Testing pipelines for genome-wide SNP calling from Genotyping-By-Sequencing (GBS) data for Pinus ponderosa

Mengjun Shu (MSHU@UCMERCED.EDU)
University of California Merced https://orcid.org/0000-0002-6323-2664
Emily V. Moran
University of California Merced

Research

Keywords: SNP, GBS, reduced-representation sequencing, pipeline, de novo, reference-based

Posted Date: June 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-32336/v1

Abstract

Background

Single Nucleotide Polymorphism (SNP) markers have rapidly gained popularity due to their abundance in most genomes and their amenability to high-throughput genotyping techniques. Reduced-representation restriction-enzyme-based sequencing methods (GBS or RADseq) have been demonstrated to be robust and cost-effective genotyping methods. While previous studies have shown that alignment of the short-read fragments to a genome sequence results in better SNP calling than *de novo* approaches, only a few tree species - and few conifers in particular - have an annotated sequence. While these could be used to align sequence fragments from related species, sequence divergence might result in SNPs being missed if they are in fragments that don't align properly. Producing a new annotated genome sequence for every conifer species before SNP analyses are conducted is still prohibitive, as many conifer genomes are huge (> 19 GB) and include a large proportion of repeat sequences, making assembly difficult. Here we compare four bioinformatics pipelines, two of which require a reference genome (TASSEL-GBS V2 and Stacks), two of which are *de novo* pipelines (UNEAK and Stacks). We used Illumina sequence data from 94 ponderosa pines, with loblolly pine as the reference genome.

Results

The number of SNPs called was much lower without a reference genome (62–196 thousand vs. 2.1–2.7 million SNPs). UNEAK was the fastest overall and identified more SNPs than Stacks *de novo*. Stacks with a reference genome produced the highest number of SNPs with lowest proportion of paralogs, while SNPs identified by TASSEL-GBS V2 exhibited the highest heterozygosity, minor allele frequency, and proportion of paralogs. More SNPs were uniquely identified by Stacks than TASSEL, though there was high overlap between methods.

Conclusion

The present case study provides a comprehensive comparison between four commonly-used SNP calling pipelines, and identifies the Stacks reference-based approach as the best overall for conifers (or other species with large repetitive genomes) that do not have a published reference genome for the same species. However, all four pipelines had distinct benefits and limitations, with Stacks for instance being less user-friendly than some of the other pipelines. In addition, researchers studying other conifer species using similar approaches should be prepared to analyze very large numbers of SNPs.

Background

Single Nucleotide Polymorphisms (SNPs) have been widely used for plant genomic studies, including genome-wide association studies, marker-assisted breeding and genomic selection, because of their

abundance in most genomes and amenability to high-throughput, cost effective genotyping technologies [1–3]. Reduced-representation sequencing approaches using multiplexed next-generation sequencing (NGS) and restriction enzyme based genome complexity reduction, often referred to as Genotyping-by-Sequencing (GBS) or RADseq (restriction site-associated DNA sequencing) have emerged as a cost-effective strategies for genome-wide SNP discovery and genotyping without the need for a reference genome [4–7].Moreover, they have the potential to reach regions of the genome involved in transcription regulation that are inaccessible to sequence capture approaches that target coding sequences [8]. While "GBS" and "RADseq" are sometimes used as umbrella terms for all such techniques, here we use GBS to refer to a specific approach that uses fewer steps than RADseq and lower sequencing depth [5].

However, genotyping and identifying SNPs is a challenge in most conifer species, due to their extremely large (19–32 Gb) and highly repetitive genomes. GBS has been tested for conifers on small numbers of individuals (< 10) and has been found to produce tens of thousands of SNPs with high coverage [9, 10]. However, the use of GBS on conifers species is still largely limited by the difficulty of genome-wide SNP calling from the massively parallel short-read sequences [11, 12]. Even though GBS only sequences a fraction of the genome, because conifer genomes are so large and repetitive the datasets produced still present a computational challenge.

Studies now commonly use advanced analysis pipelines to filter, sort, and align the GBS raw data to get SNP data. There are two general types of pipelines for handling GBS data: reference-based and *de novo* approaches. Reference-based pipelines call SNPs by mapping the raw GBS data to an existing reference genome to identify the position of sequences and compare the sequences from the same position to call SNPs [13]. Several reference-based pipelines have been widely used, including: TASSEL-GBS (v1 and v2), Stacks, IGST, and Fast-GBS [14–17]. In the absence of a reference genome, *de novo* pipelines identify pairs of nearly identical reads (presumed to represent alternative alleles of a locus) to call SNPs. Two *de novo* pipelines are commonly used: the Universal Network Enabled Analysis Kit (UNEAK) [18], and Stacks [16].

Previous studies have generally found that alignment to a reference genome from the same species increases the number of identifiable SNPs compared to the *de novo* pipelines [19]. However, it is unknown which pipeline is best for SNP calling in species that lack a sequenced genome. This includes most conifer species [20–22]. Though aligning sequences to the reference genome of a closely related species could allow for more SNPs to be identified if sequences are fairly conserved, it could also result in many sequence fragments being rejected (and therefore SNPs in these fragments not being identified) if this is not the case.

No reference genome is available for ponderosa pine (*Pinus ponderosa*), but one does exist for loblolly pine (*Pinus taeda)* [22, 23]. Of the conifers that have been sequenced to date, *P. taeda* is the most closely related to *P. ponderosa*; both are classified in the subgenus *Pinus* as opposed to *Strobus* [24, 25], which contains the other sequenced pine, *P. lambertiana* [21]. Furthermore, the *P. taeda* reference genome was used to successfully used to design probes for sequence capture in *P. contorta* [26, 27]. Recent studies

show that within this subgenus, *P. taeda* and *P. ponderosa* diverged more recently from each other than either did from lodgepole pine (*P. contorta*) [28, 29], suggesting that there is likely substantial sequence similarity between *P. taeda* and *P. ponderosa* as well. Previous studies have used *de novo* pipelines such as UNEAK to identify > 10,000 SNP loci in conifers that lack a full genome sequence [9, 10]. However, these earlier studies were based on a small number of samples, usually 6 individuals. Inclusion of more individuals will likely increase the number of SNPs identified – but by how much, and will the inclusion of more individual-level variation change the relative efficiency of different pipelines?

Despite the many advantages of GBS data, its reliability for SNP calling is compromised by the presence of paralogous genomic regions. Especially for the large genomes of conifers, involving both polyploidy and repetitive element activity [30], it is challenging to separate multiple copies in a genome (e.g. paralogs) from variants at a single locus due to sequence similarity and the short sequences obtained. Moreover, it is largely unknown which pipeline does a better job at filtering out the paralogs.

In this study, we sequenced 94 individual *P. ponderosa* using GBS and compared four pipelines for SNP calling, including two reference based pipelines (TASSEL-GBS V2, Stacks), and two *de novo* pipelines (UNEAK, Stacks). We first tested the performance of various restriction enzymes for fragmentation of *P. ponderosa* genome, and then used the best for GBS library construction. Then we applied the TASSEL-GBS V2 [17] and Stacks [16] pipelines using the reference genome of *P. taeda*, as well as the Stacks and UNEAK [18] pipelines without a reference genome. Our aim was to determine which method produced the most SNPs, which produced the least amount of missing data for the SNPs identified, and how much overlap there is in the SNPs called between method, as well as the proportion of paralogs among the SNPs called by different pipelines.

Results

Restriction enzyme selection

Figure 1 shows the amplified fragment size distributions of libraries from ponderosa pine DNA digested with different restriction enzymes. *Ape*KI yielded a high smooth curve of fragment sizes between 150 and 500, the sequencing size range for GBS. There were no discrete peaks suggesting repetitive DNA fragments present in *Ape*KI library, while other three REs had a few discrete peaks. *Pst*I performed similarly, though the curve was more jagged. EcoT22I produced a lower, more jagged curve, while EcoT22I + P*st*I had the worst performance of all. We therefore selected *Ape*KI. This enzyme does not cut CpG methylated sequences [31], and therefore tends to avoid stably silenced portions of the genome, hopefully increasing the proportion of SNPs from more actively transcribed regions.

Sequence quality of raw reads

Quality control for the raw reads involves the analysis of sequence quality, GC content, sequence length distribution, the presence of adaptors, overrepresented sequences, sequence duplication levels in order to detect sequencing errors, PCR artifacts, or contamination. Reducing the error rate of base calls and

improving the accuracy of the per-base quality score are integral to having reliable GBS raw data [13]. The sequence quality of the raw reads was high, with the per base sequence quality score over 32 and the most frequently observed mean quality score per sequence over 40. This indicates that the sequencing error is less than 0.1%.

The per base GC content module measures the GC content across the whole length of each sequence in a file and compares it to a modeled normal distribution of GC content. For the raw data of our study, the per base GC content is a roughly normal distribution with both the shape and peak corresponding to the distribution of GC content of the underlying genome, which indicates a normal random library without any bias.

According to the sequence length distribution plot, the length for all the sequences is, as expected, 90 bp for one set of samples and 100 bp for the other. The duplicate sequences analysis issues an error since non-unique sequences make up more than 50% of the total, which is in line with the high proportions of repetitive sequences in conifers [32, 33]. This is a feature that could be problematic for SNP calling; however, each pipeline has its own method for cleaning the data that can be more or less effective at removing repetitive sequences. No sequence represented more than 0.1% of the total, indicating that the library was not contaminated.

Comparison of 4 SNP-calling pipelines

Even with reduced representation sequencing, due to the large genome size of pines, the raw data for each of the two sets of 47 samples was over 19 GB after compression. Large computing resources (a remote cluster or supercomputer) are needed to run these pipelines. The performance of the four SNP-calling pipelines differed in many respects (Table 1). Of the two *de novo* pipelines, Stacks identified fewer SNPs than UNEAK (62,882 vs. 196,698) and took much longer to run than any of the other three pipelines (over 53 hours). Of the two reference-based analyses, Stacks identified 25% more SNPs than TASSEL-GBS V2 and took about 57% longer to run. The two reference-based pipelines identified 1–2 orders of magnitude more SNPs than the two *de novo* pipelines. For the Stacks pipeline, the reference-based version identified over forty times as many SNPs as the *de novo* one with a shorter run time.

Table 1
Comparison of different SNP-calling approaches.

Approach	*de novo*		reference-based	
Pipeline	**Stacks**	**UNEAK**	**TASSEL-GBS V2**	**Stacks**
Run time (hours: min)	53:26	2:17	21:8	33:45
Number of good reads (billion)	7.5	7.3	6.0	7.5
Percent of good reads (%)	96.2	93.6	76.9	96.2
Total SNPs	62,882	196,698	2,131,362	2,705,038
Missing data (%)	72.3	73.9	47.4	76.0
Average MAF	0.275	0.093	0.273	0.217
Observed heterozygosity	0.258	0.044	0.306	0.066
Expected heterozygosity	0.334	0.147	0.348	0.288
Average read depth per individual (Standard Deviation)	13.2 (2.2)	4.6 (0.6)	22.5 (5.5)	5.8 (1.0)
Paralogs (%)	1.5	1.1	18.5	1.0

The SNP quality data includes good reads, missing data, average MAF, and average observed and expected heterozygosity, average read depth per individual, and the proportion of paralogs. There were 7.8 billion total reads for the 94 samples. All the five pipelines used the same quality score (20) and same length (64 bp) to clean and trim the raw data. However, the number of reads considered "good" differed between pipelines, with TASSEL-GBS V2 keeping only 76.9% of reads, while the others kept at least 93.6% (Table 1). This resulted in TASSEL-GBS V2 having a much lower missing genotype rate (47.4% vs > 72%). The TASSEL-GBS V2 pipeline produced the largest average read depth per individual (22.5 vs. < 5). The relatively low read depth of Stacks reference-based pipeline (5.8) and Stacks *de novo* pipeline (4.6) is consistent with their high percentages of missing genotype calls.

The UNEAK pipeline produced a much smaller average MAF than the other pipelines (0.093 vs. > 0.21). This is likely due to UNEAK employing a network filter to discard repeats and paralogs. Accordingly, UNEAK produced a small proportion of paralogs (1.1%). The proportion of paralogs of TASSEL-GBS V2 pipeline is much higher than the other pipelines (18.5% vs. < 1.5%). The higher numbers of SNPs identified by TASSEL-GBS V2 pipeline is partly due to paralogs.

Interestingly, reference-based Stacks identified a very low average observed heterozygosity despite having SNPs with a relatively high minor allele frequency. Stacks *de novo* and TASSEL-GBS V2 had similar minor allele frequencies and expected heterozygosity, but the observed heterozygosity was higher for TASSEL-GBS V2. For all pipelines, the average observed heterozygosity is lower than expected heterozygosity,

which suggests that at least some loci are out of Hardy Weinberg equilibrium. This may be due to selection or genetic drift operating across the Sierra Nevada mountains, as the sampled individuals are widely distributed and do not represent a single random-mating population.

There are 1,888,913 overlapping SNPs identified by both reference-based pipelines (Fig. 2). Of the SNPs identified by TASSEL-GBS V2 11.4% were unique, while of those identified by the Stacks reference pipeline 30.2% were unique. Because the positions of SNPs were identified based on the reference genome, using the vcf-compare function, we were only able to compare the SNPs found using the two reference-based pipelines. Efforts to map SNPs identified by the *de-novo* approaches to the genome were stymied by the fact that the loblolly genome has not been fully assembled into chromosomes, and we were not able to develop a work-around for this that would enable software like VCFtools to be used.

Discussion

The repetitive DNA content in conifers affects the efficiency of SNP calling [10] and requires strategies for reducing the complexity and repetitive DNA content of GBS libraries. Selection of a restriction enzyme is one of the critical steps in GBS [5, 34]. In our study, the commonly-used restriction enzyme *Ape*KI performed well for ponderosa pine, with *Pst*I offering a decent second choice. Similarly, for lodgepole pine (*Pinus contorta*) and white spruce (*Picea glauca*), Chen et al. (2013) found that the size distribution curve was smoothest for *Ape*KI compared to *Pst*I and *Eco*T22I. *Ape*KI was also used for other conifers such as interior spruce, a hybrid complex of white spruce (*Picea glauca*) and Engelmann spruce (*Picea engelmannii*) [35]. Thus, *Ape*KI seems to be a good choice for conifers in general.

Table 2
SNP-calling approaches ranked

Approach	***de novo***		**reference-based**	
Pipeline	**Stacks**	**UNEAK**	**TASSEL-GBS V2**	**Stacks**
Run time	Highest	Lowest	Medium	Medium
Ease of use	Poor	Best	Medium	Poor
# of SNPs identified	Lowest	Low	High	Highest
Missing data	High	High	Lowest	High
% paralogs	Low (good)	Low (good)	Highest (poor)	Lowest (best)

As Table 2 indicates, no one pipeline was superior for all criteria that might be of concern for a researcher, though the reference-based Stacks did the best in terms of identifying a large number of SNPs with a low number of paralogs. However, it was more complex to use. All the steps in TASSEL-GBS V2 could deal with all the 94 samples together and assign the SNP data into each sample in the final VCF file. However, some steps in Stacks (e.g. ustacks, SAMtools) require separate codes for each sample instead of the one

code for the whole library, which takes much more effort due to individual code assignment. Some of the difference in performance can be explained by the alignment methods used.

All these pipelines assemble identical reads as tags/stacks before the alignment. The reference-based pipelines then align the tags/stacks with the reference genome to find their position, and then compare the tags/stacks in the same positions to identify SNPs with 1 bp mismatch (Fig. 3). Thus, the reference genome helps to ensure that tags from the same position are compared to identify SNPs. The *de novo* pipelines directly compare the tags/stacks with each other to identify SNPs with 1 bp mismatch (Fig. 4). In this situation, some of the reads from the same general position may not be identified as pairs because not enough of their short sequences overlap, and therefore some of the SNPs are missed.

Torkamaneh et al. (2016) conducted a comparison between different SNP calling pipelines on soybean (*Glycine max*) and found that four reference-based pipelines (TASSEL-GBS V1, IGST, TASSEL-GBSV2, Fast-GBS) identified more SNPs than either of two *de novo* pipelines (Stacks, UNEAK). However, the differences between the methods were much smaller than the differences found in our study. This suggests that for other non-model species without available genome sequences, SNP calling using a reference genome from closely related species can be an effective option.

The number of SNPs identified by the two *de novo* pipelines were very different. Torkamaneh et al. (2016) also found that the UNEAK pipeline identified more SNPs than the Stacks *de novo* pipeline. One possible explanation for this difference is the different way of assembling the identical reads as tags/stacks. For Stacks, the default setting for the maximum number of stacks at a single *de novo* locus in the program ustacks is 3. If there are over 3 stacks in the same locus, it will be blacklisted, meaning that locus will not be available for insertion into, or matching against, the catalogue. This is done as a means of rejecting repetitive sequences. However, the UNEAK merges the identical reads as tags without this limit. As a result, UNEAK pipeline can potentially identify most of SNPs because fewer stacks are rejected, but could also have more errors involving not properly separating paralogs. However, as discussed below, this did not appear to be the case; the percentage of paralogs was similar. Given this, and the more efficient identification of SNPs, we would recommend UNEAK over Stacks for *de novo* SNP identification. As noted in the methods, however, UNEAK is available in TASSEL V3.0, but not in the more recently updated versions of this software.

The different number of SNPs identified by the two reference-based pipelines is likely caused by a difference in how they assemble tags/stacks. TASSEL-GBS V2 assembles the identical reads first as tags first, and then align the tags to the reference genome. Stacks directly aligns the trimmed reads directly to the reference genome, which may lead to more alignments and a greater number of SNPs identified. TASSEL-GBS V2 rejected a higher proportion of reads initially (lower % considered "good") but produced a much lower percentage of missing data by either locus or individual than the other methods, which would mean less imputation will be needed at later steps in an association or genetic structure analysis. However, despite this thinning of reads, TASSEL-GBS V2 appears to be more likely to incorrectly identify SNPs from paralogs than the other three methods. Thus, for reference-based assembly, we would

recommend Stacks based on lower paralog percentages and higher SNP number, with the caveat that it is somewhat less user-friendly.

There are 1,888,931 SNPs identified by both of the reference-based methods. These SNPs comprised most (88.6%) of those identified by TASSEL-GBS V2. This pipeline exhibited the highest heterozygosity, MAF and proportion of paralogs, so some of the loci identified that did not overlap (11.4%) likely had unusually high heterozygosity. Stacks, which produced the highest number of loci, did so in part by identifying 816,107 SNPs that were not identified by TASSEL-GBS V2.

Finally, while earlier studies making use of < 10 individual conifers identified < 20,000 SNPs [9, 10], this study identified between 62,882 and 2,705,038 SNPs from 94 individuals. This indicates the high degree of genetic variation that is present in ponderosa pine [36], consistent with observations in other widespread conifer species [37]. While these individuals came from multiple populations within the Sierra Nevada mountains, this represents only a tiny fraction of the total range of this species, which extends from northern Mexico to southern Canada and from the Pacific to the Rocky Mountains. Future studies, especially those considering range-wide variation, should be prepared to analyze very high numbers of SNPs.

Conclusion

With the Illumina data generated by GBS from ponderosa pine, we compared four SNP calling pipelines, and identified the Stacks reference-based pipeline as the best in terms of identifying large numbers of SNPs while reducing false calls from paralogs. Use of a reference genome from a related pine species greatly increased the number of SNPs identified. Researchers studying other conifer species should be prepared to analyze very large numbers of SNPs, and to consider the benefits and limitations of different pipelines.

Methods

Sample preparation

In the 1970’s, the Forest Service's Pacific Southwest Regional Genetic Resources Program planted clones of 302 wild ponderosa pines in Chico, California. They came from diverse climate conditions in the central portion of California's Sierra Nevada mountains and are now reproductively mature, thus presenting an excellent resource for genetic studies. Although *P. ponderosa* contains multiple subdivisions, with the most important being between the Pacific and Rocky Mountain groups, based on their source locations the trees within the orchard likely do not cross any subdivision boundaries [24, 36, 38, 39].

For this study, we chose 94 individual *P. ponderosa* genotypes from the orchard collection. The source locations of these 94 genotypes are shown in Fig. 5 and the location information are listed in Table S1. The sample preparation included three steps: dry needle preparation, DNA extraction, and quantification.

Fresh needles were collected and dried with silica gel desiccant. Total genomic DNA was extracted from the dried needle tissue using DNeasy Plant Mini Kit (250) following the protocol from the manufacturer (Qiagen, Hilden, Germany) with two main modifications. First, to reduce protein contamination, for the step of grinding needles we added 1.5 ul of Proteinase K (20 mg/ml) along with the Buffer AP and RNase A. The MiniG 1600 from SPEX SampePrep (Metuchen, NJ, USA) was used to grind needles with automated mechanical disruption through bead beating. Second, at the very last step, the amount of AE elution buffer was changed from 100 µl to 50 µl to get a higher concentration of DNA (averagely 200 ng/µl). The DNA concentration was quantified using an Eppendorf BioSpectrometer (Eppendorf, AG, Germany).

Restriction enzyme selection

When working on a new species, it is beneficial to determine which enzyme produces the most fragments within the desired size range (100–400 bp). For optimization of the GBS protocol, 1000 ng samples of *P. ponderosa* genomic DNA were digested separately with *Ape*KI, *Pst*I, and *Eco*T22I, and with a combination of *Pst*I and *Eco*T22I (double digest) following the instructions of the enzyme manufacturer (New England Biolabs). These three restriction enzymes are methylation sensitive and have been previously used for construction of reduced complexity GBS libraries in conifers [9, 10]. Fragment size distributions of each test library were visualized using an Agilent BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA) with the High Sensitivity DNA Kit for quantification. For each test library, we used three samples with the same DNA concentration (50 ng/ul). We selected the enzyme based on the smoothness of the distribution and the size of the fragment sequences produced. Once this was done, all of the post-extraction steps were carried out at the UC Davis Genome Center.

Illumina libraries preparation and sequencing

For *Pinus contorta* and *Picea glauca* 47-plex GBS libraries yielded good results [9]. Therefore, a 48-plex GBS library consisting of 47 DNA samples and a negative control (no DNA) was prepared in our study. The GBS protocol was slightly modified from the standard protocol [5] and that of Chen et al. (2013). The library preparation and sequencing includes 6 steps: digestion, ligation, pooling samples, PCR, clean-up, and single-end read sequencing. DNA extracts (100 ng) were digested with the restriction enzyme *Ape*KI for at 75 °C for 2 hours. Each of the 47 ponderosa pine DNA samples was tagged with a unique barcode. The barcodes for each individual are listed in Table S2. Sequences for the *Ape*KI barcode adapters and the common adapters, and the temperature cycles, were as described in Chen et al. (2013). After the digestion, the samples were cooled to 4 °C, and then adapters were ligated onto restriction fragments. This was done using T4 DNA Ligase (Life Technologies, Burlington, ON, Canada) at 16 °C for 1 hour, after which samples were "heat killed" at 65 °C for 20 minutes. The pool was quantified via qPCR using the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA) for Illumina sequencing platforms, with 0.9X bead cleanup to remove small fragments (< 250 bp). Additional DNA purification using the Zymo DNA Clean & Concentrator kit (Zymo Research, Irvine, CA) was performed to further increase the purity of the extracted DNA. The fragments from all 47 samples were then sequenced

(single-end read 90 bp) on one lane of an Illumina HiSeq 4000 (Illumina, San Diego, CA). The same procedure was repeated for the other 47 samples (single-end read 100 bp). We then assessed the sequence quality of the raw reads using FastQC analysis [40].

SNP calling

We used the reference genome of loblolly pine v2.0 (https://treegenesdb.org/FTP/Genomes/Pita/) for the reference-based pipelines.

TASSEL-GBS V2 is implemented in TASSEL V5.0, a program originally developed for maize to facilitate genotype-phenotype comparisons (https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline). This pipeline requires a reference genome to call SNPs. The steps involved are illustrated on the left side of Fig. 3. The raw sequence data in FASTQ file are first trimmed to same length (64 bp) and then identical reads are assembled into tags (unique DNA sequences). These distinctive tags are saved into FASTQ file. Then alignment program bowtie2 [41] is used to align these tags with the reference genome of loblolly pine. Based on the position of the tags against the reference genome, SNPs are produced by identifying tags aligned in the same position that have a 1 bp mismatch. Finally, the SNP information within each tag for each sample is output as a VCF file. Each step is performed internally with TASSEL- GBS V2 plugins except for alignment, which is carried out externally with bowtie2. We used the default parameter settings for our analysis except that the minimum quality score was set to 20 to make the base call accuracy more than 99%.

Stacks is a software package developed for restriction site-associated DNA sequencing that identifies SNPs and calculates population statistics from any restriction enzyme-based, reduced-representation sequence data with short-read sequences (http://catchenlab.life.illinois.edu/stacks/). It was developed with population genomics in mind, and so aims to assemble loci in large numbers of individuals and read haplotypes from them. Stacks allows for SNP calling with or without a reference genome; we chose to do both. The detailed steps of Stacks reference pipeline are represented on the right side of Fig. 3. There are two main differences from the TASSEL-GBS V2 pipeline. First, the Stacks reference pipeline aligns the reads directly against the reference genome, while TASSEL-GBS V2 assembles the same reads into tags and then performs the alignment. Second, the BWA alignment program [42] is used instead of bowtie2. Each step in the Stacks reference pipeline is performed internally in Stacks algorithms except alignment with BWA and the SAMtools [43] step used to get read position.

The steps involved in the Stacks *de novo* pipeline are shown on the right side of Fig. 4. First, reads are demultiplexed, cleaned and trimmed to 64 bp, and identical reads are assembled as "stacks". The stacks in each sample are merged as catalogs, which then are grouped together across samples. Third, SNPs are identified by matching reads to the catalogs and assigning SNPs to each sample when there is a 1 bp mismatch. SNP information is saved in a VCF file. Optional additional steps include the creation of genetic maps and calculation of population statistics. Every step in Stacks *de novo* pipeline uses the Stacks internal algorithms. For both Stacks reference and *de novo* pipeline, we used the default

parameter settings except that the quality score limit is set to 20 instead of 15, for greater accuracy and to be consistent with TASSEL-GBS V2.

The UNEAK (Universal Network Enabled Analysis Kit) pipeline can be implemented in TASSEL V3.0 (https://tassel.bitbucket.io/TasselArchived.html), but it is not available in the more recent V5.0. UNEAK is a *de novo* pipeline that can call SNPs without a reference genome. The steps in the UNEAK pipeline are on the left side of Fig. 4. The general design of UNEAK is as follows: 1) raw Illumina DNA sequence data were first trimmed to 64-bp; 2) identical 64-bp reads for each individual are collapsed into tags; 3) pair-wise alignment identifies tag pairs having a single base pair mismatch. These single base pair mismatches are candidate SNPs, which are then assigned to each sample and saved as VCF file. As in the Stacks *de novo* pipeline, every step in UNEAK pipeline uses the internal algorithms. We again used the default parameter settings except that the base call accuracy is changed from 0.03 to 0.01, which is equivalent to the first two methods.

We ran most of the step on the Multi-Environment Research Computer for Exploration and Discovery (MERCED) cluster, a shared resource for UC Merced researchers, which has 128 GB RAM in each compute node. The exception was the step cstacks (Fig. 4, merge stacks as catalogs) in the Stacks *de novo* pipeline, which requires large memory and RAM. For this, we used the XSEDE supercomputing resource [44], which has 3000 GB RAM in each computer node.

SNP quality and comparison

To evaluate the quality of the SNPs in each VCF output file, six parameters were chosen: good reads, the missing genotype rate, minor allele frequency (MAF), heterozygosity, read depth, and the proportion of paralogs. We used PLINK 1.9 [45], a widely used open-source C/C++ toolset in population genetics, to calculate the missing genotype rate, MAF, and heterozygosity. Recent whole-genome duplications have occurred in conifers [30, 46], resulting in multiple paralogs. Such paralogs could yield false SNPs if incorrectly identified as a single locus based on short GBS sequence reads. To address this issue and distinguish real allelic variation from paralogs, we tested for deviations in ratio of read depth for each allele within heterozygotes in the GBS data [47]. The deviation of this ratio from its expected value (1:1) is expressed as a Z-score with a binomial distribution (P = 0.5). Based on these Z-scores, we declare likely paralog variants using a conservative threshold of $|Z| > 5$.

Besides the quality of the SNPs, we were also interested in how many SNPs were identified by more than one pipeline. In our study, the comparison of SNP overlap was done using VCFtools [48].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Funding:

The sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. For SNP identification, we made use of the MERCED computer cluster at UC Merced (supported by NSF Award ACI-1429783) and the Extreme Science and Engineering Discovery Environment (XSEDE; supported by NSF Award ACI-1548562).

Authors' contributions:

Mengjun Shu[¶]: Research design, performed research, analyzed data, wrote paper (corresponding author).

Emily Moran[&]: Research design, edited paper.

Acknowledgements

We thank the Forest Service's Pacific Southwest Regional Genetic Resources Program for allowing us to sample needles from their seed orchard. We also thank Jeffrey Lauder and Melaine Aubry-Kientz for their comments on this manuscript.

Data accessibility statement

Raw DNA sequencing data: available at National Center for Biotechnology Information under BioProject number PRJNA527618. https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA527618&go=go

Individual SNP genotypes: available on Dryad. DOI:

https://doi.org/10.5061/dryad.6fv8fb4

References

1. Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, et al. Association Genetics of Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-Hardiness Related

Traits. Genetics. 2009;182:1289–302.

2. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012;44:808–11.
3. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. PNAS. 2013;110:453–8.
4. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81–92.
5. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011;6:1–10.
6. Poland JA, Brown PJ, Sorrells ME, Jannink J-L. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLOS ONE. 2012;7:e32253.
7. Poland JA, Rife TW. Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome. 2012;5:92–102.
8. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. SNP Markers and Their Impact on Plant Breeding [Internet]. International Journal of Plant Genomics. 2012 [cited 2019 Mar 14]. Available from: https://www.hindawi.com/journals/ijpg/2012/728398/.
9. Chen C, Mitchell SE, Elshire RJ, Buckler ES, El-Kassaby YA. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics Genomes. 2013;9:1537–44.
10. Pan J, Wang B, Pei Z-Y, Zhao W, Gao J, Mao J-F, et al. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Molecular Ecology Resources. 2015;15:711–22.
11. Glenn TC. Field guide to next-generation DNA sequencers. Molecular Ecology Resources. 2011;11:759–69.
12. Goto S, Kajiya-Kanegae H, Ishizuka W, Kitamura K, Ueno S, Hisamoto Y, et al. Genetic mapping of local adaptation along the altitudinal gradient in Abies sachalinensis. Tree Genetics Genomes. 2017;13:104.
13. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12:443–51.
14. Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, et al. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS ONE. 2013;8:1–9.
15. Torkamaneh D, Laroche J, Bastien M, Abed A, Belzile F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics. 2017;18:5.

16. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013;22:3124–40.
17. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE. 2014;9.
18. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, et al. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLOS Genetics. 2013;9:e1003215.
19. Torkamaneh D, Laroche J, Belzile F. Genome-Wide SNP. Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. PLoS One [Internet]. 2016 [cited 2018 Mar 12];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4993469/.
20. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, et al. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics. 2013;29:1492–7.
21. Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, Cardeno C, et al. Sequence of the Sugar Pine Megagenome Genetics. 2016;204:1613–26.
22. Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marçais G, et al. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics. 2014;196:875–90.
23. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014;15:R59.
24. Willyard A, Cronn R, Liston A. Reticulate evolution and incomplete lineage sorting among the ponderosa pines. Molecular Phylogenetics and Evolution [Internet]. 2009 [cited 2016 Jan 5];52:498–511. Available from: http://www.sciencedirect.com/science/article/pii/S105579030900044X.
25. Gernandt DS, Hernández-León S, Salgado-Hernández E, Pérez de La Rosa JA. Phylogenetic relationships of Pinus subsection Ponderosae inferred from rapidly evolving cpDNA regions. Syst Bot. 2009;34:481–91.
26. Suren H, Hodgins KA, Yeaman S, Nurkowski KA, Smets P, Rieseberg LH, et al. Exome capture from the spruce and pine giga-genomes. Molecular Ecology Resources. 2016;16:1136–46.
27. Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, et al. Convergent local adaptation to climate in distantly related conifers. Science. 2016;353:1431–3.
28. Eckert AJ, Hall BD. Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): Phylogenetic tests of fossil-based hypotheses. Molecular Phylogenetics and Evolution [Internet]. 2006 [cited 2019 Aug 18];40:166–82. Available from: http://www.sciencedirect.com/science/article/pii/S1055790306000911.
29. Gernandt DS, Geada Lopez G, Garcia SO, Liston A. Phylogeny and classification of *Pinus*. Taxon. 2005;54:29–42.

30. Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, et al. Early genome duplications in conifers and other seed plants. Science Advances. 2015;1:e1501084.
31. Castel AL, Nakamori M, Thornton CA, Pearson CE. Identification of restriction endonucleases sensitive to 5-cytosine methylation at non-CpG sites, including expanded (CAG)n/(CTG)n repeats. Epigenetics. 2011;6:416–20.
32. Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, et al. The Pinus taeda genome is characterized by diverse and highly diverged repetitive sequences. BMC Genom. 2010;11:420.
33. Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, et al. Evolution of Genome Size and Complexity in Pinus. PLOS ONE. 2009;4:e4332.
34. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLOS ONE. 2012;7:e37135.
35. Gamal El-Dien O, Ratcliffe B, Klápště J, Chen C, Porth I, El-Kassaby YA. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genom. 2015;16:370.
36. 10.1007/s11295-015-0865-y
Potter KM, Hipkins VD, Mahalovich MF, Means RE. Nuclear genetic variation across the range of ponderosa pine (Pinus ponderosa): Phylogeographic, taxonomic and conservation implications. Tree Genetics & Genomes [Internet]. 2015 [cited 2015 Jun 22];11. Available from: http://link.springer.com/10.1007/s11295-015-0865-y.
37. Potter KM, Jetton RM, Dvorak WS, Hipkins VD, Rhea R, Whittier WA. Widespread inbreeding and unexpected geographic patterns of genetic variation in eastern hemlock (Tsuga canadensis), an imperiled North American conifer. Conserv Genet. 2012;13:475–98.
38. Burns RM, Honkala BH. Silvics of North America. Washington DC: USDA Forest Service; 1990.
39. Conkle MT, Critchfield WB. Genetic variation and hybridization of ponderosa pine. Ponderosa pine: the species and its management [Internet]. 1988 [cited 2014 Nov 6];27:43. Available from: http://www.fs.fed.us/psw/publications/conkle/psw_1988_conkle001.pdf.
40. Simon A. FastQC: a quality control tool for high throughput sequence data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
43. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93.
44. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, et al. XSEDE: Accelerating scientific discovery. Computing in Science Engineering. 2014;16:62–74.
45. Purcell S, Chang C. PLINK 1.9. URL https://www cog-genomics org/plink2. 2015.

46. Prunier J, Verta J-P, MacKay JJ. Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. New Phytol. 2016;209:44–62.
47. McKinney GJ, Waples RK, Seeb LW, Seeb JE. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Molecular Ecology Resources. 2017;17:656–69.
48. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

Figures

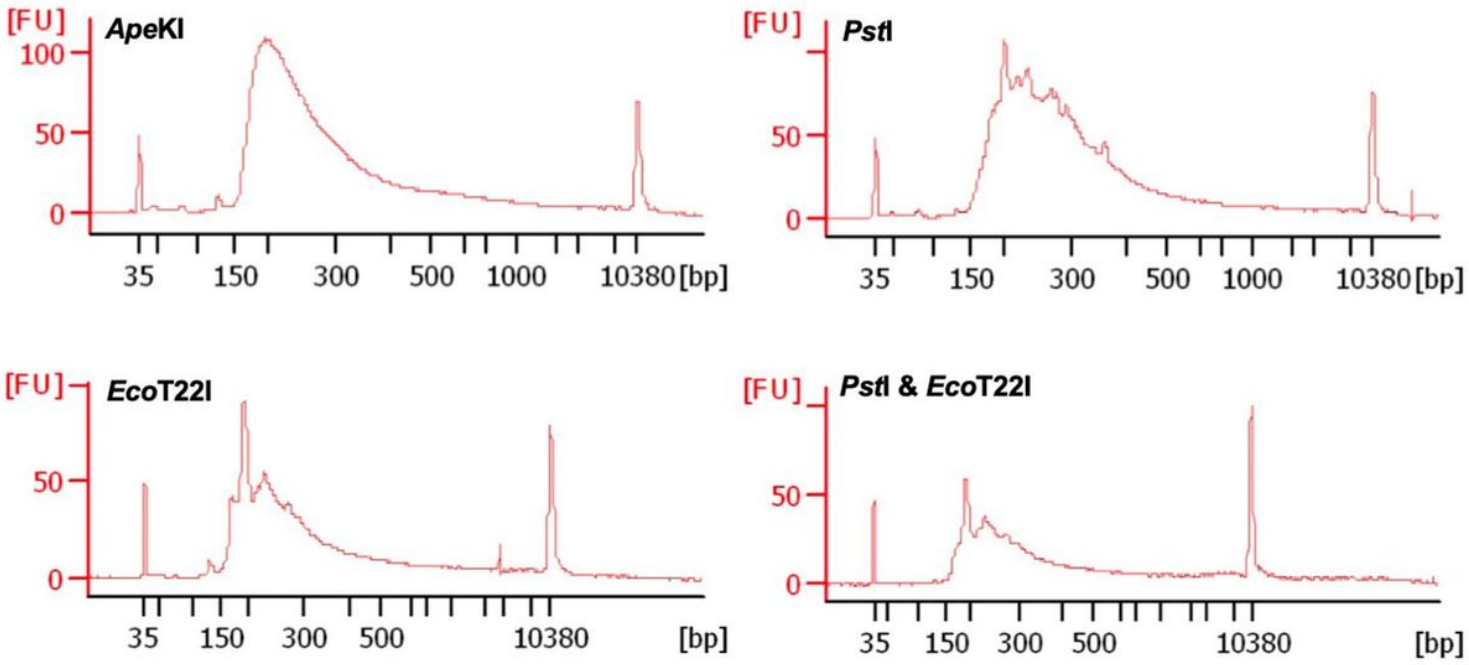

Figure 1

Fragment size distribution of GBS libraries with different restriction enzyme. The y-axis shows fluorescence units, indicating amount of DNA. Numbers below hatch marks on the x-axis indicate fragment size (bp).

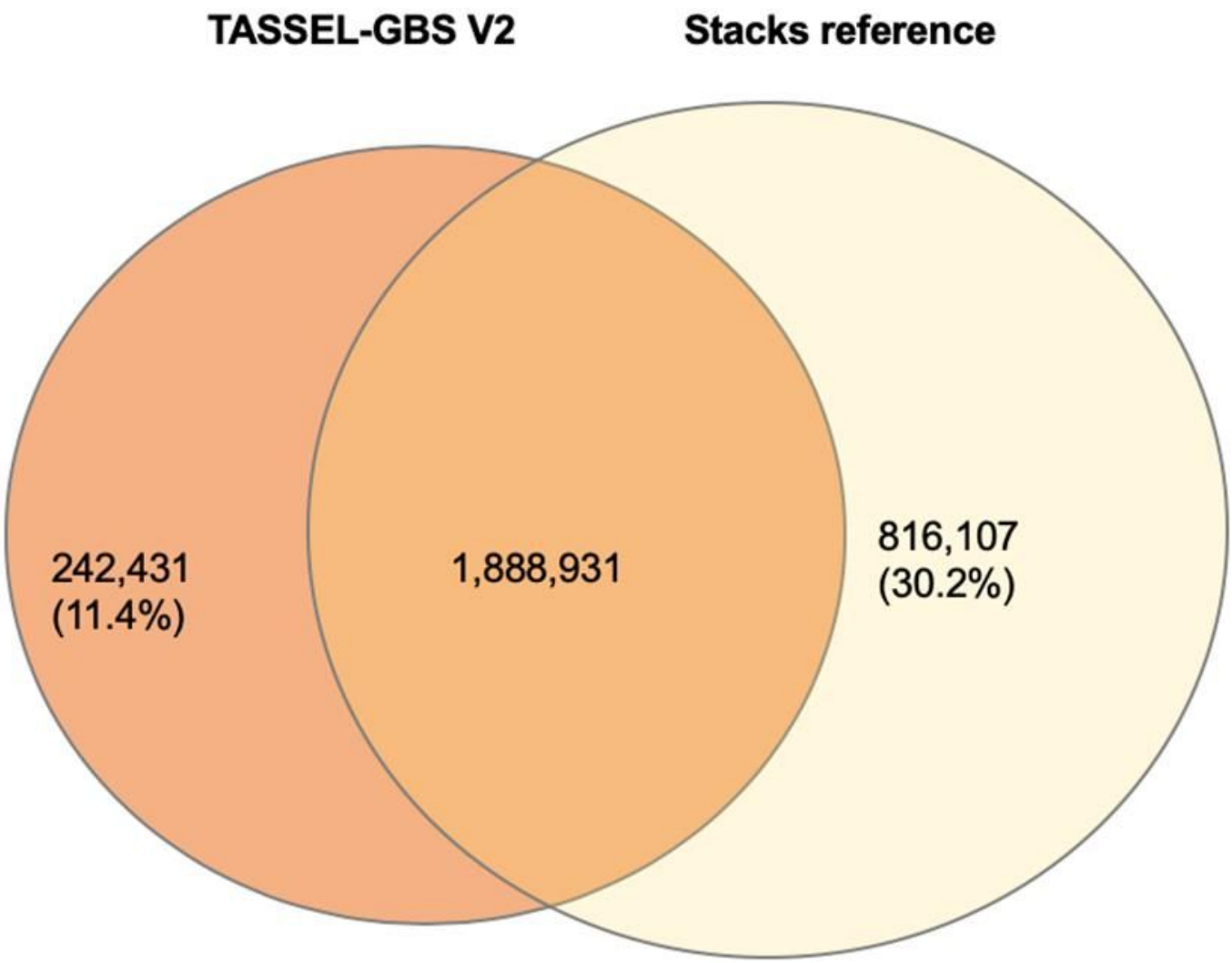

Figure 2

Venn diagram comparing SNPs overlap between the two reference-based pipelines. The circle on the left side represents the SNPs produced by TASSEL-GBS V2 pipeline. The circle on the right side represents the SNPs produced by Stacks reference-based pipeline.

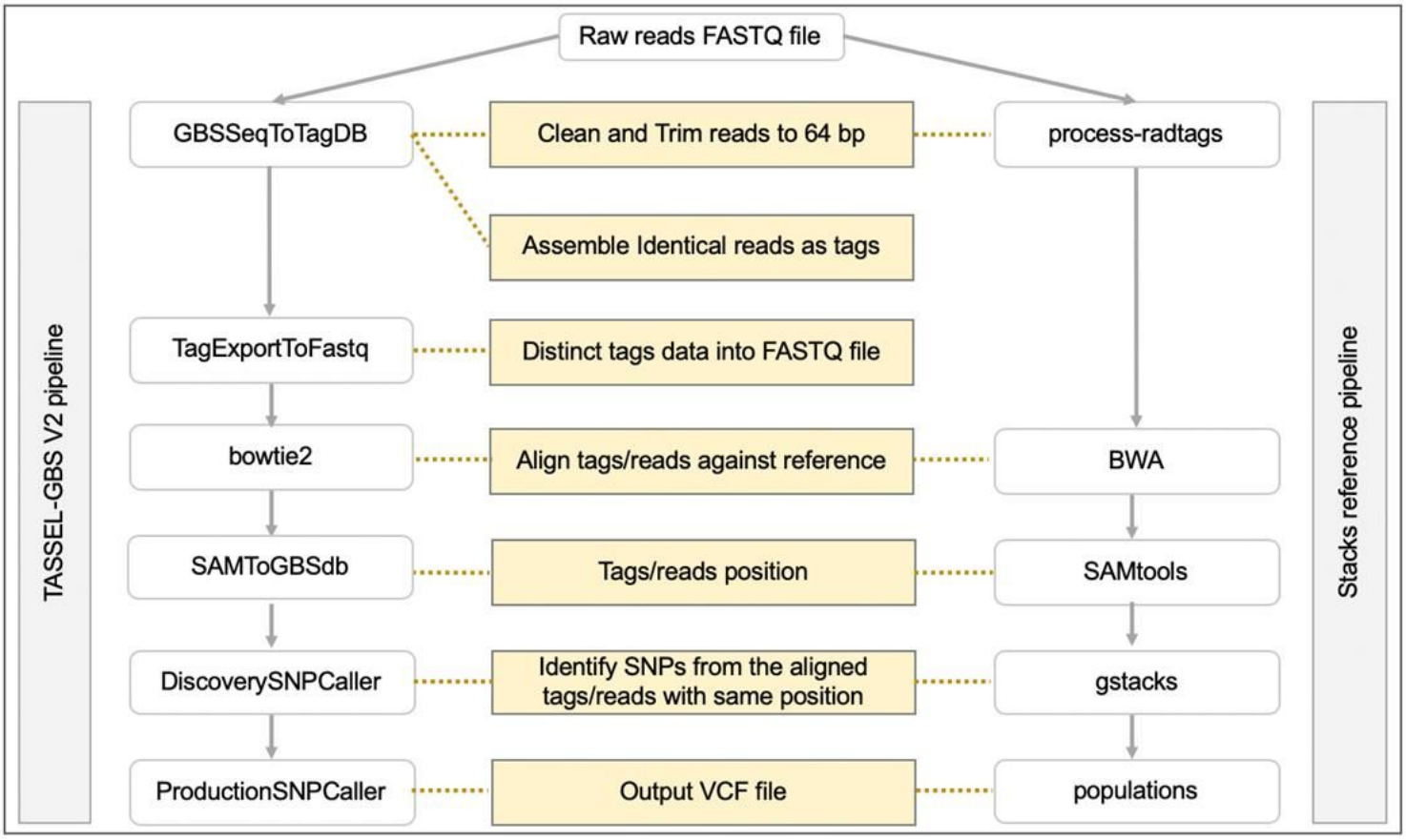

Figure 3

Comparison of the two reference-based pipelines. The horizontal boxes on the left side represent the programs in GBS V2. The horizontal boxes on the right side represent the programs in the Stacks reference pipeline. The yellow boxes in the middle represent potential program functions, while the yellow dotted lines specify the main function for each program in the two pipelines.

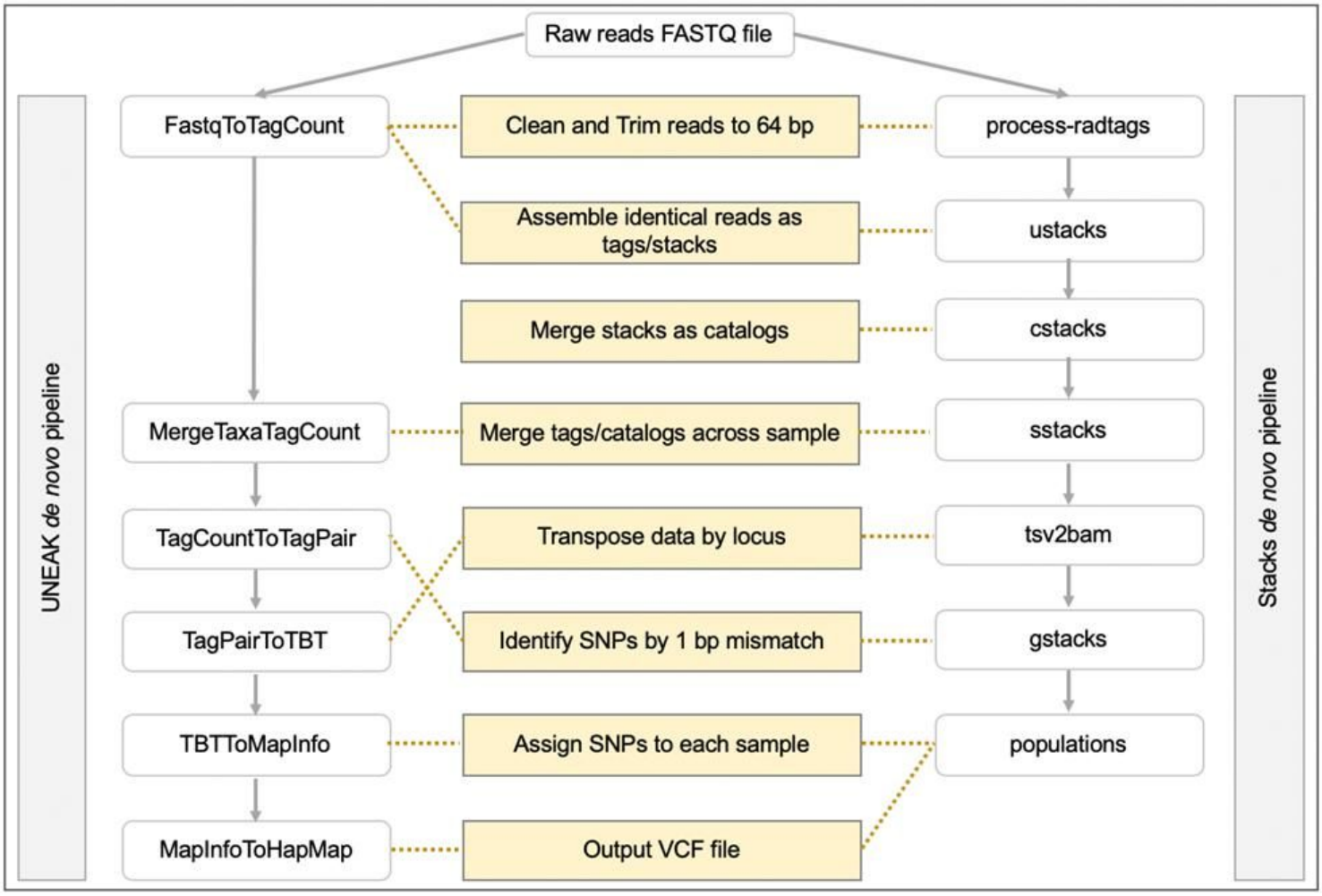

Figure 4

Comparison between two de novo pipelines. The horizontal boxes on the left side represent the programs in UNEAK de novo. The horizontal boxes on the right side represent the programs in Stacks de novo. The yellow boxes in the middle represent the functions of the program, while the yellow dotted lines specify the main function for each program in the two pipelines.

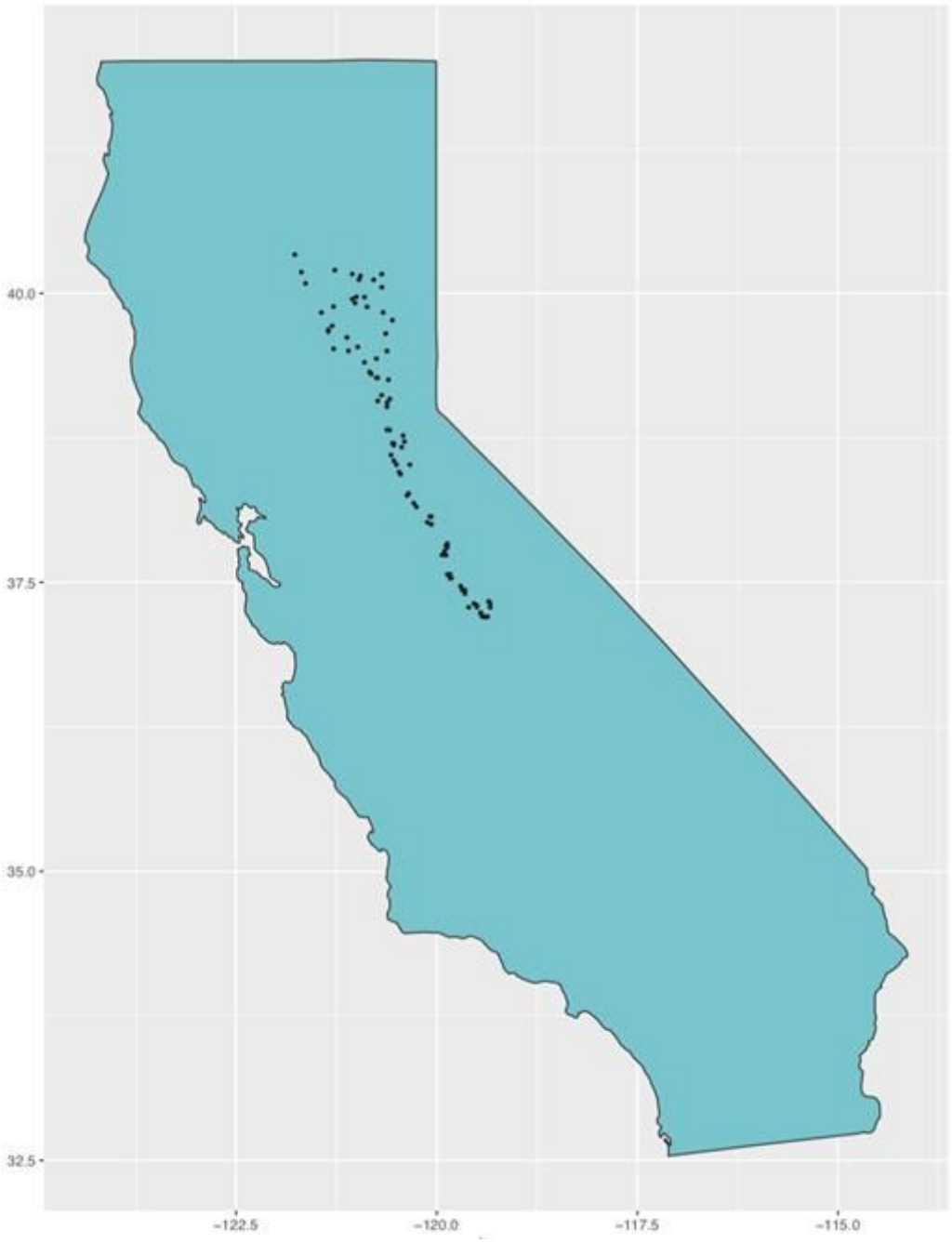

Figure 5

Geographic distribution of the 94 samples. The black dots represent original genotype source locations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1Table.xlsx
- S2Table.xlsx