), and haploid megagametophyte tissue sequenced from a single individual (hereafter megaSeq). We use the indSeq data set to validate allele frequency estimates from our poolSeq data, and the megaSeq data to probe our data for apparent heterozygote SNPs (i.e., potential false-positive SNPs) caused by the misalignment of diverged paralogs that could affect our allele frequency estimates (Table 1; see also Section 2.6).

For each data set we extracted DNA from either diploid needle tissue or haploid megagametophyte tissue (see Appendix S1: Section 1.3). From these extractions, approximately 100 ng of DNA from each individual or pooled DNA sample was used for a barcoded (Kapa, Dual-Indexed Adapter Kit) library with an approximately 350-bp mean insert size. SeqCap library preparation was performed using custom NimbleGen SeqCap probes (described above in 2.1) according to the NimbleGen SeqCap EZ HyperCap Workflow User's Guide Ver 2 (Roche Sequencing Solutions, Inc.). Following capture, each library was sequenced in a 150 bp paired-end format on an Illumina HiSeq4000 instrument at the Centre d'expertise et de services Génome Québec, Montreal, Canada.

2.4 | Bioinformatic SNP calling pipelines

Raw paired-end sequence reads from all data sets were trimmed with FASTP (v0.19.5, Chen et al., 2018) by trimming reads that did not pass quality filters of <20 Ns, a minimum mean Phred quality score of 30 for sliding windows of five base pairs (bp), and a final length of 75 bp with no more than 20 bp called as N (`-n 20 -M 30 -W 5 -l 75 -g -3`). Trimmed reads were mapped with BWA MEM (v0.7.17, Li & Durbin, 2009) to reference assemblies; we mapped jack pine to the loblolly reference (v2.01, Wegrzyn et al., 2014) and Douglas-fir to the conarietal reference (v1.0, Neale, McGuire, at al., 2017). The resulting `.sam` files were converted to binary with SAMTOOLS v1.9 (`view, sort, index`; Li et al., 2009) and subsequently filtered for proper pairs and a mapping quality score of 20 or greater (`view -q 20 -f 0x0002 -F 0x0004`). Using PICARDTOOLS v2.18.9 (http://picard.sourceforge.net), read groups were added and duplicates subsequently removed from filtered bam files.

We then called SNPs using the Genome Analysis Toolkit (GATK v4.1.0.0; McKenna et al., 2010) for indSeq data, and `VarScan` (v2.4.3; Koboldt et al., 2012) for both poolSeq and megaSeq data sets (Table 1) for comparisons since data sets that stem from a larger project are all poolSeq (and we will therefore only be using `VarScan`). For SNPs called with `GATK4`, we used HaplotypeCaller (`--genotyping-mode DISCOVERY -ERC GVCF`) and GenotypeGVCFs. We then filtered data with SelectVariants (`--select-type-to-include SNP`), VariantFiltration (`--filter-expression "QD <2.0 || FS >60.0 || MQ <40 || MQRankSum < -12.5"`), and finally with `vcftools` v0.1.14 (`--maf 0.00 -minGQ 20 -max-missing 0.75`; Danecek et al., 2011). BQSR was not carried out in our analysis due to the lack of a high-quality reference set of SNPs for our species. Note that no further filtering (e.g., for depth) was done for this initial baseline filtering strategy (further filtering is described in 3.4).

Before calling SNPs with `VarScan`, we first realigned indels with GATK 3.8 (McKenna et al., 2010)—`RealignerTargetCreator` then `IndelRealigner`—and then passed a SAMTOOLS mpileup object directly to `VarScan::mpileup2cns` with a minimum coverage set to 8, *p*-value < .05, minimum variant frequency of 0.00, ignoring variants with >90% support on one strand, a minimum average genotype quality of 20, and a minimum allele frequency of 0.80 to call a site homozygous (`--min-coverage 8 --p-value .05 --min-var-freq 0.00 --strand-filter 1 --min-avg-qual 20 --min-freq-for-hom 0.80`). Output was then filtered with a custom `python` (v3.7, www.python.org) script to filter out indels, keep only biallelic loci, and to ensure a genotype quality score >20. From the megaSeq data, we then isolated heterozygous SNP calls (hereafter megaSNPs) that represent errors in genotype calling given the haploid nature of the tissue sequenced—to keep only heterozygous calls, we ignored any biallelic cases where only the non-reference allele was called. Such apparent SNPs are probably false due to misalignments. We have published our complete SNP calling pipelines in publicly available repositories (Lind, 2021a; Lind, 2021b).

2.5 | Validation of megaSNPs as indicators of paralogy artefacts

To check whether heterozygous sites (megaSNPs) called from `VarScan` megaSeq are following expectations of patterns from paralogs, we investigated read ratio deviations from a binomial expectation for these `VarScan` megaSNP sites at the same sites in our `GATK` indSeq data using heterozygous diploid individuals (sensu McKinney et al., 2017; see also Rellstab et al., 2019). For true positive SNPs, heterozygous diploid individuals should have, on average, an even ratio of reference (REF) and alternative (ALT) read counts. If the SNP is due to a bioinformatic error arising from the misalignment of paralogs (i.e., a false positive SNP), the read ratio will differ significantly from this expectation when there is a SNP at a given position in only one paralog copy (McKinney et al., 2017). Similarly, if there is a fixed difference at a given position between two copies, then all individuals in a population will present as heterozygotes with balanced read counts for REF and ALT at that site. If we are sequencing (and then post-hoc correctly identifying) paralogs in our poolSeq data using megaSNP sites, misalignment of either duplicated or diverged paralogs will cause read ratio deviations in these loci (and affect allele frequency estimates from poolSeq, and downstream analyses), which we should be able to detect in our indSeq data. As described by McKinney et al. (2017), subsequent to whole-genome duplication during the rediploidization phase as homeologous chromosomes diverge, tetrasomically inherited sets of paralogs (duplicates) organize into distinct disomic loci (diverged duplicates).

We calculated these read ratio statistics for sites within the intersection of (1) megaSNPs, indSeq, and poolSeq SNPs, and (2) poolSeq and indSeq SNPs alone; hereafter intersections I1 and I2. The purpose of (1) is to see how paralogs could affect our poolSeq data (leveraging information in our indSeq data to do so), and of (2) is to visualize the potential influence of paralogs in our data independent of sites identified as megaSNP sites, as well as to compare poolSeq allele frequency estimates with those estimated from the indSeq data set. For these sites, we queried the indSeq data to record the frequency of heterozygous individuals (H), the allele depth ratio $\left(D = \frac{\text{REF depth}}{\text{total depth of coverage}}\right)$, and the deviation of allele depth from expectation $(\text{REF depth} - 0.5^{*}\text{total depth})$ standardized by properties of a binomial distribution with n = depth of coverage, and $p = .5$ (i.e., the z-score for the allele ratio deviation) following McKinney et al. (2016) and McKinney et al. (2017) with modifications to correctly account for missing data when calculating the proportion of heterozygotes at a particular locus. We compare our results with simulations carried out by McKinney et al. (2017). We used custom `python` code to replicate the methods of McKinney et al. (2016) with modifications, which is available on GitHub (003_testdata_validate_megaSNPs.ipynb, Lind, 2021c).

2.6 | Comparison of sequencing approaches

To study the utility of our pooled exome capture approach, we compared estimates of allele frequency from our indSeq data with estimates from our poolSeq data. To do so, we took the baseline-filtered SNPs from poolSeq and indSeq (see Section 2.4) and identified common SNPs (i.e., intersection I2). To quantify and visualize congruence between allele frequencies estimated with these methods, we report Pearson's correlation coefficient, plot histograms to visualize the congruence across the minor allele frequency (MAF) spectrum, and further plot 2D histograms to visualize congruence of allele frequency estimates. To visualize how filtering poolSeq SNPs affects the congruence between indSeq and poolSeq allele frequency estimates, we plot the allele frequency differences between methods (hereafter *AFdiff*, calculated as the difference in allele frequency methods of poolSeq and indSeq: $poolSeq_{AF} - indSeq_{AF}$) against poolSeq MAF, poolSeq depth of coverage, H, and the z-score of read ratio deviation (where H and z were calculated using indSeq data). The code for this section can be found on GitHub (002_testdata_compare_AFs.ipynb, Lind, 2021c).

3 | RESULTS

3.1 | Sequencing, mapping, and probe efficiency

Sequencing of the prepared libraries resulted in high quality data sets, with the average base quality above 30 before trimming having a mean of 86.99% across data sets and species, and a mean of 89.43% after trimming (Table S1). The number of sequenced reads varied across data sets but was similar within data sets (on average 405 million reads for indSeq, 130 million reads for poolSeq, 202 million reads for megaSeq). Mapping rates generally reflected the phylogenetic relationship between the sequenced individuals and the reference used, where rates were high for all coastal Douglas-fir data sets mapping to the convarietal reference (mean 85.11%) with lower rates for jack pine data sets mapping to the congeneric *Pinus taeda* reference (mean 35.36%; Table S1).

After filtering, the jack pine transcriptome has a size of 53 Mbp and contains 31,282 transcripts ranging from 300 to 16,688 bp with a mean length of 1695 bp; the Douglas-fir transcriptome has a size of 51 Mbp and contains 39,616 transcripts ranging from 300 to 15,302 bp with a mean length of 1310 bp. The BUSCO analysis to assess completeness of transcripts used in exome-capture probe design resulted in recovery of 87% of the 1375 BUSCOs in Douglas-fir transcripts, including 70% complete and single-copy BUSCOs, 2% complete and duplicated BUSCOs, and 15% fragmented BUSCOs. For jack pine transcripts, 93% of the BUSCOs were recovered, including 85% complete and single-copy BUSCOs, 2% complete and duplicated BUSCOs, and 6% fragmented BUSCOs. We aligned the transcripts to reference genomes to select exons and design probes. The final capture probe size are 41 Mbp for jack pine (design name: 180215_jackpine_v1_EZ_HX1) and 39 Mbp for Douglas-fir (design name: 80215_DOUGFIR_V1_EZ), corresponding to 32,208 genes in jack pine and 37,787 genes in Douglas-fir.

We counted the number of captured reads on-target, near-target (≤500 bp from target), and off-target regions for indSeq and poolSeq samples. As the DNA of each sample was sheared to approximately 350 bp, the 500 bp up- or downstream of target regions (near-target) can be directly captured by probes, whereas reads arise from outside the 500 bp margin are most often the unintended regions of the genome (off-target). The poolSeq samples had more reads than the indSeq samples and off-target regions had the most aligned reads (Figure 1). Reliable SNP calling is dependent on sequencing depth, so we calculated the cumulative numbers of bases on different regions. For Douglas-fir poolSeq, we obtained over 40 million bases in on-target regions with at least 20× sequencing depth (Figure 2a). For jack pine poolSeq, we obtained over 30 million bases in on-target regions with at least 20× sequencing depth (Figure 2b). Sequencing depths in near- and off-target regions were dramatically diminished compared to the on-target regions.

3.2 | SNP calling

The total number of SNPs after baseline filtering varied across data sets and species (Table 2). Douglas-fir generally had a higher number of SNPs called than jack pine, except for poolSeq data. However, there were more jack pine megaSNPs intersecting with poolSeq (25,500 SNPs) and indSeq (7408 SNPs) than for Douglas-fir data sets (825 SNPs and 293 SNPs, respectively). Given that megaSNPs are cases where a heterozygote call was made from a haploid sample and are therefore indicators of bioinformatic-paralogy errors,

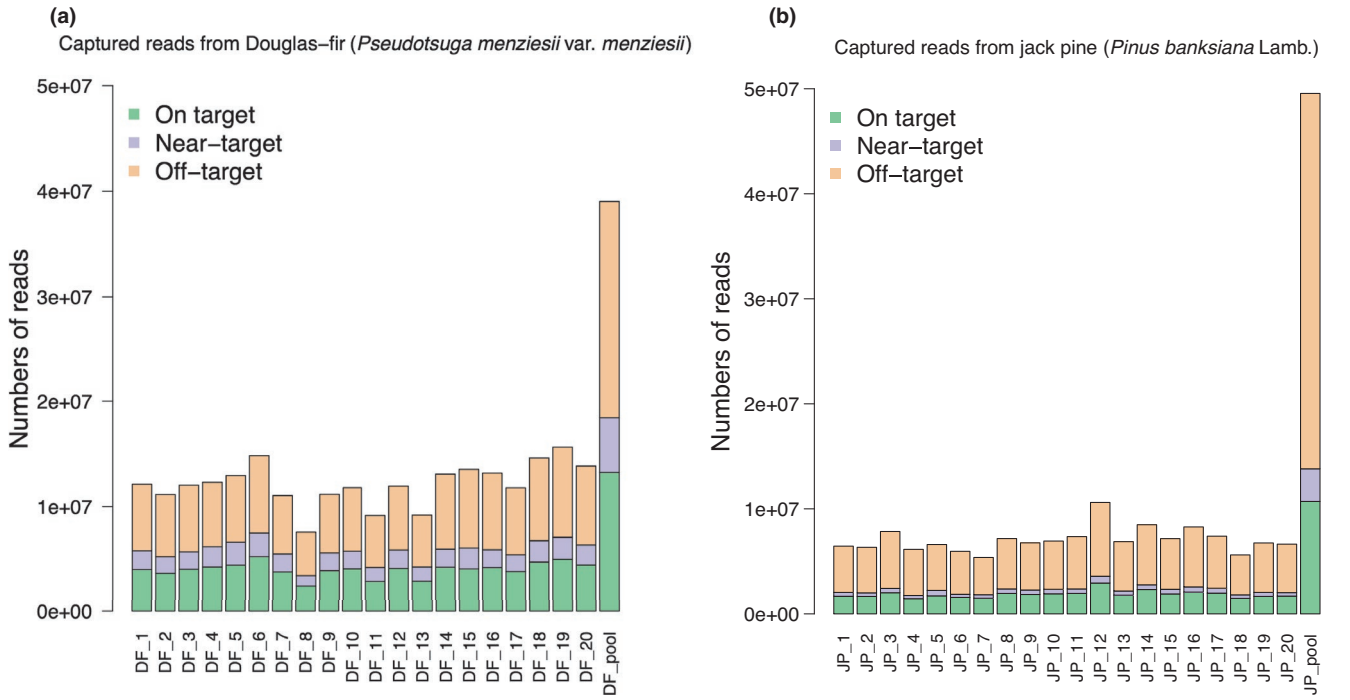

FIGURE 1 Numbers of captured reads from Douglas-fir (a) and jack pine (b) that mapped on target, near-target (≤500 bp from target) and off-target regions. On the *x*-axis, from left to right, the first 20 bars represent indSeq samples, and the last bars represents the poolSeq samples

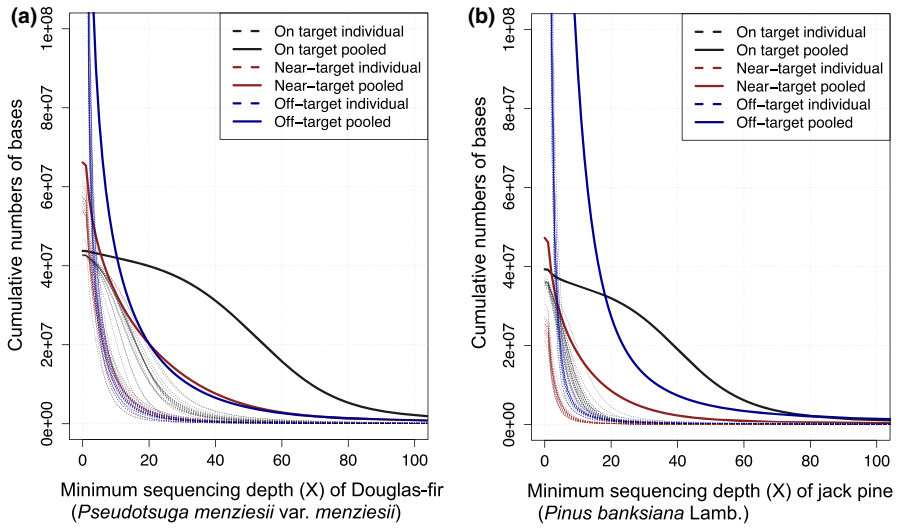

FIGURE 2 Cumulative numbers of bases in Douglas-fir (a) and jack pine (b) on target, near-target (≤500 bp from target), and off-target regions. Dashed lines represent the cumulative numbers of bases in each of 20 indSeq samples. Solid lines represent the cumulative numbers of bases in the poolSeq sample

this suggests that this error rate is much higher in jack pine. In total, several hundred thousand SNPs were found in the intersection of poolSeq and indSeq for Douglas-fir (636,279 SNPs) and jack pine (255,706 SNPs; Table 2).

3.3 | Validation of megaSNPs as indicators of paralogy artefacts

Upon inspection of our intersecting sets, patterns expected for duplicated but not diverged duplicate paralogs (McKinney et al., 2017) were apparent in both intersection I1 (megaSNPs, indSeq and poolSeq SNPs) and intersection I2 (indSeq and poolSeq SNPs), and megaSNPs did not generally typify patterns expected from nonduplicated (singleton) genes. For instance, SNPs in duplicated genes should be most distinct from SNPs in singletons when the derived allele is at intermediate frequency, and diverged duplicates are most distinct from singletons when the derived allele is fixed (Figure 3a). Sites consistent with expectations for singletons and duplicates (but not diverged duplicates) were apparent from intersection of poolSeq and indSeq sites (i.e., intersection I2; Figure 3d,e), while the indSeq sites intersecting with candidate paralog sites (megaSNPs,

TABLE 2 Output of SNPs from the conifer data sets

Data set	Species	Baseline-filtered SNPs	Baseline Intersecting SNPs	
			poolSeq	megaSeq[a]
indSeq	DF	1,526,554	636,279	293
	JP	377,080	255,706	7408
poolSeq	DF	1,601,285	—	—
	JP	3,686,528		
megaSeq[a]	DF	398,774	825	—
	JP	32,751	25,500	

Note: The intersection across all three baseline-filtered data sets were 7006 SNPs for jack pine (JP) and 248 SNPs for Douglas-fir (DF).

[a]These numbers reflect only heterozygous SNPs (i.e., megaSNPs).

i.e., intersection I1) displayed elevated levels of heterozygosity as expected from paralogs (Figure 3b,c). Indeed, patterns of deviated allele ratios were also seen in our data (Figures S1d,e and S2d,e), where the vast majority of megaSNP sites were considerably different than the 1:1 read ratio expected of heterozygous diploids (Figures S1b,c and S2b,c) as would otherwise be expected for singletons (Figures S1a and S2a). Lastly, when considering the standardized allele ratio deviation (*z*-score) we recover the same patterns of point clouds classified by McKinney et al. (2017). We observe a dense set of SNPs around the *z*-score of 0.0 for *H* values of 0.0–0.6 (Figure 4d,e) expected from singleton sites (blue in Figure 4a), another set of SNPs with elevated *H* and/or absolute *z*-score (Figure 4b,c) that is expected from duplicate loci (red in Figure 4a), and a third set of SNPs with $H > 0.9$ (Figure 4b,c) that is expected for diverged duplicates (green in Figure 4a; compare to Figures 5 and 8 in McKinney et al., 2017).

3.4 | Comparisons of sequencing approaches

Loci within the intersection of baseline-filtered indSeq and poolSeq data sets (i.e., intersection I2) showed a strong positive association between allele frequencies estimated from indSeq and poolSeq (Pearson's $r = .9760$, $p = .0000$ for jack pine; Pearson's $r = .9483$, $p = .0000$ for Douglas-fir; Figure 5a,b). Comparison of the MAF spectrum from these estimates also revealed good agreement by frequency bins (Figure S3). After exploring various filtering strategies (see Appendix S1: Section 1.4, Figures S4–S7), we applied filters that (1) showed a positive effect on congruence between allele frequency estimates in our data (removing megaSNP sites and indSeq sites with $H > 0.6$), (2) that resulted in removing sites with extreme values of *AFdiff* (removing indSeq sites with *z*-score > 10; filtering $H > 0.6$ alone also had this effect), and (3) that gave us the best estimate of indSeq allele frequency—our standard of comparison—and thus the best impression of the performance of our poolSeq approach (removing indSeq sites with >20% missing data). The correlation of allele frequencies estimated from indSeq and poolSeq data increased after this filtering (Pearson's $r = .9876$, $p = .0000$ for jack pine; Pearson's $r = .9703$, $p = .0000$ for Douglas-fir) with relatively fewer sites with extreme differences in the estimates from each method (see top-left and bottom-right corners of 2D histograms, Figure 5a–d). While some differences remain in the estimates of the minor allele frequency spectrum (Figure 5e,f), these two methods largely agree, suggesting a robust poolSeq data set for further biological hypothesis testing.

4 | DISCUSSION

The pooling of individuals to obtain next-generation sequence data is often motivated by cost savings at the expense of losing (phased) haplotype information, direct estimates of linkage disequilibrium, and rare alleles. The validation of pool-seq approaches, however, commonly involves model organisms with complete or near-complete chromosome-scale reference genomes (e.g., see Table 1 in Rellstab et al., 2013). Indeed, there are few studies that explore this congruence in non-model organisms such as conifers with large and highly fragmented reference genomes, and histories of whole genome duplications, repetitive elements, and gene family evolution (which could exacerbate misalignments through assembly errors in the reference). Here we show that combining exome capture and pool-seq can be an efficient method for quantifying genetic polymorphisms in two such species, and that heterozygous SNPs from haploid data (megaSNPs) consistently uncover sites with patterns expected from the misalignment of paralogs (Figures 3 and 4, S1 and S2). Further, we appear to uncover more false-positive variation in jack pine than in Douglas-fir (Table 2), likely due to the relative divergence between the species sequenced and reference genome used. Yet, concordance of allele frequency estimates from baseline-filtered indSeq and poolSeq data sets (i.e., intersection I2) was strong in both species ($r > .948$). Despite this high correlation, there were many loci that had extreme differences in the estimated minor allele frequency. The correlation improved further after these sites were removed with increased filtering, including the filtering of potential false-positive sites ($r > .970$, Figure 5), highlighting the utility of this method across taxa with differing demographic histories and genomic resources. These values are well within the range expected from previous pool-seq studies (Table 1 in Rellstab

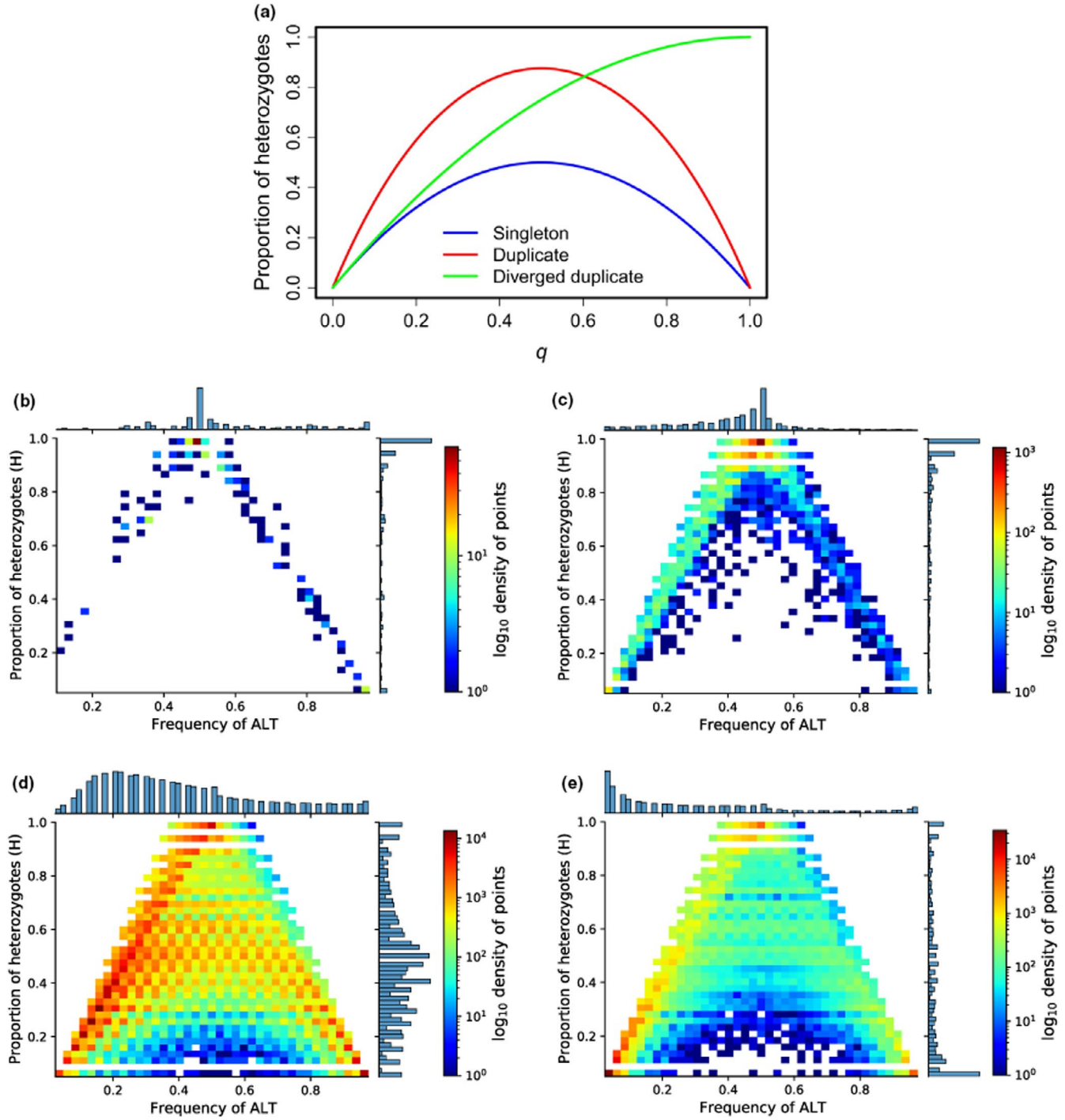

FIGURE 3 The proportion of heterozygotes, *H*, and the alternative (ALT) allele frequency calculated from indSeq data distinguish paralog misalignments according to expectations (a, Figure 1 from McKinney et al., 2017—*q* is the frequency of the ALT allele), and empirically for Douglas-fir (b, d) and jack pine (c, e). (b, c) Empirical distribution of megaSNP sites (candidate paralog sites identified as heterozygote calls from haploid tissue) calculated using indSeq data for those sites that were also called in poolSeq data (i.e., intersection I1). (d, e) Empirical distribution of intersection I2 (indSeq and poolSeq intersection) calculated using indSeq data. Note colour scale changes for each figure to accentuate patterns in the data. Frequency of ALT was binned for visualization purposes. Code to create these figures is available on GitHub (003_testdata_validate_megaSNPs.ipynb, Lind 2021c)

et al., 2013), and in some cases perform better than these model organisms.

Despite their role in adaptation and speciation (Allendorf et al., 2015; Lynch & Conery, 2000), the exclusion of potentially paralogous sites from next generation sequencing data sets is commonplace due to the difficulty in differentiating genetic polymorphisms from differences present among copies from single or diverged gene families (Dou et al., 2012; Dufresne et al., 2014;

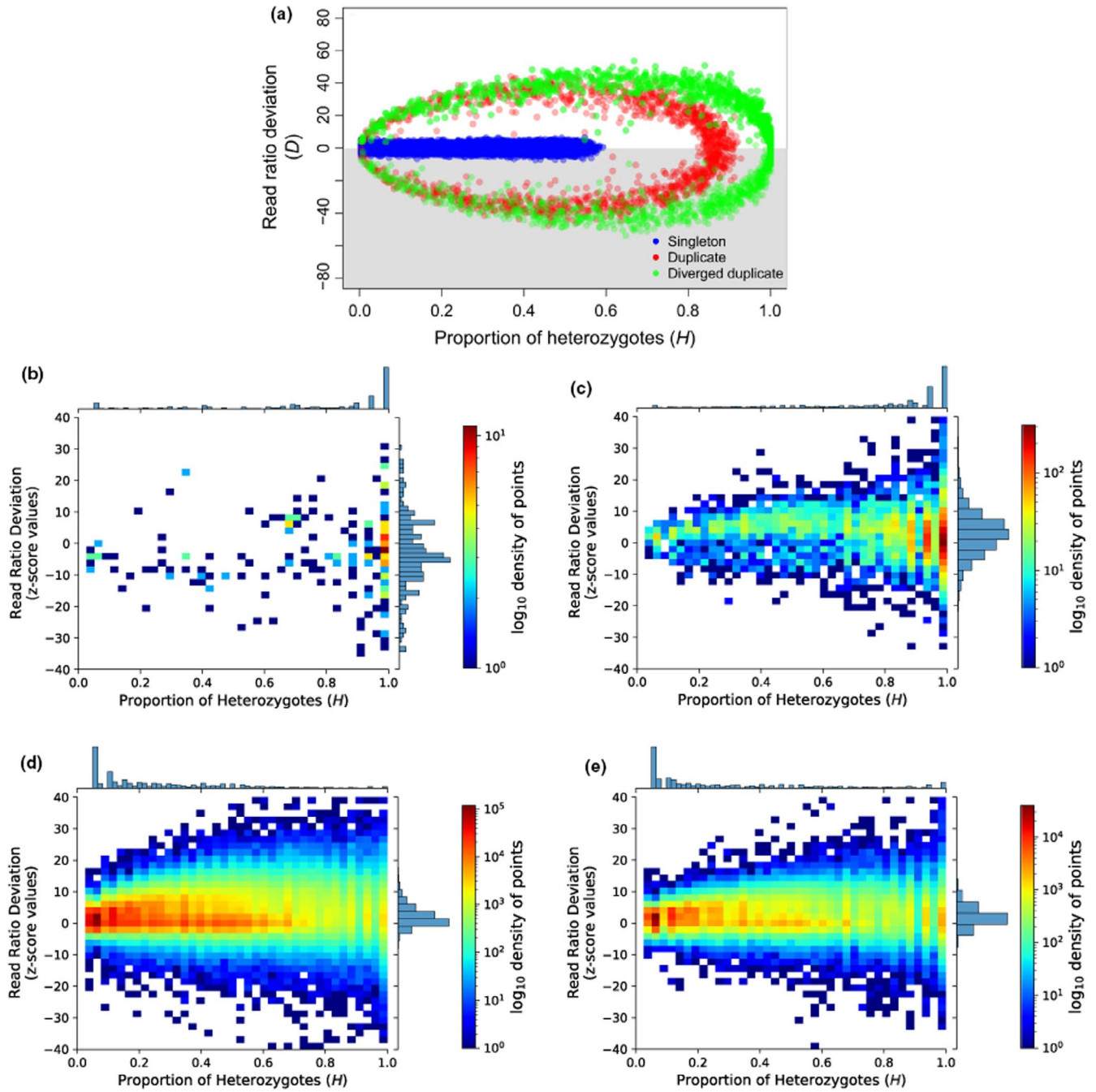

FIGURE 4 Standard deviation of read ratio (*z*-score) and the percentage of heterozygotes (*H*) calculated from indSeq data distinguish paralog misalignments according to expectations (a, figure from McKinney et al., 2017), and empirically for Douglas-fir (b, d) and jack pine (c, e). (b, c) Empirical distribution of megaSNP sites (candidate paralog sites identified as heterozygote calls from haploid tissue) calculated using indSeq data for those sites that were also called in poolSeq data (i.e., intersection I1). (d, e) Empirical distribution of intersection I2 (indSeq and poolSeq intersection) calculated using indSeq data. Some of the distribution found in the grey box in 4a will be found in the upper white panel because we used the reference allele instead of randomly choosing the allele for each locus. Note colour scale changes for each figure to accentuate patterns in the data. Code to create these figures is available on GitHub (003_testdata_validate_megaSNPs.ipynb, Lind 2021c)

Hohenlohe et al., 2012). There are several methods by which to detect such problematic sites, such as filtering by coverage (Dou et al., 2012), disomic models such as Hardy-Weinberg proportions (Catchen et al., 2013; Chen et al., 2014; Hohenlohe et al., 2011), or gene annotation, though there are several shortcomings (see descriptions of these shortcomings in Table 1 of McKinney et al., 2017). When individual sequencing data is available for the same individuals or populations, such information can be used to isolate potentially paralogous sites from pool-seq exome capture studies (e.g., Rellstab et al., 2019; Shu & Moran, 2020). However, a potentially cost-saving alternative would be to sequence the haploid tissue of a single individual (if available). Even so, there may

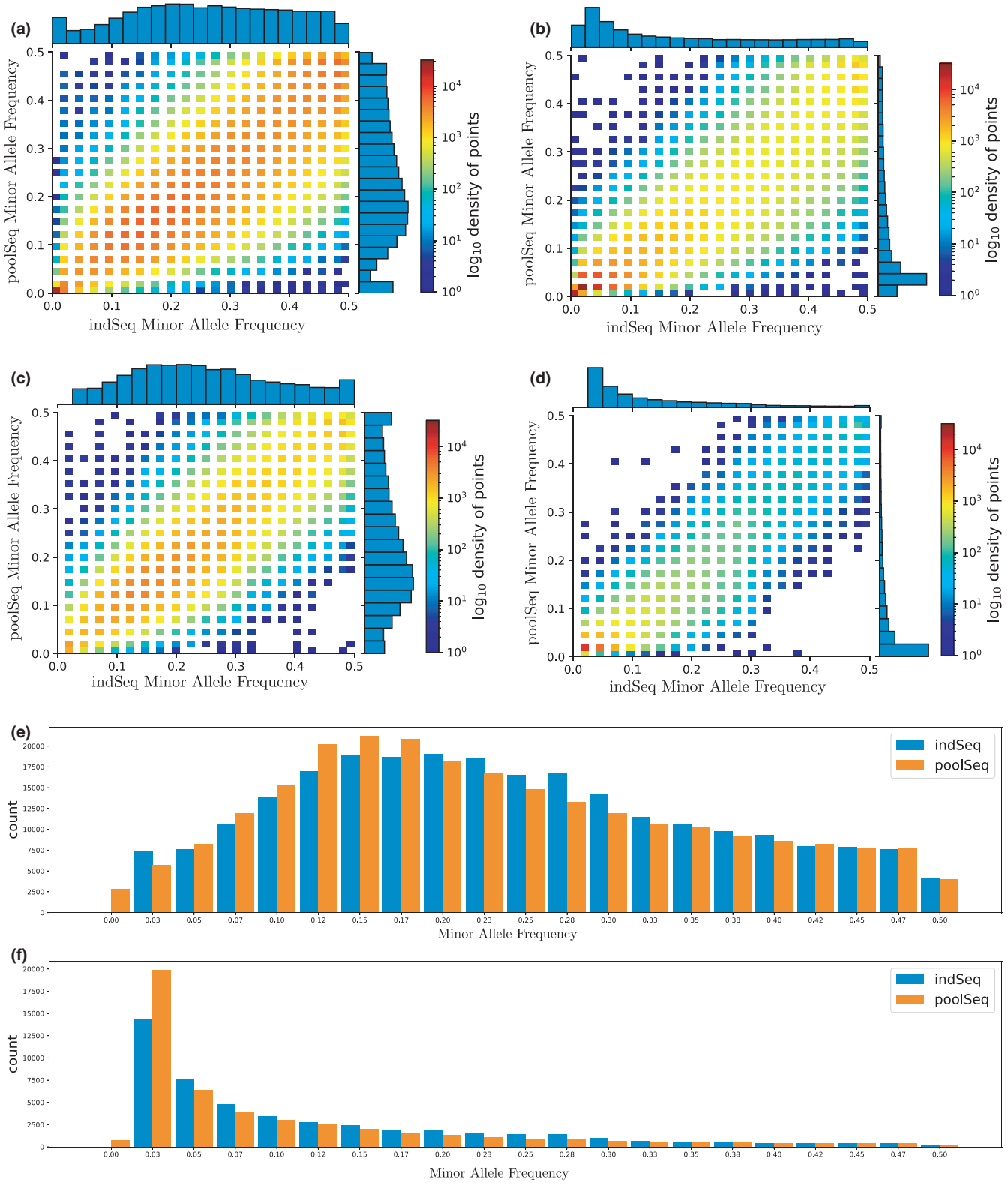

FIGURE 5 Congruence between indSeq and poolSeq (*x*- and *y*-axes, respectively a–d) minor allele frequency (MAF) estimates from Douglas-fir (a, c, e) and jack pine (b, d, f). (a, b) Two-dimensional (2D) histogram of baseline-filtered intersection between indSeq and poolSeq (i.e., intersection I2). (c, d) 2D histogram for SNPs after filtering intersection I2 for megaSNP sites,$H > 0.6$, $abs(z\text{-score}) > 10$, and indSeq sites with >20% missing data. (e, f) Congruence of minor allele frequency spectra from SNPs in (c, d). Colour scale is standardized to visualize differences in density between filtering steps. Minor allele frequency was binned for visualization purposes. Code to create these figures is available on GitHub (002_testdata_compare_AFs.ipynb, Lind 2021c)

be reduced power to detect recently diverged paralogs (i.e., when derived alleles are at low frequency and therefore not readily detected in a single individual), and an exploration varying the number and source population of haploid tissue for future studies could be used to more precisely quantify the effect and consistency of such data across sample sizes. As such, heterozygous SNPs called from our haploid data (megaSNPs) allowed us to identify variation from putative paralogous misalignments that infrequently displayed patterns expected of singleton gene copies. Indeed, high quality heterozygous calls from haploid sequencing are a reliable method for identifying misalignments due the known monoallelic state of the sequenced site (Limborg et al., 2016). While metrics from sequences of individuals are reliable (McKinney et al., 2017), they can falsely flag potentially paralogous sites as SNPs due to the stochastic nature of the sequencing process and may result in the exclusion of biologically meaningful information.

The accurate estimation of allele frequencies from pool-seq data will often depend on adequate depth of coverage and individuals, as well as thoughtful consideration of wetlab procedures and aspects of genomic resources and organismal biology. As pointed out by Rellstab et al. (2019), use of exome capture in many *Pinaceae* species will require particular care to exclude potentially paralogous sites from downstream analysis to avoid biased results. This is particularly true for pool-seq data sets relying on read counts for allele frequency estimation or population genetic inferences such as genotype-environment associations. While individually sequenced data sets may be one path forward to identifying such problematic sites (as in Rellstab et al., 2019; Shu & Moran, 2020), the sequencing of sufficient quantities of DNA from haploid gametophyte tissue available for some plants, including conifers, seedless vascular plants, and bryophytes, offers an alternate path forward to balance sequencing cost and data reliability, particularly for organism using diverged and or highly fragmented reference genomes.

ACKNOWLEDGEMENTS

The CoAdapTree project is funded by Genome Canada (241REF; Co-Project Leaders SNA, SY and Richard Hamelin), with co-funding from Genome BC and 16 other sponsors (http://coadaptree.forestry.ubc.ca/sponsors/), including Genome Alberta, Génome Québec, the BC Ministry of Forests, Natural Resources Operations and Rural Development, Alberta Innovates Bio Solutions, Vernon Seed Orchard Company, University of Alberta, University of British Columbia, Compute Canada, Mosaic Forest Management, and Western Forest Products. We thank Centre d'expertise et de services Génome Québec for sequencing service, University of Calgary Information Technologies for system support, Dr. Jürgen Ehlting and Jessica Wyatt at the University of Victoria for providing Douglas-fir samples infected with Swiss needle cast disease and preparing RNA samples for sequencing, as well as Dr. Pia Smets and Christine Chourmouzis for technical assistance. We also thank CoAdapTree Scientific Advisory Board members Drs. John Davis, Matias Kirst, and Graham Coop for their guidance, and three reviewers who provided helpful comments and suggestions. Dr. Sam Yeaman is also funded by the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN/03950-2017) and Alberta Innovates (20150252). Dr. Sally Aitken is also funded by NSERC Discovery Grant (RGPIN-2020-05136). The funding bodies did not have any role in the design of the study, collection, analysis, or interpretation of data in writing the manuscript.

AUTHOR CONTRIBUTIONS

Sally Aitken and Sam Yeaman obtained funding for the study. Sally Aitken, Sam Yeaman, Brandon Lind, and Mengmeng Lu conceived the research design with contributions from Tom Booker. Dragana Obreht Vidakovic generated the sequence data which Brandon Lind analysed, with contributions from Mengmeng Lu, Pooja Singh, and Tom Booker. Mengmeng Lu designed probes with contributions from Sam Yeaman. Brandon Lind and Mengmeng Lu designed SNP calling pipelines, which were coded by Brandon Lind and reviewed by Mengmeng Lu. Brandon Lind wrote the manuscript with contributions from Mengmeng Lu. All authors contributed to the editing of this manuscript.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

At the time of acceptance, all code needed to process raw sequence data through figure generation has been made available on GitHub and archived on Zenodo (Lind, 2021a, 2021b, 2021c). Commands used in probe design used default arguments unless otherwise specified. Raw sequence data was deposited on the Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA; bioproject PRJNA744263). Post-pipeline code in unstripped jupyter notebook format (Kluyver et al., 2016) has been archived to Zenodo (Lind, 2021c) following the addition of post-acceptance code needed to create and upload data to SRA and DataDryad.org—specific package versions used are often at the top of these notebooks via a dropdown html menu. The DataDryad.org archive (Lind et al., 2021) also includes the filtered SNP sets used for our analyses presented here, the assembled transcriptomes, configuration files (which includes ploidy and sample information) needed for each data set to run our GATK (Lind, 2021a) and `VarScan` (Lind, 2021b) pipelines using the fastq files in the SRA archive, as well as files with the metadata for SRA and biosample information (including accession information). NimbleGen SeqCap probes are available for jack pine under design name 180215_jackpine_v1_EZ_HX1, and for Douglas-fir under design name 80215_DOUGFIR_V1_EZ.

ORCID

Brandon M. Lind https://orcid.org/0000-0002-8560-5417
Mengmeng Lu https://orcid.org/0000-0001-5023-3759
Dragana Obreht Vidakovic https://orcid.org/0000-0003-1529-3347
Pooja Singh https://orcid.org/0000-0001-6576-400X

Tom R. Booker https://orcid.org/0000-0001-8403-6219
Sally N. Aitken https://orcid.org/0000-0002-2228-3625
Sam Yeaman https://orcid.org/0000-0002-1706-8699

REFERENCES

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: Climate change outcomes for tree populations. *Evolutionary Applications*, *1*, 95–111. https://doi.org/10.1111/j.1752-4571.2007.00013.x.

Alberto, F. J., Aitken, S. N., Alía, R., González-Martínez, S. C., Hänninen, H., Kremer, A., Lefèvre, F., Lenormand, T., Yeaman, S., Whetten, R., & Savolainen, O. (2013). Potential for evolutionary responses to climate change – Evidence from tree populations. *Global Change Biology*, *19*, 1645–1661. https://doi.org/10.1111/gcb.12181.

Allendorf, F. W., Bassham, S., Cresko, W. A., Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2015). Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *Journal of Heredity*, *106*, 217–227. https://doi.org/10.1093/jhered/esv015.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81. https://doi.org/10.1038/nrg.2015.28.

Casola, C., & Koralewski, T. E. (2018). *Pinaceae* show elevated rates of gene turnover that are robust to incomplete gene annotation. *Plant Journal*, *95*, 862–876. https://doi.org/10.1111/tpj.13994.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*, 3124–3140. https://doi.org/10.1111/mec.12354.

Chen, N., Van Hout, C. V., Gottipati, S., & Clark, A. G. (2014). Using Mendelian inheritance to improve high-throughput SNP discovery. *Genetics*, *198*, 847–857. https://doi.org/10.1534/genetics.114.169052.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

[dataset] Cronn, R., Dolan, P. C., Jogdeo, S., Wegrzyn, J. L., Neale, D. B., Clair, J. B., & Denver, D. R. (2017a). *3B24_fc1136_ln7_GTGGCC*. Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA). psme_diurnal3B-2400. BioSample: SAMN04168923; Sample name: Diurnal3B-2400; SRA: SRS1117152.

[dataset] Cronn, R., Dolan, P. C., Jogdeo, S., Wegrzyn, J. L., Neale, D. B., Clair, J. B., & Denver, D. R. (2017b). *Multi-genotype #2 (MG2_I)*. Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA). Multi-genotype #2. BioSample: SAMN00849794; Sample name: MG2; SRA: SRS308266.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R. & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics*, *14*, 436–440. https://doi.org/10.1214/aoms/1177731363.

Dou, J., Zhao, X., Fu, X., Jiao, W., Wang, N., Zhang, L., Hu, X., Wang, S., & Bao, Z. (2012). Reference-free SNP calling: Improved accuracy by preventing incorrect calls from repetitive genomic regions. *Biology Direct*, *7*, 17. https://doi.org/10.1186/1745-6150-7-17.

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, *23*, 40–69. https://doi.org/10.1111/mec.12581.

Fracassetti, M., Griffin, P. C., & Willi, Y. (2015). Validation of pooled whole-genome re-sequencing in *Arabidopsis lyrata*. *PLoS One*, *10*, e0140462–e140515. https://doi.org/10.1371/journal.pone.0140462.

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*, 207–218. https://doi.org/10.1534/genetics.110.114397.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., & Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Molecular Ecology*, *22*, 3766–3779. https://doi.org/10.1111/mec.12360.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*, 644–652.

[dataset] Hess, M., Wildhagen, H., Junker, L. V., & Ensminger, I. (2016). *GSM1893337: LA9S7; Pseudotsuga menziesii; RNA-Seq*. Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA). LA9S7. BioSample: SAMN04111182; SRA: SRS1089467; GEO: GSM1893337.

Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W., & Luikart, G. (2011). Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and west-slope cutthroat trout. *Molecular Ecology Resources*, *11*, 117–122. https://doi.org/10.1111/j.1755-0998.2010.02967.x.

Hohenlohe, P. A., Catchen, J., & Cresko, W. A. (2012). Population genomic analysis of model and non-model organisms using sequenced RAD tags. *Methods in Molecular Biology*, *888*, 235–260.

Holliday, J. A., Aitken, S. N., Cooke, J. E., Fady, B., González-Martínez, S. C., Heuertz, M., Jaramillo-Correa, J. P., Lexer, C., Staton, M., Whetten, R. W., & Plomion, C. (2017). Advances in ecological genomics in forest trees and applications to genetic resources conservation and breeding. *Molecular Ecology*, *26*, 706–717. https://doi.org/10.1111/mec.13963.

Janes, J. K., & Hamilton, J. A. (2017). Mixing it up: The role of hybridization in forest management and conservation under climate change. *Forests*, *8*, 237. https://doi.org/10.3390/f8070237.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks – A publishing format for reproducible computational workflows. In F. Loizides, & B. Schmidt (Eds.), *Positioning and Power in academic publishing: Players, agents and agendas* (pp. 87–90). IOS Press.

Koboldt, D., Zhang, Q., LarsoN, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, R., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*, 568–576. https://doi.org/10.1101/gr.129684.111.

Kovach, A., Wegrzyn, J. L., Parra, G., Holt, C., Bruening, G. E., Loopstra, C. A., Hartigan, J., Yandell, M., Langley, C. H., Korf, I., & Neale, D. B. (2010). The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics*, *11*, 420. https://doi.org/10.1186/1471-2164-11-420.

Krutovsky, K. V., Troggio, M., Brown, G. R., Jermstad, K. D., & Neale, D. B. (2004). Comparative mapping in the *Pinaceae*. *Genetics*, *168*, 447–461. https://doi.org/10.1534/genetics.104.028381.

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, Z., Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., Rieseberg, L. H., & Barker, M. S. (2015). Early genome duplications in conifers and other seed plants. *Science Advances*, *1*, e1501084. https://doi.org/10.1126/sciadv.1501084.

Limborg, M. T., Seeb, L. W., & Seeb, J. E. (2016). Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Molecular Ecology*, *25*, 2117–2129. https://doi.org/10.1111/mec.13601.

[dataset] Lind, B. M. (2021a). *GitHub.com/CoAdapTree/gatk_pipeline: Publication release (Version 1.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.5083321.

[dataset] Lind, B. M. (2021b). *GitHub.com/CoAdapTree/varscan_pipeline: Publication release (Version 1.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.5083302.

[dataset] Lind, B. M. (2021c). *GitHub.com/CoAdaptree/testdata_validation: Publication release (Version 1.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.5083292.

[dataset] Lind, B. M., Lu, M., Vidakovic, D. O., Singh, P., Booker, T., Aitken, S., & Yeaman, S. (2021). Data from: Haploid, diploid, and pooled exome capture recapitulate features of biology and paralogy in two non-model tree species. *Dryad Data Repository*, https://doi.org/10.5061/dryad.k0p2ngf7w.

Lind, B. M., Menon, M., Bolte, C. E., Faske, T. M., & Eckert, A. J. (2018). The genomics of local adaptation in trees: Are we out of the woods yet? *Tree Genetics & Genomes*, *14*, *29*. https://doi.org/10.1007/s11295-017-1224-y.

[dataset] Little, S. A., Boyes, I. G., Donaleshen, K., vonAderkas, P., & Ehlting, J. (2016) *RNA-Seq of Douglas-fir megagametophytes at different development points and of cone bracts and scales*. Sequence Read Archive of the National Center for Biotechnology Information (NCBI SRA). June10_Poll. BioSample: SAMN05255397; SRA: SRS1697751; GEO: GSM2202773.

Liu, Y.-Y., Yang, K.-Z., Wei, X.-X., & Wang, X.-Q. (2016). Revisiting the phosphatidylethanolamine-binding protein (PEBP) gene family reveals cryptic FLOWERING LOCUS T gene homologs in gymnosperms and sheds new light on functional evolution. *New Phytologist*, *212*, 730–744.

Lu, M., Krutovsky, K. V., Nelson, C. D., Koralewski, T. E., Byram, T. D., & Loopstra, C. A. (2016). Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). *BMC Genomics*, *17*, 1. https://doi.org/10.1186/s12864-016-3081-8.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicated genes. *Science*, *290*, 1151–1155.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

[dataset] McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2016). Data from: Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping by sequencing data from natural populations. *Dryad Data Repository*, https://doi.org/10.5061/dryad.cm08m.

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*, 656–669. https://doi.org/10.1111/1755-0998.12613.

Morse, A. M., Peterson, D. G., Islam-Faridi, M. N., Smith, K. E., Magbanua, Z., Garcia, S. A., Kubisiak, T. L., Amerson, H. V., Carlson, J. E., Nelson, C. D., & Davis, J. M. (2009). Evolution of genome size and complexity in *Pinus*. *PLoS One*, *4*, e4332. https://doi.org/10.1371/journal.pone.0004332.

Neale, D. B., & Kremer, A. (2011). Forest tree genomics: Growing resources and applications. *Nature Reviews Genetics*, *12*, 111–122. https://doi.org/10.1038/nrg2931.

Neale, D. B., Martínez-García, P. J., De La Torre, A. R., Montanari, S., & Wei, X.-X. (2017). Novel insights into tree biology and genome evolution as revealed through genomics. *Annual Review of Plant Biology*, *68*, 457–483. https://doi.org/10.1146/annurev-arplant-042916-041049.

Neale, D. B., McGuire, P. E., Wheeler, N. C., Stevens, K. A., Crepeau, M. W., Cardeno, C., Zimin, A. V., Puiu, D., Pertea, G. M., Sezen, U. U., Casola, C., Koralewski, T. E., Paul, R., Gonzalez-Ibeas, D., Zaman, S., Cronn, R., Yandell, M., Holt, C., Langley, C. H., … Wegrzyn, J. L. (2017). The Douglas-fir genome sequence reveals specialization of the photosynthetic apparatus in *Pinaceae*. *G3 Genes|genomes|genetics*, *7*, 3157–3167. https://doi.org/10.1534/g3.117.300078.

Neale, D. B., & Savolainen, O. (2004). Association genetics of complex traits in conifers. *Trends in Plant Science*, *9*, 325–330. https://doi.org/10.1016/j.tplants.2004.05.006.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., … Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, *497*, 579–584. https://doi.org/10.1038/nature12211.

Parchman, T. L., Jahner, J. P., Uckele, K. A., Galland, L. M., & Eckert, A. J. (2018). RADseq approaches and applications for forest tree genetics. *Tree Genetics & Genomes*, *14*, 39. https://doi.org/10.1007/s11295-018-1251-3.

Pavy, N., Lamothe, M., Pelgas, B., Gagnon, F., Birol, I., Bohlmann, J., Mackay, J., Isabel, N., & Bousquet, J. (2017). A high-resolution reference genetic map positioning 8.8K genes for the conifer white spruce: Structural genomics implications and correspondence with physical distance. *Plant Journal*, *90*, 189–203.

Puritz, J. B., & Lotterhos, K. E. (2018). Expressed exome capture sequencing: A method for cost-effective exome sequencing for all organisms. *Molecular Ecology Resources*, *18*, 1209–1222. https://doi.org/10.1111/1755-0998.12905.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rellstab, C., Dauphin, B., Zoller, S., Brodbeck, S., & Gugerli, F. (2019). Using transcriptome sequencing and pooled exome capture to study local adaptation in the giga-genome of *Pinus cembra*. *Molecular Ecology Resources*, *19*, 536–551.

Rellstab, C., Zoller, S., Tedder, A., Gugerli, F., & Fischer, M. C. (2013). Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. *PLoS One*, *8*, e80422. https://doi.org/10.1371/journal.pone.0080422.

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*, 749–763. https://doi.org/10.1038/nrg3803.

Scott, A. D., Zimin, A. V., Puiu, D., Workman, R., Britton, M., Zaman, S., Caballero, M., Read, A. C., Bogdanove, A. J., Burns, E., Wegrzyn, J., Timp, W., Salzberg, S. L., & Neale, D. B. (2020). A reference genome sequence for giant sequoia. *G3 Genes|genomes|genetics*, *10*, 3907–3919. https://doi.org/10.1534/g3.120.401612.

Shu, M., & Moran, E. V. (2020). Testing pipelines for genome-wide SNP calling from genotyping-by-sequencing (GBS) data for *Pinus ponderosa*. *Researchsquare*. https://doi.org/10.21203/rs.3.rs-32336/v1.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Sniezko, R. A., & Winn, L. A. (2017). Conservation and restoration of forest trees impacted by non-native pathogens: The role of genetics and tree improvement. In R. A. Sniezko, G. Man, V. Hipkins, K. Woeste, D. Gwaze, J. T. Kliejunas, & B. A. McTeague tech. cords. *Gene conservation of tree species—Banking on the future* (Vol. *963*, p. 68). Proceedings of a workshop. Gen. Tech. Rep. PNW-GTR-963. US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., Aitken, S. N., & Holliday, J. A. (2016). Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, *16*, 1136–1146. https://doi.org/10.1111/1755-0998.12570.

Voronova, A., Viktorija, B., Korica, A., & Rungis, D. (2017). Retrotransposon distribution and copy number variation in gymnosperm genomes. *Tree Genetics & Genomes*, *13*, 88. https://doi.org/10.1007/s11295-017-1165-5.

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., Dougherty, W. M., Lin, B. Y., Zieve, J. J., Martínez-García, P. J., & Holt, C. (2014). Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics*, *196*, 891–909.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Lind, B. M., Lu, M., Obreht Vidakovic, D., Singh, P., Booker, T., Aitken, S., & Yeaman, S. (2022). Haploid, diploid, and pooled exome capture recapitulate features of biology and paralogy in two non-model tree species. *Molecular Ecology Resources*, 22, 225–238. https://doi.org/10.1111/1755-0998.13474