PLOS ONE

RESEARCH ARTICLE

Validation of Pooled Whole-Genome Re-Sequencing in *Arabidopsis lyrata*

Marco Fracassetti[1]*, Philippa C. Griffin[1,2], Yvonne Willi[1]

1 Institute of Biology, Evolutionary Botany, University of Neuchâtel, Neuchâtel, Switzerland, **2** School of BioSciences, University of Melbourne, Parkville, Victoria, Australia

* marco.fracassetti@unine.ch

Citation: Fracassetti M, Griffin PC, Willi Y (2015) Validation of Pooled Whole-Genome Re-Sequencing in *Arabidopsis lyrata*. PLoS ONE 10(10): e0140462. doi:10.1371/journal.pone.0140462

Editor: Ulrich Melcher, Oklahoma State University, UNITED STATES

Received: January 30, 2015

Accepted: September 25, 2015

Published: October 13, 2015

Data Availability Statement: The raw data is available via the European Nucleotide Archive (accession number PRJEB8335).

Funding: This work was supported by the Swiss National Science Foundation (PP00P3-123396 and PP00P3-146342 to YW) and the Fondation Pierre Mercier pour la Science (to YW), Lausanne. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Sequencing pooled DNA of multiple individuals from a population instead of sequencing individuals separately has become popular due to its cost-effectiveness and simple wet-lab protocol, although some criticism of this approach remains. Here we validated a protocol for pooled whole-genome re-sequencing (Pool-seq) of *Arabidopsis lyrata* libraries prepared with low amounts of DNA (1.6 ng per individual). The validation was based on comparing single nucleotide polymorphism (SNP) frequencies obtained by pooling with those obtained by individual-based Genotyping By Sequencing (GBS). Furthermore, we investigated the effect of sample number, sequencing depth per individual and variant caller on population SNP frequency estimates. For Pool-seq data, we compared frequency estimates from two SNP callers, VarScan and Snape; the former employs a frequentist SNP calling approach while the latter uses a Bayesian approach. Results revealed concordance correlation coefficients well above 0.8, confirming that Pool-seq is a valid method for acquiring population-level SNP frequency data. Higher accuracy was achieved by pooling more samples (25 compared to 14) and working with higher sequencing depth (4.1× per individual compared to 1.4× per individual), which increased the concordance correlation coefficient to 0.955. The Bayesian-based SNP caller produced somewhat higher concordance correlation coefficients, particularly at low sequencing depth. We recommend pooling at least 25 individuals combined with sequencing at a depth of 100× to produce satisfactory frequency estimates for common SNPs (minor allele frequency above 0.05).

Introduction

The method of pooling biological samples for downstream analysis has been used for more than seventy years [1]. The main advantage of pooling is that more samples can be analyzed in a cost-effective way. Pooling has been widely used in population genetics analysis for the estimation of single-nucleotide polymorphism (SNP) frequencies (reviewed in Sham et al. [2]). More recently, the field of population genetics has been revolutionized by the development of next-generation sequencing (NGS), as it is now possible to study genetic variation at the whole-genome level [3–7]. Whole-genome sequencing of pooled DNA is more recent and

known as Pool-seq [8]. While this method has become popular in the last few years, it has also been questioned, particularly in regard to the accuracy of SNP frequency data it produces [9,10]. To address this criticism, we investigated the robustness of Pool-seq in estimating SNP frequencies depending on sample size, sequencing depth and the SNP caller used.

So far, Pool-seq has been used in the study of bacteria [11], yeast [12], flatworm [13], sea urchins [14], plants [15,16], *Drosophila* [17–19], fish [20], birds [21] and mammals [22–25]. The approach has been applied to identify genomic loci affecting a trait of interest [19], to infer the demographic history of populations [20], to detect the signature of selection [17,18,25] and to perform genome-wide association studies (GWAS) [15,16]. In many cases, the pooling of samples is used to reduce costs. But pooling can be obligatory in other cases, such as when separating individuals is problematic [14,26] or when there is insufficient DNA to make individual libraries.

Several weaknesses of the method have been discussed. Low individual numbers, rough DNA quantification, and low sequencing depth can add error to polymorphism frequency estimates [27,28]. While these problems can be resolved and/or the magnitude of impact estimated, there are two more systemic, less easily resolvable limitations. When DNA of individual samples is pooled, information on haplotypes is lost. It is no longer possible to link a polymorphism with the individual to which it belongs [8], which is a problem for studies that require information on linkage disequilibrium, for example. The other limitation is that sequencing errors cannot easily be distinguished from true rare alleles [9]. Several authors have developed statistical approaches to tackle these two issues [29–33], which have been implemented in software programs to analyse pooled data [34–37]. In line with the intention of such improvements, the goal must be to assess the impact of problems of Pool-seq and to come up with procedures to resolve them, especially as whole-genome re-sequencing of individuals for population genomics is still expensive for species with medium-sized to large genomes.

This study focused on validating Pool-seq for population genomics by comparing SNP frequencies revealed by pooling and re-sequencing with those revealed by individual-based Genotyping By Sequencing (GBS) [38]. Comparisons were based on field-sampled plants of *Arabidopsis lyrata*. Library preparation required very little DNA and was performed with standard laboratory equipment. The three main questions we addressed were: (1) What is the increase in accuracy of Pool-seq SNP frequency estimates when increasing pool size? (2) What is the sequencing depth per individual required to obtain reliable population SNP frequencies with Pool-seq? And, (3) what is the difference in accuracy of SNP calling between a heuristic approach as implemented in the software VarScan [39] and a Bayesian approach as implemented in Snape [35]?

Materials and Methods

The *A*. *lyrata* plants of population A were collected in Presque Isle State Park (Erie, PA, USA) with a permit granted by the Commonwealth of Pennsylvania. The *A*. *lyrata* plants of population B were collected in the Clark Reservation State Park (Jamesville, NY, USA) with a permit granted by the New York State Office of Parks, Recreation and Historic Preservation. DNA of field-collected plants was extracted from silica-dried leaves with the DNeasy 96 Plant Kit (Qiagen, Hombrechtikon, Switzerland). Each DNA sample was quantified twice with the DNA quantification kit Quant-IT^{TM} DNA HS (Invitrogen, Paisley, UK), a method based on fluorimetry with a DNA-specific dye. Samples were only accepted as suitable for the study if the average concentration was at least 0.25 ng/ml and the coefficient of variation between the two rounds of quantification was smaller than 0.1. We sampled 14 individuals from population A and 25 from population B (Fig 1). The same individuals of these two populations were analysed by pooled (Pool-seq) and individual (GBS) sequencing.

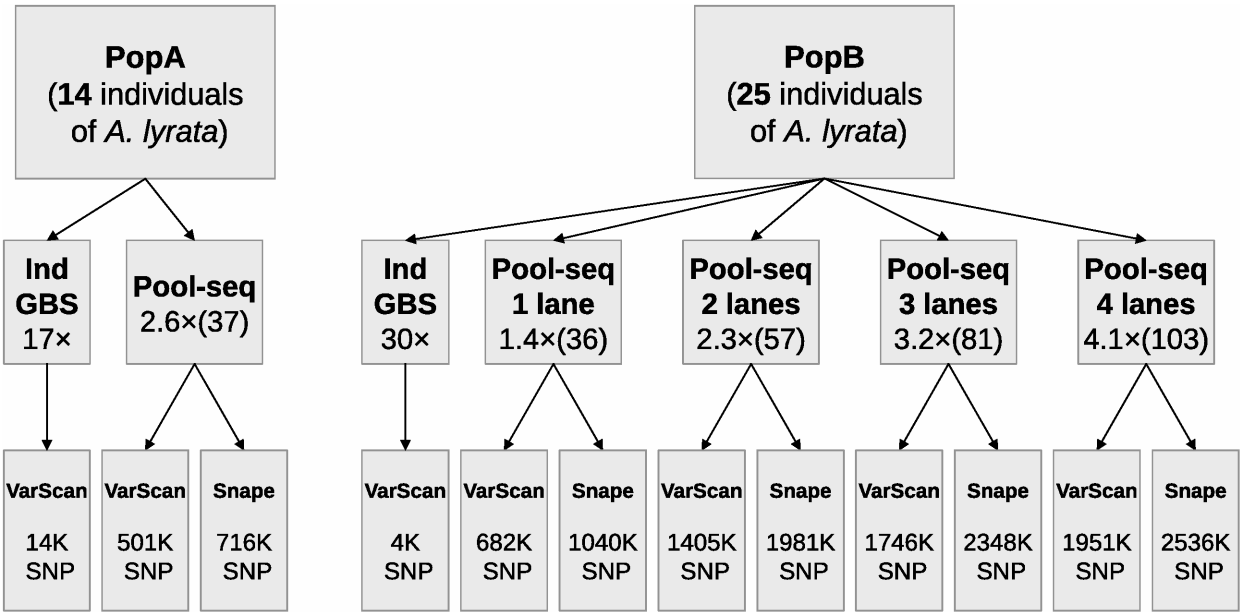

Fig 1. Diagram presenting the data sets produced to validate pooled whole-genome re-sequencing (Pool-seq) by individual-based Genotyping By Sequencing (GBS). The three rows of boxes contain the following information: top row: name of *Arabidopsis lyrata* population and number of individuals per population; second row: sequencing method, number of lanes merged (Pool-seq, population B only), the sequencing depth per individual and per pool (in parentheses); third row: the number of SNPs called by VarScan and Snape for each data set. Note that for GBS data, only the SNP caller VarScan was used.

doi:10.1371/journal.pone.0140462.g001

Library preparation: Pool-seq

Libraries for Pool-seq were prepared with the Nextera Kit (Illumina, San Diego, CA, USA) from equimolar-pooled DNA samples for each population. For each library a total of 40 ng of DNA was used, 2.8 ng per individual for population A and 1.6 ng per individual for population B. The protocol was customized to work with strips of 8 PCR tubes. The tagmentation time was increased from the manufacturer's protocol of 5 min to 10 min. The number of PCR cycles was increased to 8 (instead of 5) and the elongation time was decreased to 2 min (instead of 3 min). Library A was paired-end sequenced for 100 bases (PE100) on half a lane of Illumina HiSeq2000. Library B was PE100 sequenced on four lanes, each time constituting one quarter of the lane. Data of the lanes of population B were merged to create combinations from one to four lanes together (lane 1, lanes 1+2, lanes 1+2+3, lanes 1+2+3+4; Fig 1).

Library preparation: GBS

Genomic DNA (50 ng per individual) was digested at 37°C for 65 min in a 20 μL reaction volume with 5 U *Msp*I (NEB, Ipswich, MA, USA) in 10× NEBuffer 4. Following heat inactivation of the restriction enzyme (65°C, 20 min), tubes were allowed to cool slowly to room temperature covered with tinfoil. Adapter ligation was then performed immediately, using the following reaction mixture: 5 μL 10× NEBuffer 2, 1.93 μL P1 adapter (10 μM; sequence as per Elshire et al. [39] but with a CG instead of a CWG sticky end, and containing a 4–9 base barcode sequence), 1.93 μL P2 adapter (10 μM; sequence as per Elshire et al. [38] but with a CG instead of CWG sticky end), 1.8 μL rATP (100 mM), 1.5 μL T4 DNA ligase (2×10^6 U/mL), made up to 50 μL with ddH_2O. Ligation reactions were incubated at room temperature for 45 min, then heat-inactivated at 65°C for 20 min. Tubes were allowed to cool slowly as before.

To multiplex barcoded samples, 5 μL of each ligation mix was pooled. The mixture was cleaned with a Clean and Concentrator -5 Kit (Zymo Research, Irvine, CA, USA), eluted in 50 μL Buffer

EB. The pooled and cleaned DNA was used as template in 25 parallel PCR amplifications (replicated to minimise template bias). Each well included 2 μL template DNA, 2.5 μL of each PCR primer (as per Elshire et al. [38]), 5 μL dNTPs (2 mM), 0.5 μL Taq polymerase (Promega, Madison, WI, USA), 5x GoTaq buffer (Promega) and ddH_2O to a final volume of 50 μL. Cycling protocol was as follows: 72°C for 5 min, 96°C for 30 s, 18 cycles of [96°C for 30 s, 65°C for 30 s, 72°C for 30 s], and a final extension of 72°C for 5 min. All replicate PCR reactions were pooled, and cleaned a second time as before, eluting in 30 μL of buffer per ~200 μL of PCR product. Size selection was performed with the Caliper LapChip XT (PerkinElmer, Waltham, MA, USA), set to collect two peaks (first peak: 350 bp, second peak: 455 bp), which effectively collected fragments between 301–519 bp due to the machine's size accuracy limit of 14%. A third cleanup was performed, eluting in 17 μL Buffer EB. Sequencing was performed in a single Illumina HiSeq2000 lane.

Bioinformatics pipelines and SNP frequency comparison

The bioinformatics pipelines for Pool-seq and GBS sequence data were kept as similar as possible to minimize differences due to software used (pipelines accessible at: http://github.com/fraca). The sequences are stored at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) with the accession number PRJEB8335. Demultiplexing of the GBS data was performed with the preprocess_radtags script of Stacks [40], which retains reads with the proper barcode and restriction cut sites.

The Pool-seq and GBS sequences were trimmed using the script trim-fastq.pl of the software program PoPoolation [34] with a base quality threshold of 20, trimmed only from the 3' end to allow the subsequent removal of duplicates. Reads were mapped with BWA-MEM using default parameters [41]. The first 8 scaffolds of the published genome of *A. lyrata* v1.0 [42] were used as the reference genome. Data of the Pool-seq lanes of population B were merged to create the different combinations. Duplicate reads were removed with the MarkDuplicates tool of Picard [43]. Only proper paired reads with a mapping quality score above 20 were retained to create a pileup file with SAMtools [44]. The pileup file of Pool-seq data was filtered to retain regions with depth of coverage per site of 14–500 for population A and 25–500 for population B. The pileup file of GBS data was filtered for regions with depth of coverage per site of 5–500 for an individual and for data available for at least 90% of the individuals of a population. The regions near insertions and deletions were identified (identify-genomic-indel-regions.pl) and removed (filter-pileup-by-gtf.pl) with PoPoolation [34]. The genomic interspersed repeats were identified in the reference genome with RepeatMasker [45] using the default settings for "arabidopsis" and removed from the pileup files.

Finally, the filtered pileup files were used to call SNPs with the program VarScan with a significance (P) value ≤ 0.05, minimum base quality of 20 and a minimum allele count of two reads. For the Pool-seq data, SNPs were additionally called with Snape [35]. We retained SNPs with a posterior probability of segregation > 0.9 and minimum allele count of two reads. The nucleotide diversity and the genetic differentiation from the reference genome that are needed to set prior probabilities in the Bayesian model of Snape were calculated by NPStat [37]. We used the BEDTools software [46] to calculate sequencing depth or depth of coverage per site, defined as the number of times each base was sequenced per individual or per population pool. We applied the same thresholds for SNP calling and genome coverage calculation. Fig 1 presents the final 12 data sets used for further analysis. Allele frequency estimates were calculated as the fraction of reads carrying the non-reference allele for Pool-seq data, and the fraction of the non-reference allele across GBS-derived genotypes.

Three statistics were used to compare Pool-seq-based SNP frequencies with those obtained by GBS. First, the concordance correlation coefficient (CCC) was calculated using the epiR

package [47]. This test statistic can be used to evaluate the agreement between two variables [48]. The CCC combines precision (deviation from best-fit-line) and accuracy (deviation of best-fit-line from 45° line through origin) to determine how far the observed data deviate from the line of perfect concordance. Second, the absolute value of the difference between the estimated SNP frequencies with the two methods ($|\Delta f|$) was calculated and its distribution investigated. Third, a false negative rate was calculated as the fraction of SNPs called in GBS but not in the pooled sample, relative to the total number of SNPs called by GBS. This calculation included only genomic regions covered by both GBS and Pool-seq data, and considered SNP frequencies estimated from GBS to represent the true population frequencies. Because sequencing depth of GBS reads did not meet the minimum threshold of five reads for all the individuals, data did not allow the reliable estimation of a false positive rate of SNP calling.

Results

Sequencing statistics

Pooled sequencing of population A yielded 34 million paired-end reads. Prior to restricting the reads falling within an informative range of coverage depth, 50% of reads mapped unambiguously to 74% of the *A*. *lyrata* nuclear genome, at a mean depth of 27×. After applying the read depth cutoff (min 14×, max 500×) and removing duplicates, 46% of the reads mapped to 41% of the *A*. *lyrata* nuclear genome. The mean sequencing depth of population A was 37× in the final data set, which is equivalent to a mean depth of 2.6× per individual. Pooled sequencing of population B was performed on four lanes, each of which yielded ~40 million paired-end reads. We unambiguously mapped 59% of the total reads to cover 80% of the *A*. *lyrata* nuclear genome, at a mean depth of 25× per lane. After applying the read depth cutoff (min 25×, max 500×) and removing duplicates, the percentage of the genome covered by one lane was on average 36%, while the four lanes together covered 70%. The mean sequencing depth (post-cutoff) of population B depended on the number of lanes merged; depth was on average 36× for one lane and 103× for four lanes (Fig 1). Accordingly, sequencing depth per individual varied between 1.4× and 4.1×. Individual sequencing by GBS yielded 105 million paired-end reads (population A and B together) that were correctly barcoded and trimmed. We unambiguously mapped 40% of reads to cover 2% of the *A*. *lyrata* nuclear genome. Once the read depth cutoff (min 5×, max 500×) was applied, the mean sequencing depth per individual for population A in the final data set was 17× (range across individuals: 10×-30×). For population B the mean sequencing depth per individual was 30× (range across individuals: 11×-113×).

Number of SNPs

Table 1 shows the number of SNPs called by GBS and Pool-seq. For the Pool-seq protocol and population A, the software VarScan called 0.50 million SNPs, while Snape called 0.72 million SNPs. Increasing the depth from 1.4× to 4.1× (from one to four lanes) for population B increased the number of SNPs called. Using VarScan, the SNPs called increased from 0.68 million to 1.95 million. Using Snape, the SNPs called increased from 1.04 million to 2.54 million. Fig 2 shows the fraction of SNPs called with both VarScan and Snape in population B using one or four lanes. Almost all the SNPs called by VarScan were also called by Snape. The percentage of SNPs called by both programs relative to the total number of SNPs called by either Snape or VarScan, increased from 65% to 76% when the input data were increased from one lane to four lanes.

GBS led to more SNPs for population A than for population B. The smaller sample of individuals in population A (14 instead of 25 in population B) made it easier to attain the processing threshold of five or more reads for at least 90% of individuals. Therefore, population A had

Table 1. Comparison of SNP numbers and frequency estimate accuracy revealed by Pool-seq and by GBS. Columns report: library/lane identity (population A or B, estimation of sequencing depth per individual in Pool-seq, and software used to detect SNPs of Pool-seq data set), number of SNPs detected by GBS (SNP_{GBS}) and Pool-seq ($SNP_{Pool\text{-}seq}$), overlapping number of SNPs detected (SNP_{both}), concordance correlation coefficient (CCC) with lower and upper 95% confidence limit (LCL; UCL) of CCC, the mean of the absolute difference in SNP frequency estimates of the two methods (|Δf|), false negative rate (FN rate), that is, the fraction of SNPs called by GBS but not by Pool-seq, and their mean minor allele frequency (FN MAF).

Library/lane ID	$SNP_{Pool\text{-}seq}$	SNP_{GBS}	SNP_{both}	CCC	LCL	UCL	|Δf|	FN rate	FN MAF
A 2.6× VarScan	500'515	13'843	5731	0.827	0.819	0.835	0.109	0.270	0.115
A 2.6× Snape	716'483	13'843	7102	0.864	0.858	0.870	0.103	0.137	0.075
B 1.4× VarScan	682'317	4177	1333	0.887	0.876	0.898	0.092	0.385	0.077
B 1.4× Snape	1'039'746	4177	1754	0.911	0.902	0.918	0.083	0.212	0.054
B 2.3× VarScan	1'405'122	4177	2166	0.931	0.926	0.937	0.073	0.287	0.059
B 2.3× Snape	1'981'376	4177	2636	0.941	0.937	0.946	0.067	0.146	0.043
B 3.2× VarScan	1'745'682	4177	2413	0.946	0.942	0.950	0.063	0.211	0.049
B 3.2× Snape	2'348'269	4177	2738	0.951	0.948	0.955	0.059	0.116	0.038
B 4.1× VarScan	1'950'679	4177	2536	0.952	0.948	0.955	0.058	0.170	0.045
B 4.1× Snape	2'536'178	4177	2771	0.955	0.952	0.958	0.055	0.101	0.036

doi:10.1371/journal.pone.0140462.t001

higher overlap among individuals in genomic regions with sufficient sequencing depth and a higher total number of called SNPs. Moreover, the number of SNPs identified by both GBS and Pool-Seq was low (column SNP_{both}, Table 1) because GBS revealed SNP information for a small fraction of the genome and that fraction overlapped incompletely with genomic regions also covered with acceptable depth by Pool-seq.

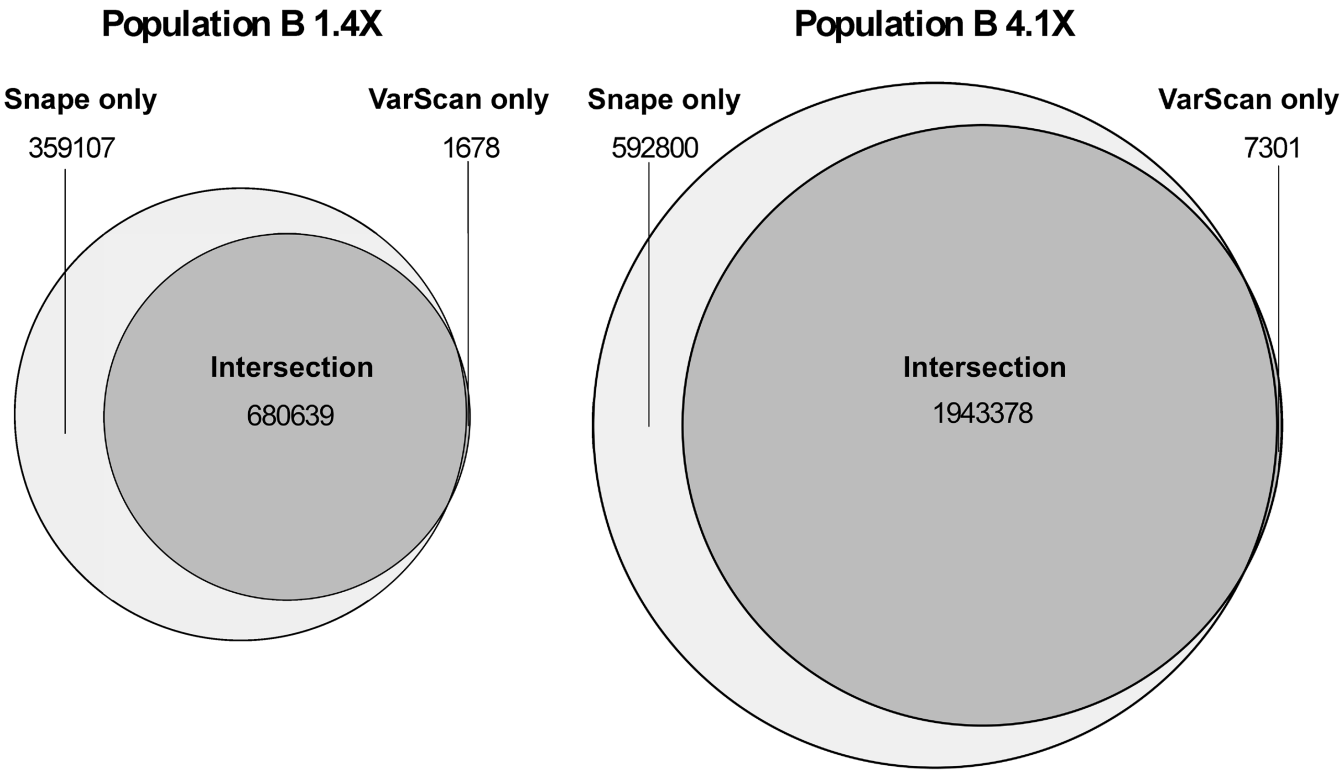

Fig 2. Venn diagram of Pool-seq SNPs called with VarScan (dark grey) and Snape (light grey). The left-hand panel shows the SNPs called for population B using data from lane 1 only. The right-hand panel shows the SNPs called for population B with the data from all four lanes. The figure was produced with the R package VennDiagram [49].

doi:10.1371/journal.pone.0140462.g002

Comparison of SNP frequencies revealed by Pool-seq versus GBS

First, SNP frequencies obtained with Pool-seq and GBS were compared by the use of the concordance correlation coefficient (CCC), which captures the agreement between two variables by accounting for precision and accuracy and which can range from 0 to 1. Fig 3 illustrates CCC values with upper and lower 95% confidence ranges for all library/lane combinations studied. CCC values for population A were 0.827 for SNPs called with VarScan and 0.864 for those called with Snape (Table 1). For population B, CCC values increased with increasing depth of coverage per site from 0.887 (1.4×) to 0.952 (4.1×) with VarScan and from 0.911 (1.4×) to 0.955 (4.1×) with Snape. S1 Fig illustrates the correlation between SNP frequency estimates of Pool-seq and those of GBS. The correlation between the two increased when more samples were pooled, and when the depth of coverage per site was increased.

Second, SNP frequencies revealed with Pool-seq and GBS were compared based on the absolute difference between the SNP frequency estimates of the two methods ($|\Delta f|$ in Table 1). The mean $|\Delta f|$ for population A was 0.109 with VarScan and 0.103 with Snape. The mean $|\Delta f|$ for population B decreased with increasing sequencing depth, from 0.092 to 0.058 with VarScan and from 0.083 to 0.055 with Snape. Fig 4 shows the distribution of $|\Delta f|$ for each library/lane combination, and S2 Fig presents the distribution of the difference between the SNP frequency estimates of the two methods across the achieved read depth at SNP sites for each library/lane combination. The difference in SNP frequencies between methods was generally lower when read depth was high, both across and within library/lane combinations. Furthermore, the distribution of the difference was not appreciably biased towards either negative or positive values (S2 Fig).

Third, the false negative rate (FN rate in Table 1) decreased with increasing sequencing depth, from 0.385 (1.4×) to 0.170 (4.1×) with VarScan and from 0.212 (1.4×) to 0.101 (4.1×) with Snape. At the same time, the mean frequency of minor alleles at GBS SNPs that were missed by Pool-seq (FN MAF in Table 1) decreased from 0.077 to 0.045 with VarScan and from 0.054 to 0.036 with Snape. Fig 5 illustrates that the minor allele frequency at SNP sites missed by Pool-seq was mostly lower than 5% when the number of sequenced individuals and the sequencing depth per individual were both high.

Discussion

Pooled whole-genome re-sequencing (Pool-seq) has only recently been adopted for population genomics in eukaryotes, so validation studies are needed, together with test of aspects of the wet-lab protocol and effects of the bioinformatics pipeline on results. Several studies have addressed the validation of this method (see Table 1 in [28]) but very few have examined the kind of large data sets now common in population genomics, containing more than a few thousand SNPs [22,27,32,51]. Here we analysed two populations of *Arabidopsis lyrata* by sequencing pools of individuals, and sequencing the same individuals separately by GBS. The main objective was to compare SNP frequencies obtained by Pool-seq with GBS-based SNP frequencies. Overall, we found that concordance correlation coefficients between SNP frequencies based on the two methods were high, between 0.827 and 0.955. These values are well within the range of other validation studies of pooled sequencing (e.g. Table 1 in [28]). Concordance increased with the pool size, with mean individual sequencing depth in the pool, and with the use of Snape as compared to VarScan as SNP calling software for the pooled samples.

The comparison of different numbers of individuals pooled was based on comparing 14 individuals with sequencing depth per individual of 2.6× and 25 individuals sequenced on two lanes with sequencing depth per individual of 2.3×. With the frequentist SNP caller VarScan, the concordance correlation coefficient increased from 0.827 to 0.931, while the mean absolute

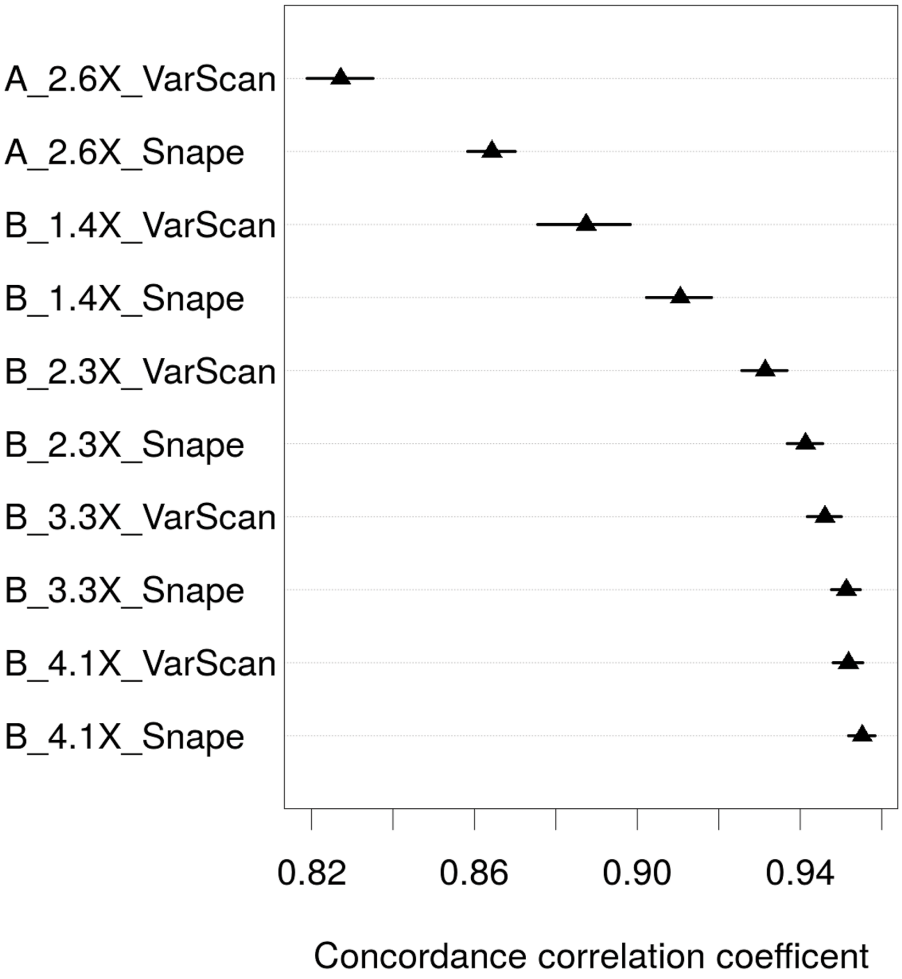

Fig 3. Concordance correlation coefficient between SNP frequencies estimated with Pool-seq and GBS for each library/lane combination and SNP caller. Mean CCC values with upper and lower 95% confidence ranges are shown. The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq (either VarScan or Snape; for GBS, only VarScan was used).

doi:10.1371/journal.pone.0140462.g003

difference between SNP frequency estimates from the two methods decreased from 0.109 to 0.073 (Table 1, Figs 3 and 4). With the Bayesian-based SNP caller Snape, the concordance correlation coefficient increased from 0.864 to 0.941, while the mean absolute difference between SNP frequency estimates from the two methods decreased from 0.103 to 0.067. These results clearly show that an increase in the number of individuals that are pooled—at least for the range we worked with—improves the accuracy of SNP frequency estimation, as predicted by several theoretical studies [8,10,33]. Similar to our results, those of another study on pooling different numbers of isofemale lines of *Drosophil*a revealed increases in concordance correlation coefficients from 0.822–0.867 with 22 lines to 0.906–0.934 and 0.911–0.936 with 42 lines [27]. Aside from this, we found that increasing the number of pooled individuals did not greatly increase the chance of detecting SNPs, at least not with sequencing depth per individual used here. The false negative rate remained almost unchanged, increasing slightly from 0.270 to 0.287 with VarScan, and from 0.137 to 0.146 for Snape.

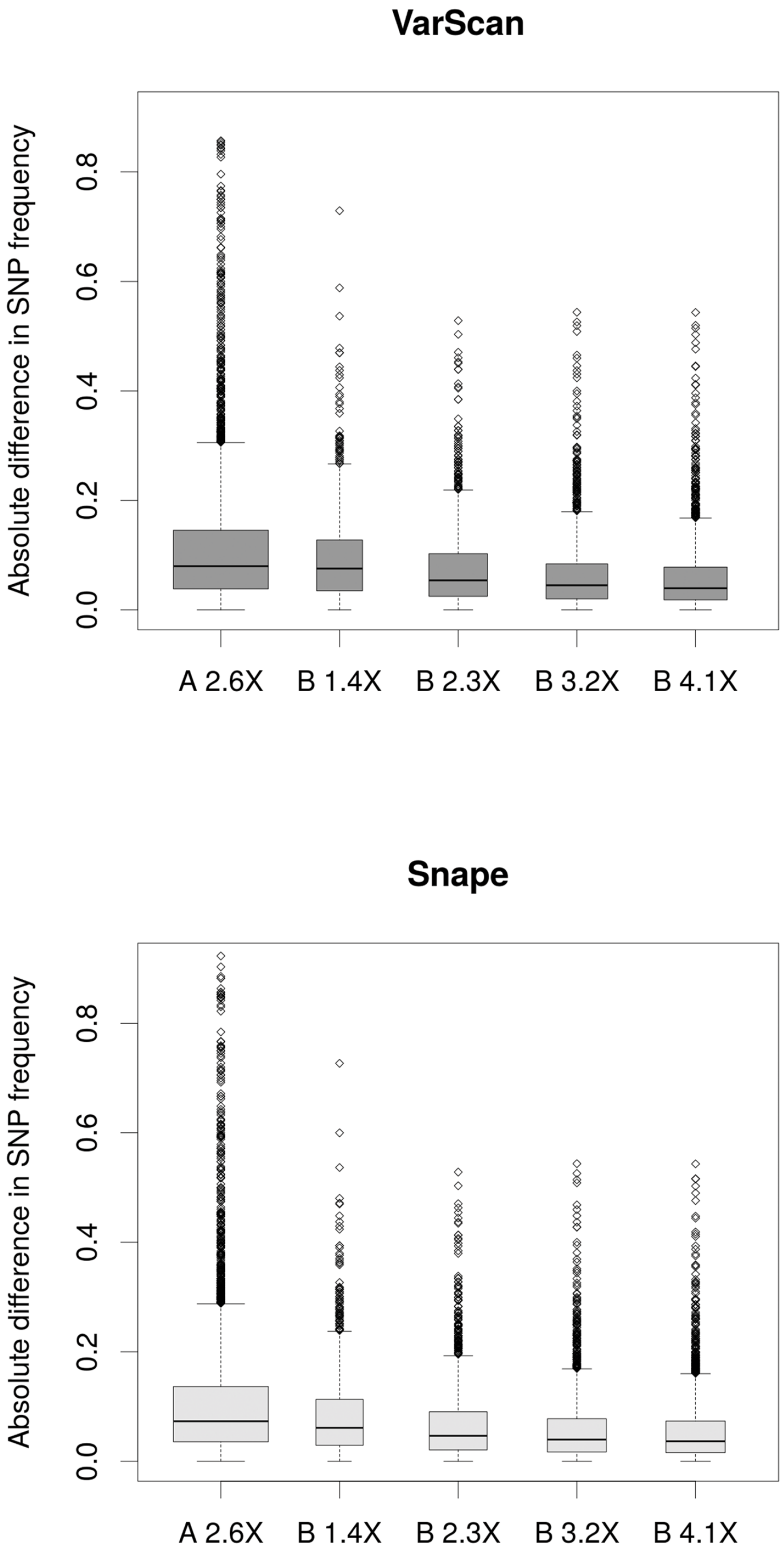

Fig 4. Box plot illustrating the distribution of the absolute difference in SNP frequency estimates between Pool-seq and GBS. The upper panel (dark grey) shows distributions when SNPs were called with VarScan for Pool-seq, the lower panel (light grey) shows distributions with Snape. Library names contain information on: the population (A or B), and the sequencing depth by Pool-seq. The band inside each box shows the median, while the lower and upper ends indicate the first and third quartile, respectively. The

lower whisker is -1.5x the interquartile range from the first quartile, while the upper whisker is +1.5x the interquartile range from the third quartile. The diamonds represent outliers.

doi:10.1371/journal.pone.0140462.g004

The comparison of varying depth of coverage per site revealed further potential for improving SNP frequency estimates. An increase of the depth of sequencing per individual from 1.4× to 2.3×, 3.2×, and 4.1×, led to an increase in concordance of Pool-seq with GBS (Fig 3) and a decrease in the absolute difference between SNP frequency estimates between methods (Fig 4) and false negative rate (Table 1). In line with our results, a sequencing study on a pool of 30 individuals of the pine processionary moth [32] revealed improved frequency estimates when the sequencing depth was increased from a range of 10×-50× to >200×, equivalent to a depth per individual of 0.3×-1.7× to >6.7×. The authors observed an increase in the correlation coefficient from 0.93 to >0.99 (across different sequencing depths per individual for individual sequencing) and a decrease of the median of the absolute difference between individual-based and pooled-based frequency estimates from 0.067 to 0.007.

A major issue with the Pool-seq technique is a lack of power to detect rare alleles [9,27,33], which is unimportant for some applications but important for others. For example, rare alleles may be important for explaining phenotypic variation within populations [52] and therefore desirable to detect in genome-wide association studies. We investigated this issue by analyzing the minor allele frequency of false negative SNPs (SNPs that were called only in GBS but not in the Pool-seq samples). In all library/lane combinations, the majority of false negative SNPs had low minor allele frequencies (Fig 5). At the sequencing depth of 4.1× per individual in the pool with 25 individuals the majority of GBS SNPs not detected by Pool-seq had a frequency below 0.05 (mean = 0.045 for VarScan and mean = 0.036 for Snape; Table 1). For higher GBS-based SNP frequencies, the number of SNPs missed by Pool-seq rapidly decreased. This result supports the utility of our upper pool size and maximum depth of sequencing. It has been suggested that to detect a minor allele with near-certainty, its frequency must be larger than 10 divided by the number of pooled diploid individuals [33], which in our study would have been 0.4 for the larger population. We appeared able to detect all minor alleles with frequency > 0.15 at the largest pool size and sequencing depth tested (Fig 5). The discrepancy is likely due to the difference in variant calling approaches and the fact that we used a $P = 0.05$ threshold for detection as opposed to the $P = 0.001$ level used by Lynch et al. [33]. For some population genetics studies this detection threshold is likely to be acceptable and our results confirm that this kind of pooled data is useful for detecting common minor alleles. Of course, those considering Pool-seq should be aware of the limitation of this approach in detecting rare alleles.

Several SNP callers can be applied to pooled data (reviewed in [8]). We used VarScan [38], which uses a frequentist approach, and Snape [35], which uses a Bayesian approach. Both take into account sequencing depth, base quality, and statistical significance, while Snape includes information on nucleotide diversity and divergence from the reference genome to detect SNPs. Our results show that Snape called considerably more SNPs than VarScan (Fig 2). The number of SNPs called by Snape that were confirmed by GBS was on average 20% higher than the number of SNPs called by VarScan confirmed by GBS (column SNP_{both} in Table 1). Furthermore the false negative rate was found to be systematically lower with Snape. Therefore, it can be argued that Snape is more powerful at detecting SNPs than is VarScan. This may however be accompanied by an increase in the false positive rate, which is an important avenue for further investigation. Also, the concordance correlation coefficients between GBS and Pool-seq SNP frequencies were slightly higher with Snape than with VarScan, although this difference between SNP callers declined with increasing sequencing depth (Table 1, Fig 3). The absolute

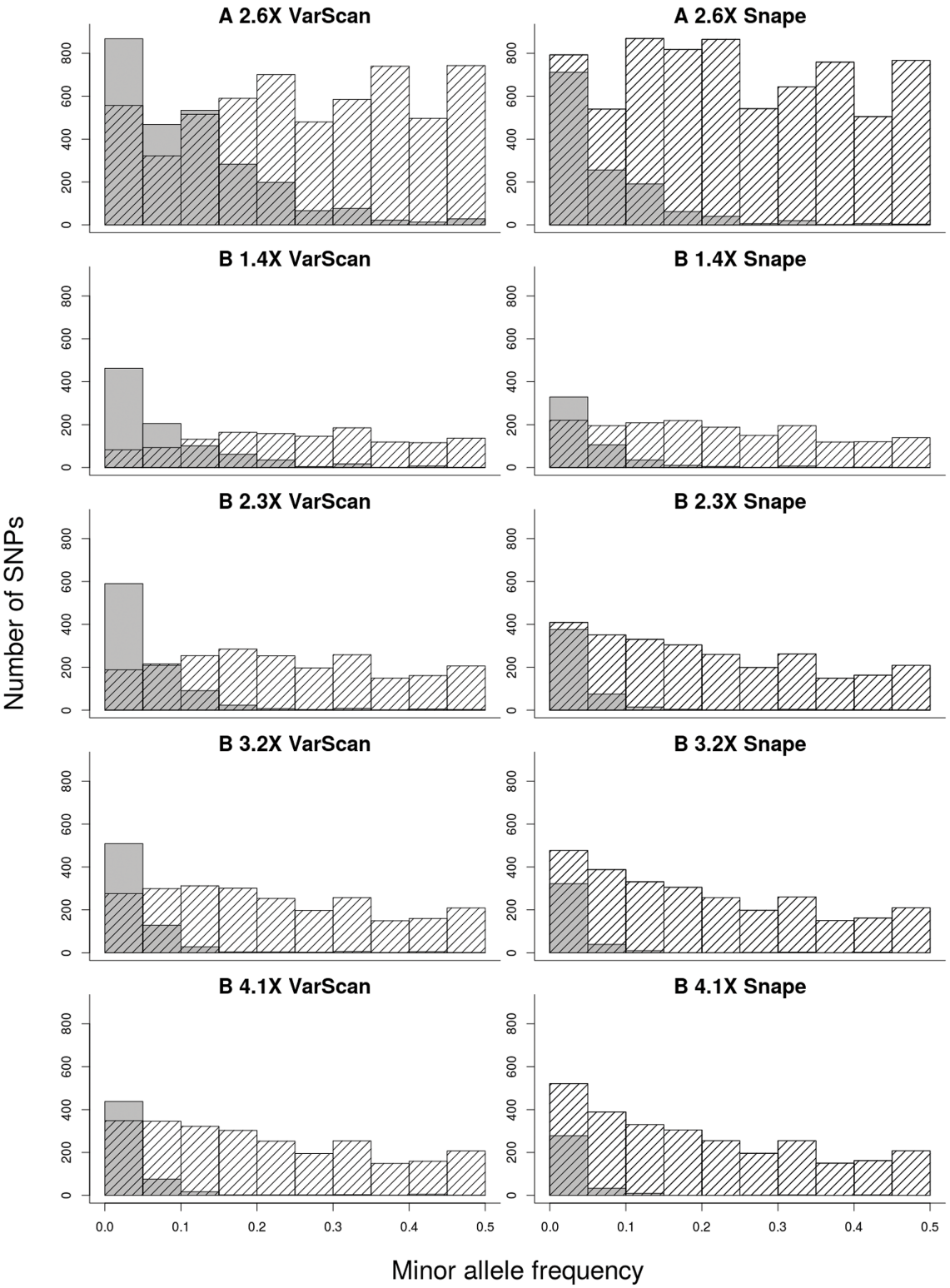

Fig 5. Histogram of minor allele frequency of GBS. The grey bars represent the SNPs present only in GBS. The striped bars represent the SNPs sequenced in the GBS and Pool-seq samples. The 10 panels show the results for the various Pool-seq library/lane combinations and the two SNP callers. The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq (either VarScan or Snape; for GBS, only VarScan was used).

doi:10.1371/journal.pone.0140462.g005

difference in SNP frequencies between methods was lower with Snape than with VarScan. These results indicate that the use of priors for nucleotide diversity and divergence contribute positively to the calling of SNPs.

In conclusion, we have presented a method that uses low input DNA (1.6 ng per individual) and widely-available commercial kits to perform pooled whole-genome re-sequencing. Thanks to the tagmentation step, we avoided fragmentation by sonication, which requires more input DNA. We validated SNP frequencies by comparison with GBS data. Our study strengthens the conclusion that the quality of pooled sequencing data sets relies on two critical parameters: the number of individuals that are pooled, and sequencing effort. In a recent review on Pool-seq [8], the authors recommend pools of at least 40 individuals with sequencing depth of more than 50× per pool. Lynch et al. [33] used a maximum likelihood estimator and suggested more than 100 individuals and a sequencing depth of 100× per pool to obtain high confidence in allele frequency estimates. Based on the empirical comparison we performed, we find that a pool of 25 individuals combined with a sequencing depth of 100× produces SNP frequency data with satisfactory precision and accuracy. We confirm that Pool-seq is a useful method to detect genomic variants with a frequency of about 0.05 and larger.

Supporting Information

S1 Fig. Scatter plots of SNP frequency estimates based on GBS and Pool-seq for the various library/lane combinations and the two SNP callers. The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq (either VarScan or Snape; for GBS, only VarScan was used). The solid line indicates the expectation of equal frequency with both sequencing approaches.
(EPS)

S2 Fig. Hexbin plots of the difference in SNP frequency estimates between Pool-seq and GBS with respect to the total read depth at SNP sites of Pool-seq. The name of a library/lane combination contains information on: the population (A or B), sequencing depth per individual by Pool-seq, and the software used to detect SNPs for Pool-seq (either VarScan or Snape; for GBS, only VarScan was used). Hexagons are shaded by SNP count according to the scale shown on the right. The figure was produced with the hexbin package in R [50].
(TIFF)

Acknowledgments

Christian Beisel and Daniel Berner gave advice on wet lab protocols, Luca Ferretti and Robert Kofler on bioinformatics pipelines and data analysis. We thank two anonymous reviewers for helpful comments that improved this article. Wet lab work and sequencing was done at: the Genetic Diversity Centre, ETH Zürich; the Functional Genomics Centre Zürich, ETH Zürich and University of Zürich; the Quantitative Genomics Facility Basel, ETH Zürich-Basel and University of Basel.

Author Contributions

Conceived and designed the experiments: MF PCG YW. Performed the experiments: MF PCG. Analyzed the data: MF. Contributed reagents/materials/analysis tools: MF YW. Wrote the paper: MF YW PCG.

References

1. Dorfman R. The detection of defective members of large populations. Ann Math Stat. 1943; 14: 436–440.
2. Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA pooling: a tool for large-scale association studies. Nat Rev Genet. 2002; 3: 862–871. doi: 10.1038/nrg930 PMID: 12415316
3. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437: 376–380. doi: 10.1038/nature03959 PMID: 16056220
4. Pandey V, Nutter RC, Prediger E. Applied Biosystems SOLiDTM System: ligation-based sequencing. In: Janitz M, editor. Next Generation Genome Sequencing: Towards Personalized Medicine. 2008. pp. 29–42. doi: 10.1002/9783527625130.ch3
5. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456: 53–59. doi: 10.1038/nature07517 PMID: 18987734
6. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009; 323: 133–138. doi: 10.1126/science.1162986 PMID: 19023044
7. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011; 475: 348–352. doi: 10.1038/nature10242 PMID: 21776081
8. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014; 15: 749–763. doi: 10.1038/nrg3803 PMID: 25246196
9. Cutler DJ, Jensen JD. To pool, or not to pool? Genetics. 2010; 186: 41–43. doi: 10.1534/genetics.110.121012 PMID: 20855575
10. Anderson EC, Skaug HJ, Barshis DJ. Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. Mol Ecol. 2014; 23: 502–512. doi: 10.1111/mec.12609 PMID: 24304095
11. Holt KE, Teo YY, Li H, Nair S, Dougan G, Wain J, et al. Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. Bioinformatics. 2009; 25: 2074–2075. doi: 10.1093/bioinformatics/btp344 PMID: 19497932
12. Burke MK, Liti G, Long AD. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *Saccharomyces cerevisiae*. Mol Biol Evol. 2014; 31: 3228–3239. doi: 10.1093/molbev/msu256 PMID: 25172959
13. Clément JAJ, Toulza E, Gautier M, Parrinello H, Roquis D, Boissier J, et al. Private selective sweeps identified from next-generation pool-sequencing reveal convergent pathways under selection in two inbred *Schistosoma mansoni* strains. PLoS Negl Trop Dis. 2013; 7: e2591. doi: 10.1371/journal.pntd.0002591 PMID: 24349597
14. Pespeni MH, Sanford E, Gaylord B, Hill TM, Hosfelt JD, Jaris HK, et al. Evolutionary change during experimental ocean acidification. Proc Natl Acad Sci. 2013; 110: 6937–6942. doi: 10.1073/pnas.1220673110 PMID: 23569232
15. Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin S V. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet. 2010; 42: 260–263. doi: 10.1038/ng.515 PMID: 20101244
16. Fischer MC, Rellstab C, Tedder A, Zoller S, Gugerli F, Shimizu KK, et al. Population genomic footprints of selection and associations with climate in natural populations of *Arabidopsis halleri* from the Alps. Mol Ecol. 2013; 22: 5594–5607. doi: 10.1111/mec.12521 PMID: 24102711
17. Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS Genet. 2012; 8: e1002487. doi: 10.1371/journal.pgen.1002487 PMID: 22291611
18. Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlötterer C, et al. Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Mol Ecol. 2012; 21: 4748–4769. doi: 10.1111/j.1365-294X.2012.05731.x PMID: 22913798
19. Bastide H, Betancourt AJ, Nolte V, Tobler R, Stöbe P, Futschik A, et al. A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. PLoS Genet. 2013; 9: e1003534. doi: 10.1371/journal.pgen.1003534 PMID: 23754958
20. Corander J, Majander KK, Cheng L, Merilä J. High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. Mol Ecol. 2013; 22: 2931–2940. doi: 10.1111/mec.12174 PMID: 23294045

21. Rubin C-J, Zody MC, Eriksson J, Meadows JRS, Sherwood E, Webster MT, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature. 2010; 464: 587–591. doi: 10.1038/nature08832 PMID: 20220755

22. Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods. 2008; 5: 247–252. doi: 10.1038/nmeth.1185 PMID: 18297082

23. Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, et al. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature. 2013; 495: 360–364. doi: 10.1038/nature11837 PMID: 23354050

24. Rubin C-J, Megens H-J, Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A. 2012; 109: 19529–19536. doi: 10.1073/pnas.1217149109 PMID: 23151514

25. Esteve-Codina A, Paudel Y, Ferretti L, Raineri E, Megens H-J, Silió L, et al. Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs. BMC Genomics. 2013; 14: 148. doi: 10.1186/1471-2164-14-148 PMID: 23497037

26. Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. Who is eating what: diet assessment using next generation sequencing. Mol Ecol. 2012; 21: 1931–1950. doi: 10.1111/j.1365-294X.2011.05403.x PMID: 22171763

27. Zhu Y, Bergland AO, González J, Petrov DA. Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*. PLoS One. 2012; 7: e41901. doi: 10.1371/journal.pone.0041901 PMID: 22848651

28. Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC. Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. PLoS One. 2013; 8: e80422. doi: 10.1371/journal.pone.0080422 PMID: 24244686

29. Futschik A, Schlötterer C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics. 2010; 186: 207–218. doi: 10.1534/genetics.110.114397 PMID: 20457880

30. Pérez-Enciso M, Ferretti L. Massive parallel sequencing in animal genetics: wherefroms and wheretos. Anim Genet. 2010; 41: 561–569. doi: 10.1111/j.1365-2052.2010.02057.x PMID: 20477787

31. Boitard S, Schlötterer C, Nolte V, Pandey RV, Futschik A. Detecting selective sweeps from pooled next-generation sequencing samples. Mol Biol Evol. 2012; 29: 2177–2186. doi: 10.1093/molbev/mss090 PMID: 22411855

32. Gautier M, Foucaud J, Gharbi K, Cézard T, Galan M, Loiseau A, et al. Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol Ecol. 2013; 22: 3766–3779. doi: 10.1111/mec.12360 PMID: 23730833

33. Lynch M, Bost D, Wilson S, Maruki T, Harrison S. Population-genetic inference from pooled-sequencing data. Genome Biol Evol. 2014; 6: 1210–1218. doi: 10.1093/gbe/evu085 PMID: 24787620

34. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, et al. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One. 2011; 6: e15925. doi: 10.1371/journal.pone.0015925 PMID: 21253599

35. Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M. SNP calling by sequencing pooled samples. BMC Bioinformatics. 2012; 13: 239. doi: 10.1186/1471-2105-13-239 PMID: 22992255

36. Boitard S, Kofler R, Françoise P, Robelin D, Schlötterer C, Futschik A. Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. Mol Ecol. 2013; 13: 337–340. doi: 10.1111/1755-0998.12063

37. Ferretti L, Ramos-Onsins SE, Pérez-Enciso M. Population genomics from pool sequencing. Mol Ecol. 2013; 22: 5561–5576. doi: 10.1111/mec.12522 PMID: 24102736

38. 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: e19379. doi: 10.1371/journal.pone.0019379 PMID: 21573248

39. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22: 568–576. doi: 10.1101/gr.129684.111 PMID: 22300766

40. Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. Stacks: an analysis tool set for population genomics. Mol Ecol. 2013; 22: 3124–3140. doi: 10.1111/mec.12354 PMID: 23701397

41. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;

42. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011; 43: 476–481. doi: 10.1038/ng.807 PMID: 21478890

43. Picard: a set of tools for working with next generation sequencing data in the BAM format. [Internet]. Available: http://broadinstitute.github.io/picard/

44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

45. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0 [Internet]. 2010. Available: http://www.repeatmasker.org

46. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. doi: 10.1093/bioinformatics/btq033 PMID: 20110278

47. Stevenson M, Nunes T, Heuer C, Marshall J, Sanchez J, Thornton R, et al. epiR: tools for the analysis of epidemiological data. R package version 0.9–62 [Internet]. 2015. Available: http://cran.r-project.org/package = epiR

48. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989; 45: 255–268. PMID: 2720055

49. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 2011; 12: 35. doi: 10.1186/1471-2105-12-35 PMID: 21269502

50. Carr D, Lewin-Koh N, Maechler M. Hexbin: hexagonal binning routines. R package version 1.27.0. In: 2014 [Internet]. Available: http://cran.r-project.org/web/packages/hexbin/index.html

51. Bansal V, Tewhey R, Leproust EM, Schork NJ. Efficient and cost effective population resequencing by pooling and in-solution hybridization. PLoS One. 2011; 6: e18353. doi: 10.1371/journal.pone.0018353 PMID: 21479135

52. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A. 2006; 103: 1810–1815. doi: 10.1073/pnas.0508483103 PMID: 16449388