Monitoring rapid evolution of plant populations at scale with Pool-Sequencing

Lucas Czech[1,%], Yunru Peng[1,%], Jeffrey P. Spence[2], Patricia L.M. Lang[3], Tatiana Bellagio[1,3], Julia Hildebrandt[4], Katrin Fritschi[4], Rebecca Schwab[4], Beth A. Rowan[4], GrENE-net consortium[$], Detlef Weigel[4], J.F. Scheepens [5], François Vasseur [3,6], Moises Exposito-Alonso[1,3,4,7*]

[1] Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94305, USA

[2] Department of Genetics, Stanford University, Stanford, CA 94305, USA

[3] Department of Biology, Stanford University, Stanford, CA 94305, USA

[4] Department of Molecular Biology, Max Planck Institute for Biology Tübingen, 72076 Tübingen, Germany

[5] Faculty of Biological Sciences, Goethe University, Frankfurt, Max-von-Laue-Str. 13, 60438 Frankfurt am Main, Germany

[6] Centre d'Écologie Fonctionnelle et Évolutive (CEFE), University of Montpellier, CNRS, EPHE, IRD, Univ Paul Valéry Montpellier 3, F-34090 Montpellier, France

[7] Department of Global Ecology, Carnegie Institution for Science, Stanford, CA 94305, USA

[%] Shared co-first authors

[*] To whom correspondence should be addressed: moisesexpositoalonso@gmail.com

[$] GrENE-net consortium authors and affiliations listed in the appendix

Keywords: Evolve & Resequence in plants, Pool-Sequencing, rapid revolution, GrENE-net.org.

Abstract

The change in allele frequencies within a population over time represents a fundamental process of evolution. By monitoring allele frequencies, we can analyze the effects of natural selection and genetic drift on populations. To efficiently track time-resolved genetic change, large experimental or wild populations can be sequenced as pools of individuals sampled over time using high-throughput genome sequencing (called the Evolve & Resequence approach, E&R). Here, we present a set of experiments using hundreds of natural genotypes of the model plant *Arabidopsis thaliana* to showcase the power of this approach to study rapid evolution at large scale. First, we validate that sequencing DNA directly extracted from pools of flowers from multiple plants -- organs that are relatively consistent in size and easy to sample -- produces comparable results to other, more expensive state-of-the-art approaches such as sampling and sequencing of individual leaves. Sequencing pools of flowers from 25-50 individuals at ~40X coverage recovers genome-wide frequencies in diverse populations with accuracy $r > 0.95$. Secondly, to enable analyses of evolutionary adaptation using E&R approaches of plants in highly replicated environments, we provide open source tools that streamline sequencing data curation and calculate various population genetic statistics two orders of magnitude faster than current software. To directly demonstrate the usefulness of our method, we conducted a two-year outdoor evolution experiment with *A. thaliana* to show signals of rapid evolution in multiple genomic regions. We demonstrate how these laboratory and computational Pool-seq-based methods can be scaled to study hundreds of populations across many climates.

Introduction

How fast a species can adapt to different environments from standing within-species genetic variation is a burning question in evolutionary ecology and genetics. A powerful approach to study environment-driven adaptation is provided by field experiments in which multiple genotypes of a species are grown together and traits and fitness are measured (Clausen et al., 1941; Kingsolver et al., 2001; Savolainen et al., 2013). Such experiments, typically conducted within a single generation, have allowed measuring the strength of natural selection over phenotypic traits or genetic variants, which is often strong (Exposito-Alonso et al., 2019; Kingsolver et al., 2001; Siepielski et al., 2017; Thurman and Barrett, 2016). Such studies often cannot measure the response to selection—evolutionary change—of a population, as this depends on the genetic trait architecture (Bergland et al., 2014; Walsh and Blows, 2009) and environmental fluctuation over time (Bergland et al., 2014), which has led to inconsistent long-term trait changes in populations (Merilä et al., 2001). Highly-replicated multi-year experiments where phenotypes and genomic variation are tracked would be ideal to study these evolutionary forces and robustly test the predictability of evolution (Grant and Grant, 2002; Nosil et al., 2018) .

An opportunity to conduct multi-generational experiments to study evolution over time is the so-called "Evolve & Resequence" (E&R) approach (Schlötterer et al., 2015; Turner et al., 2011). E&R experiments leverage cost-effective, high-throughput sequencing to study the frequency of genome-wide variants or genotypes of a population over time. Such frequency trajectories capture evolutionary forces such as drift and natural selection in action. This approach has been popular in bacterial and animal model systems such as *Escherichia coli* and *Drosophila melanogaster* (Bergland et al., 2014; Good et al., 2017; Schlötterer et al., 2014). In the traditional genome sequencing approach, each individual is processed independently into one DNA sequencing library. The most common E&R approach is Pool-Sequencing, where multiple individuals sampled from the same population are processed into a single DNA sequencing library (Futschik and Schlötterer, 2010). While individual haplotypes are lost in the Pool-Seq approach, population-level allele frequencies are obtained in a cost-effective manner (Schlötterer et al., 2014). The Pool-Seq approach has been typically applied on a single population over time to study rapid selective sweeps (Iranmehr et al., 2017) and quantitative trait evolution (Endler et al., 2016). Parallel E&R experiments across large environmental gradients could enable the study of population (mal)adaptation across present climates and inform future responses (Capblancq et al., 2020). Combining Pool-Seq experiments, which subject the same starting genetic variation to an environmental condition, with landscape genomic approaches that aim to detect climate-driven natural selection or sweeps in the presence of population confounders (Günther and Coop, 2013; Hancock et al., 2011; Pfenninger et al., 2021), could be a powerful approach to depict how climate impacts evolutionary genetic processes leading to adaptation and extinction.

To enable globally-distributed E&R experiments to study climate adaptation, two key innovations are necessary beyond lowering sequencing costs: making library preparation scalable to thousands of whole-genome samples, and standardizing computational genomics software that allow researchers to analyze thousands of population samples, akin in speed to single-genome data structures and libraries such as HTSlib (Bonfield et al., 2021). To achieve the first goal, we reduced the preparation time (to ~2 h/96 pooled samples) and cost (to ~$3/pooled sample) (Rowan et al.,

2015) for genomic DNA library preparation using Tn5 transposase (Baym et al., 2015; Rowan et al., 2015), and tested these libraries for Pool-Sequencing approaches. For the second goal, we developed a new C++ implementation for fast computing of population genetic statistics for Pool-Seq, g$_r$enedalf, (Czech and Exposito-Alonso, 2022), based on the original Perl-based PoPoolation software (Kofler et al., 2011a, 2011b). Our implementation now offers ~100-fold speed improvements, allowing analyses of thousands of pooled libraries in minutes rather than days (Czech and Exposito-Alonso, 2022). These methods can be applied to any organisms, and we demonstrate the utility and power of the approach with the diploid annual plant *Arabidopsis thaliana*. We showcase our methods' efficacy for studying rapid adaptation in the context of plant evolutionary ecology, a field that typically uses individual-based methods such as common garden experiments and within-generation fitness assays to understand natural selection in different environments (Anderson and Wadgymar, 2019; Brachi et al., 2010; Exposito-Alonso et al., 2019; Fournier-Level et al., 2011; Lovell et al., 2021; Lowry et al., 2009; Monnahan et al., 2020) .

In this article, we describe our E&R design, Pool-Seq protocols, and computational approaches for a set of four experiments using natural genotypes of *A. thaliana*. We provide evidence that our simple and affordable large-scale experimental setup can generate allele frequency data with quality comparable to established small-scale approaches. In particular, we sequenced a mixture of seeds pooled from several hundreds of *A. thaliana* genotypes from the 1001 Genomes Project (1001 Genomes Consortium, 2016), which can be used as a founder population for multiple evolution experiments (**Experiment 1**). We further constructed sequencing libraries of exactly two inbred genotypes using the Pool-Seq approach to assess the deviation in the allele frequencies from the expected 50% frequencies at positions where the genotypes differ (**Experiment 2**). We conducted varying poolings of genotypes and tissue types (i.e., leaf versus flower) to describe the effect of individual pooling and coverage in allele frequency inferences (**Experiment 3**). We ran a pilot "E&R common garden" experiment to test our methods in realistic outdoor settings and analyzed whether signals of rapid evolution could be detected in a few generations (**Experiment 4**).

A Snakemake-based pipeline to streamline and parallelize frequency calling in Pool-Sequencing

To tackle the large amount of sequencing data that is needed to comprehensively test for rapid evolution across environments with Pool-Sequencing, we implemented g$_r$enepipe (Czech and Exposito-Alonso, 2021), a pipeline based on the Snakemake workflow management system (Köster and Rahmann, 2012; Mölder et al., 2021), to process raw sequence data into variant calls and allele frequencies. We used g$_r$enepipe to process the data from all our four experiments described below. Unless otherwise specified, we used g$_r$enepipe v0.6.0, with the following tools in the pipeline: trimmomatic (Bolger et al., 2014) for read trimming, bwa mem (Li and Durbin, 2009) for mapping against the reference genome, and samtools (Li et al., 2009) for working with bam and pileup files. We furthermore employed several quality control tools that are built into g$_r$enepipe to ensure that our sequence data is of sufficient quality (Andrews and Others, 2017; Ewels et al., 2016; Li et al., 2009; Okonechnikov et al., 2016). Note that g$_r$enepipe furthermore offers variant calling, using tools such as BCFtools (Li, 2011), freebayes (Garrison and Marth, 2012), and the GATK HaplotypeCaller (McKenna et al., 2010). The exact tools and parameter settings used in each run of the pipeline are available at https://github.com/lczech/grenepilot-paper.

The g_{r}enepipe automatization of single variant polymorphism (SNP) and their frequency calling allows us to test a number of variant filters and compare them in a standardized fashion. Specifically, we focused on quality controls related to:

1. Base quality filters based on Illumina PHRED scores.
2. Mapping quality filters to reduce the likelihood of false positive variant calls. These follow essentially the same curated filters of the PoolSNP pipeline used in the "Drosophila Evolution over Space and Time" resource (Kapun et al., 2021, 2020) .
3. Free discovery of genetic variants *vs* utilizing only 11,769,920 biallelic SNPs (out of 12,883,854) previously discovered in individual strains from the 1001 Genomes project (1001 Genomes Consortium, 2016) or a high-quality subset of the same genome set of 1,353,386 biallelic SNPs (Exposito-Alonso et al., 2019).
4. Coverage filters and minimal alternative allele counts to reduce sampling noise and sequencing errors (Kapun et al., 2021; Lynch et al., 2014).

The experiments described below make use of these filters, unless otherwise specified.

A new efficient command line tool for population genetic statistics using Pool-Sequencing

To efficiently analyze Pool-Seq allele frequency data for thousands of population samples, we developed a C++ based command line tool called g_{r}enedalf (Czech and Exposito-Alonso, 2022), which is able to parse .bam/.sam/.vcf/.pileup/.sync files, analyze allele counts and frequencies on the fly, and compute population genetic statistics implemented in the broadly used PoPoolation1 and PoPoolation2 (Kofler et al., 2011a, 2011b) along with new extensions of several unbiased statistics derived here and elsewhere (Hivert et al., 2018). The **Supplemental Mathematical Appendix** includes mathematical derivations and motivation of various unbiased corrections of Watterson's θ_W, π, Tajima's D, and F_{ST} that account for two main sources of noise in Pool-Seq: the finite number of individuals pooled (n), and the finite coverage per base pair along the genome (C) (see cartoon **Fig. S1**). These are two nested Binomial samplings, where, for a polymorphic site in a population, we first have a chance of sampling k individuals carrying each occurring allele out of all n individuals pooled, which is proportional to the true allele frequency f_A in the population. Then, after DNA sequencing, we have a chance of observing c reads in a pool of C (coverage) reads, which is proportional to the frequency of each allele in the pooled sample of individuals (k/n).

The first parameter that we are interested is the genetic diversity, nucleotide diversity, or observed heterozygosity, for a given SNP in the genome, expressed as:

$$\pi(c, C) = \frac{C}{C-1}\left(1 - \sum_{\tau} \frac{c_\tau^2}{C^2}\right), \qquad \text{(eq. 1)}$$

where c is the number of reads presenting each of the four possible nucleotide bases ($\tau \in \text{ACTG}$), and the ratio represents the raw sample allele frequency $f_\tau = c_\tau / C$, represents the raw sample allele frequency. Our software corrects such diversity parameters using Bessel's correction for finite coverage. An additional correction of individual sample size $\text{n}/(\text{n-}1)$ may also be applied, although can be done for all genome-wide values computed by g_{r}enedalf in downstream processes. Such a metric of diversity could be used to detect islands of low diversity appearing over time in E&R, which could be indicative of a selective sweep.

The second parameter of most interest is allele frequency differentiation between two spatial or temporal population samples, F_{ST} , for which there are multiple definitions (**Supplemental Mathematical Appendix**). Following the same notation as the nucleotide diversity, Nei's F_{ST} can be defined following the approach of PoPoolation2 as:

$$\widehat{\mathrm{F}}_{\mathrm{FST}}^{\mathrm{PoPool}} = \frac{\pi_{W(T)} - \frac{1}{2}\left(\pi_{W(1)} + \pi_{W(2)}\right)}{\pi_{W(T)}} \quad , \qquad \text{(eq. 2)}$$

where the within, between, and total diversity can be calculated based on frequencies for two populations, coverages, and number of individuals pooled (indicated with subscripts *(1) (2)* for the two populations, and *(T)* when combined, for which the coverage and individuals take the minimum of the two populations) as:

$$\begin{aligned}
\widehat{\pi}_{W(1)}^{\mathrm{PoPool}} &= \frac{n_{(1)}}{n_{(1)}-1} \cdot \frac{C_{(1)}}{C_{(1)}-1} \cdot \left(1 - \sum_{\tau} f_{\tau(1)}^{2}\right) \\
\widehat{\pi}_{W(2)}^{\mathrm{PoPool}} &= \frac{n_{(2)}}{n_{(2)}-1} \cdot \frac{C_{(2)}}{C_{(2)}-1} \cdot \left(1 - \sum_{\tau} f_{\tau(2)}^{2}\right) \\
\widehat{\pi}_{W(T)}^{\mathrm{PoPool}} &= \frac{n_{(T)}}{n_{(T)}-1} \cdot \frac{C_{(T)}}{C_{(T)}-1} \cdot \left(1 - \sum_{\tau} \frac{1}{2}\left(f_{\tau(1)} f_{\tau(2)}\right)\right) \quad . \qquad \text{(eq. 3)}
\end{aligned}$$

The above observed metrics are most useful for inferring processes within E&R experimental populations with known founders. When using Pool-Seq for natural populations, it may also be helpful to infer population parameters such as the population mutation rate $\theta(4N_e\,\mu)$ from empirical diversity estimates such as π while accounting for Pool-Seq errors. The general strategy described in PoPoolation (Kofler et al., 2011a, 2011b) and reimplemented $\mathtt{g_renedalf}$ is described in detail in the **Supplemental Mathematical Appendix**.

Experiment 1: Sequencing a seed mixture of 231 genotypes to characterize a diversity panel

Rationale: In this experiment, we established a genetically diverse panel of *A. thaliana* natural accessions. We sequenced the seed mix of this panel to assess the ability of Pool-Seq to correctly recover genome-wide allele frequencies. This was the first step to use the seed pool for further E&R experiments (see below).

Setup: The founder seed mix for this experiment was sourced from the seeds of 231 genotypes, 229 of which are part of the 1001 Genomes Project (2016) and available from the Arabidopsis Biological Resource Center (ABRC) under accession CS78942 (https://abrc.osu.edu/stocks/465820), while the remaining 2 genotypes were sourced through the Israel Plant Gene Bank (https://igb.agri.gov.il/) under accession numbers 24208 and 22863 (**Dataset S1**). Seeds were pooled at roughly equal yet variable proportions based on weight (See **Dataset S1** for estimated numbers of seeds per ecotype). Note that differences in seed proportions are intendedly captured by directly sequencing seeds below.

Analysis: Eight tubes, each containing about 2,470 seeds (estimated based on weight) from the founder seed mix (**Table S1**) were homogenized using a FastPrep-24 (MP Biomedicals, Irvine, CA, USA). DNA extraction was done using a Qiagen DNeasy Plant Mini kit (Hilden, Germany) (**Supplemental Appendix I: DNA extraction**). One TruSeq library was prepared from each DNA extract. The eight TruSeq libraries were multiplexed and sequenced together on one lane of a HiSeq 3000 sequencer (Illumina, San Diego, California, USA). The total sequencing output was 9.54×10^{10} base pairs and the average genome-wide coverage was ~500X across all seed pool sequencing data (**Fig. S5**). Raw sequence data were processed with our `g`$_{r}$`enepipe` workflow (Czech and Exposito-Alonso, 2021) to trim and map the reads against the *A. thaliana* TAIR10 reference genome (Berardini et al., 2015; Lamesch et al., 2012). Subsequently, using our `g`$_{r}$`enedalf` tool, we calculated the raw minor allele frequencies (MAF) at each biallelic position, based on bam/pileup files counting the ratio of reads containing either reference or alternative alleles (**Fig. S1**). Since users of Pool-Seq may utilize popular computationally efficient variant callers used in individual sequencing, we also ran `g`$_{r}$`enepipe` with three different variant callers: BCFtools (Li, 2011), freebayes (Garrison and Marth, 2012), and the GATK HaplotypeCaller (McKenna et al., 2010). These tools are not primarily designed for calling variants and their frequency from Pool-Seq data, but the resulting VCF file of each caller can be turned into a frequency table by extracting the Allelic Depth ("AD") format field at each genome position for each sample, a process also implemented in `g`$_{r}$`enedalf`. We also tried to run GATK HaplotypeCaller and freebayes using the average pool size as the ploidy options (`--ploidy 2470` and `--ploidy 2470 --pooled-discrete`, respectively, as well as `--pooled-continuous` in freebayes; note that *A. thaliana* is diploid although inbred, but pooling ~2,500 seeds would make the DNA library highly ploid from a computational point of view). These analyses resulted in prohibitively long runtimes even in cluster environments (GATK HaplotypeCaller) and large memory usage (freebayes), demonstrating these tools' limited capabilities for analyzing large datasets and large pool sizes. We hence ran the three callers with default ploidy of 2 to study their artifacts in Pool-Seq applications, assuming that other researchers may be required to resort to these default settings.

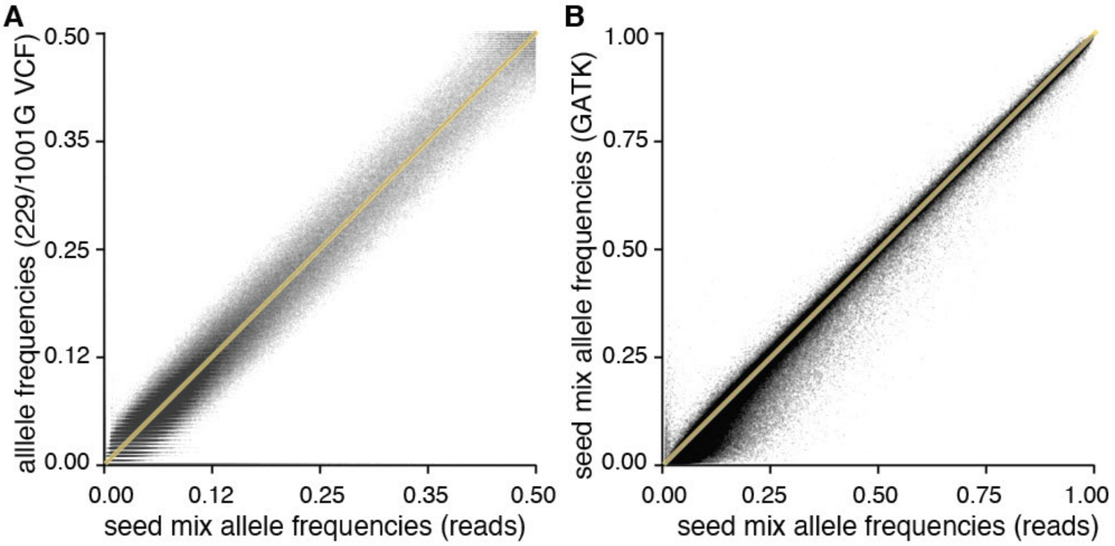

Fig. 1 | Direct sequencing of experimental founder seeds captures the 1001 Genomes variation.
***(A)** Comparison of minimum allele frequencies directly estimated from ratios of bases in reads from sequencing the seed mix (x-axis) and allele frequencies calculated* in silico *from the 1001G VCF subsetted to the genotypes shared with the seed mix (y-axis)* ***(B)** Comparison of allele frequencies from the seed mix likewise directly calculated from ratios of bases in reads (x-axis)* vs *from the allelic depth ("AD") VCF field after calling SNPs using GATK with default settings (y-axis). Yellow lines indicate y=x line.*

Results: We conducted two comparisons, the first to quantify how well direct sequencing of pools of seeds captured the variation found in the 229 genotypes from the 1001 Genomes, and the second to study the technical artifacts generated by diploid SNP callers.

The first comparison is based on the raw frequency of alternative allele counts divided by coverage in the seed sequencing bam/pileup files against the same SNPs using the 229 columns in the 1001 Genomes VCF table corresponding to the genotypes mixed at roughly equal proportions in the seed mix (comparisons conducted with 1,353,386 *bona fide* SNPs with minimum alternative allele count >2). This yielded a high correlation and low deviation from the y=x correspondence line (**Fig. 1A**, Pearson's *r* = 0.982*, SD* = 0.0214; for unfiltered comparison see **Fig. S3A-B**). Of note, comparing seed allele frequencies to all 1,135 individuals from the 1001 Genomes (i.e. not only the 229 included in the seed mix) shows a high density of alleles that are at low-to-intermediate frequencies in the 1001 Genomes but at low frequency in the seeds (**Fig. S3A,C**), likely indicative of a rare and highly divergent population group, the so-called relict accessions, comparatively underrepresented in our Pool-Seq subset of the larger set of 1001 Genomes (1001 Genomes Consortium, 2016). All in all, we were able to recover nearly all of the SNP variation present in the 229 individual founder genotypes with our Pool-Seq approach.

The second comparison assessed the variation in allele frequency recovery from raw allele counts in bam/pileup files and standard diploid SNP callers. This revealed that certain tools generated frequency estimates upwardly or downwardly biased compared to the raw allele fraction from bam/pileup files (**Fig. S6-10**). This bias, especially for alleles found at low frequency (<20%), appears most dramatic in GATK HaplotypeCaller in its default diploid likelihood mode (**Fig. 1B, Fig. S6-10**). Further, very low frequency alleles (<4%) appear missing (**Fig. S11**). GATK, which is tuned for human SNP calling, aims to call genetic variants that fit the reference homozygote, heterozygote, or alternative homozygote scheme, and utilizes local genome realignment information to reject certain reads, which may be causing unexpected biases (see asymmetry in low-frequency SNPs in **Fig. 1B**, where only 1,353,386 *bona fide* SNPs with minimum count >2 were used). While correlations between raw allele frequency and SNP calling-based allele frequencies are typically high (*r* >0.99), deviations can be substantial without filters. For instance, the standard deviation of differences between GATK and raw allele ratio frequencies suggests deviations higher than 10% (SD_{GATK}=0.094-0.204 depending on coverage cutoffs, **Fig. S6-8)**. BCFtools and freebayes appeared less biased and more consistent ($SD_{BFCtools}$=0.051-0.156 and $SD_{freebayes}$=0.043-0.097, **Fig. S6-8**). Based on these results, we recommend, despite their computational capacity and popularity, to avoid SNP callers designed for individual sequencing for Pool-Seq data. In conclusion, $g_r$enedalf offers computational speed, generates frequency tables from raw sequencing reads, allows for data manipulations such as subsets or sample comparisons, and implements quality filters shown to provide appropriate frequency estimates if evaluated in a set of *bona fide* SNPs (see Experiment 2 below) (Guirao-Rico and González, 2021).

Experiment 2: Two-genotype analysis to understand biases of DNA contribution to pooled samples and sequencing noise

Rationale: One important assumption in population inferences based on Pool-Seq data is that each individual contributes an equal amount of sequencing reads. However, the deviation in DNA contribution by pooling organs from different individuals or entire individuals has not been tested in *A. thaliana* or other model plant systems (although it is common practice in *D. melanogaster* to directly pool whole flies, see for instance Tilk *et al.* (2019)). Instead, a typical approach in many state-of-the-art Pool-Seq experiments is to extract DNA separately from different individuals and subsequently pool equal amounts of DNA, an unfeasible approach when studying thousands or tens of thousands of individuals (Gautier et al., 2013; Rellstab et al., 2013; Roda et al., 2017). Whether flower organ sizes, such as those described in *A. thaliana* across ecotypes (Juenger et al., 2000), or cell ploidy differences via endoreplication in sepals (Robinson et al., 2018) have an effect in differential DNA contributions when pooling flowers, is unknown and could be manifested in deviations of allele frequencies. In Experiment 2, we sequenced a pool of two *A. thaliana* genotypes sampling one flower each (i.e., the smallest possible pool size *n*=2) and tested it against carefully quantified and pooled DNA isolates of the same two genotypes to assess the variation in DNA contribution.

Setup: To quantify the deviation in DNA content when pooling two flowers from distinct genotypes, we sequenced three replicates of two flowers each. The first genotype was the laboratory inbred strain Col-0, which was the type strain used to assemble the reference genome of *A. thaliana* (Lamesch et al., 2012). The second, a natural accession (inbred in greenhouse propagations) from the 1001 Genomes project (1001 Genomes Consortium, 2016), was RUM-20 (#9925), which differs from Col-0 by 1,007,560 SNPs according to the 1001 Genomes data (note that the average genotypic difference of any two genotypes is 400—600K SNPs; we hence picked a relatively divergent accession). These two ecotypes did not show visible flower size differences, but were not chosen based on their flower size differences. To compare the pooled flower method with the conventional method where DNA is pooled at equal proportions, we extracted DNA from a leaf of a Col-0 individual and a leaf of a RUM-20 individual, and generated three DNA replicates via equal pooling by DNA concentration before library preparation (**Fig. 2A**, **Table S2**). DNA was extracted with the CTAB method and processed into whole-genome sequencing libraries using a modified Nextera protocol (**Supplemental Appendix I: DNA extraction** and **Library preparation**).

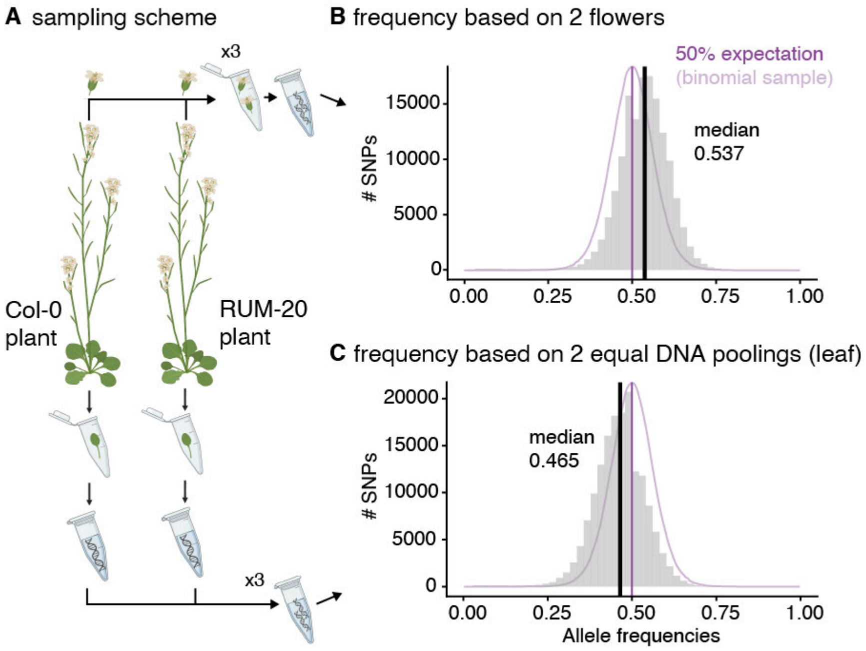

Fig. 2 | Experimental design (Exp. 2) to test the relative contribution to DNA sequencing output
(A) *Flower and leaf tissues were sampled from two genotypes, Col-0 and RUM-20. Three replicates of two flowers were collected (the Pool-Seq method) while leaves were collected individually (conventional method). Leaf DNA was pooled at equal quantity to create three replicates of DNA input for library preparation.* ***(B)*** *Distribution of allele frequencies in one of the three replicates of the directly extracted and whole-genome sequenced 2-flower pools (see all replicates in* ***Fig. S14****; allele frequencies of SNPs that passed mapping quality filters, had a minimum minor allele count >2, and were present in the 1,353,386* bona fide *SNPs).* ***(C)*** *The equivalent of (B) for two separate leaf DNA extracts carefully pooled at equal concentration.*

Analysis: Assuming that both pooled individuals from the two inbred lines are indeed homozygous, one would expect polymorphic alleles to be at exactly 50% proportion if the tissues of both individuals contributed exactly equal amounts of DNA. Mean deviations from 50% would indicate differences in DNA content and/or mapping bias if deviations are systematic for one genotype. In addition, variations in the extent of deviation across replicates would indicate sampling noise due to limited DNA sequencing coverage. To test this, we again trimmed and mapped genome-wide reads with our `grenepipe` workflow, and computed frequencies from bam/pileup files with `grenedalf`, as described in Experiment 1.

Results: This proof-of-concept analysis provided a number of clues on the power and potential biases of Pool-Seq to study population evolution in real time. Firstly, as the expected frequencies of polymorphic sites are around 50%, we could detect many low frequency alleles that are most likely artifacts (**Fig. 2B-C**). This could not be done while sequencing large populations of seeds or flowers because we expect many true low frequency alleles. We show that a lack of filters for coverage, mapping quality, and most importantly, minimum alternative allele count, leads to a majority of calls of polymorphic sites likely representing artifacts (e.g., 90% of unfiltered SNPs may be false positives, see **Fig. S15A**; these filters were applied in all experiments). We therefore implemented stringent filters for mapping quality (samtools option `-q 60`), base quality (option `-Q 30`), and matching of forward/reverse read mapping (options `--rf 0x002 --ff 0x004 --ff 0x008`), and minimum allele counts in the bam/pileup file of reads (MAC>2). In combination with a filter for the *bona fide* 1,353,386 SNPs, these filters led to the expected distribution of allele frequencies around 50% with some of the remaining variation likely explained by the binomial sampling variance caused by limited coverage (**Fig. 2B-C, Fig. S13**). Third, we could show that the deviation of average allele frequencies from 50% was small (2.2% frequency, **Fig. S14D-F**) and of similar magnitude as the deviation measured in the DNA pools generated from DNA isolates of equal concentration (1.5%, **Fig. S14A-C**). This suggests that uncontrolled factors (such as flower size, endoreplication and ploidy, differential tissue grinding) minimally affect DNA contributions of flowers, and that their magnitude is comparable to variable DNA contributions from individual samples even after DNA normalization. Such small deviations become statistically diluted when pooling large numbers of individuals (Lynch et al., 2014). For instance, for 100 flowers, errors would range from 0.0004 to 0.1% for allele frequencies from 1% to 50%. This is in agreement with previous Pool-Seq experiments with whole *D. melanogaster* flies, which indicates that allele frequency estimation per population requires 100 individuals at 50X coverage for virtually-perfect allele frequency retrieval (Gautier et al., 2013). In summary, the Pool-Seq approach using large numbers of *A. thaliana* flowers, sampling one flower per individual, should provide highly reliable allele frequency inferences in E&R experiments.

Experiment 3: Combinatorial experiments of pool sizes and tissue type sequencing to determine optimal sampling schemes

Rationale: In this experiment, we evaluated the ability of Pool-Seq to recover correct allele frequencies from pooled samples made up of 5 to 100 flowers (one per individual) and leaves sampled from *A. thaliana* plants. We studied whether (A) individual leaf DNA extraction and library preparation with equal DNA input and (B) pooled flower DNA extraction and library preparation without DNA normalization produce comparable population estimates.

Setup: We grew a mixture of seeds of 231 genotypes mixed roughly at equal proportions (**Dataset S1**) in 2,500 pots with one individual each (replicating similar conditions of large evolving populations outdoors, see Experiment 4). We then selected 50 random plants for our test. Flowers were sampled from different subsets of these 50 plants to assess the effect of increasing the number of individuals randomly sampled: 5, 10, 25, 50, and 100 flowers (**Fig. 3**; for 100 flowers, we included 2 flowers from each of the same 50 plants). For the same plants for which flowers were collected, we also removed, and separately stored, one leaf per plant for independent DNA extraction.

Tissue grinding, DNA extraction and library preparation steps are described in **Supplemental Appendix I: DNA extraction and library preparation**. Leaf DNA pooling was done for the same individual combinations for which flower subsamples of 5, 10, 25, and 50 individuals were taken and pooled (see **Table S3-4** and **Fig. S2** for combinations). Therefore, we expect the allele frequencies of the equimolar pool of leaf DNA and that of the flower extracts to be close to identical (as in Experiment 2), unless scaling the Pool-Seq method to many individuals incurs systematic biases.

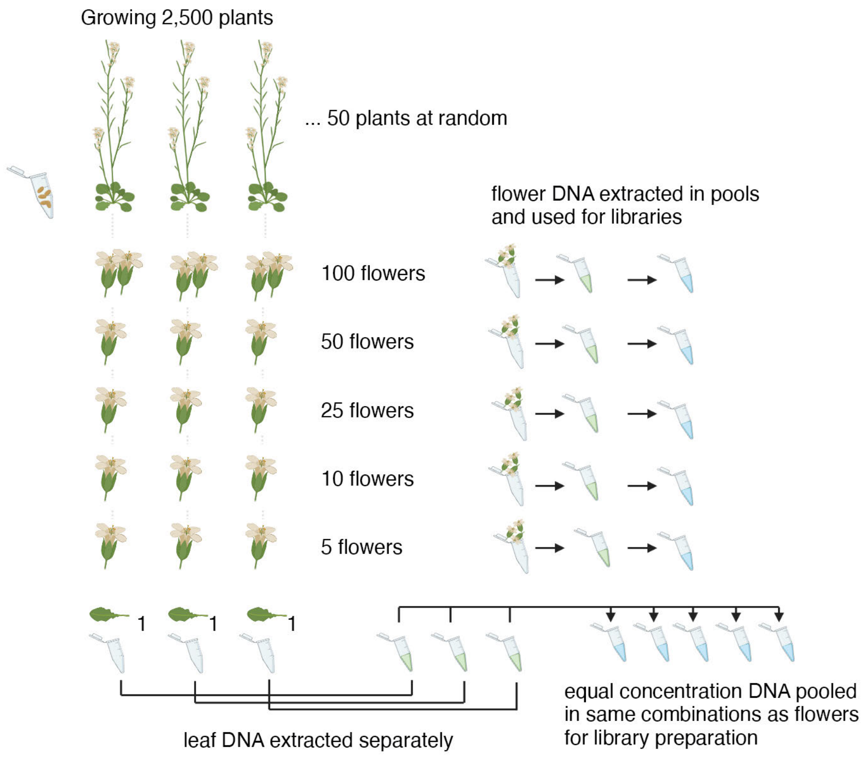
Growing 2,500 plants
... 50 plants at random
flower DNA extracted in pools and used for libraries
100 flowers
50 flowers
25 flowers
10 flowers
5 flowers
1
1
1
leaf DNA extracted separately
equal concentration DNA pooled in same combinations as flowers for library preparation

Fig. 3 | Experimental design (Exp. 3) to test Pool-Seq with plant flowers

A total of 50 plants sown at random from 231 diverse genotypes of A. thaliana *(**Table S2**) were individually grown and sampled in different combinations and in replication (5, 10, 25, 50. In the 100 flower sampling, 50 individuals sampled twice) (**Table S4**).*

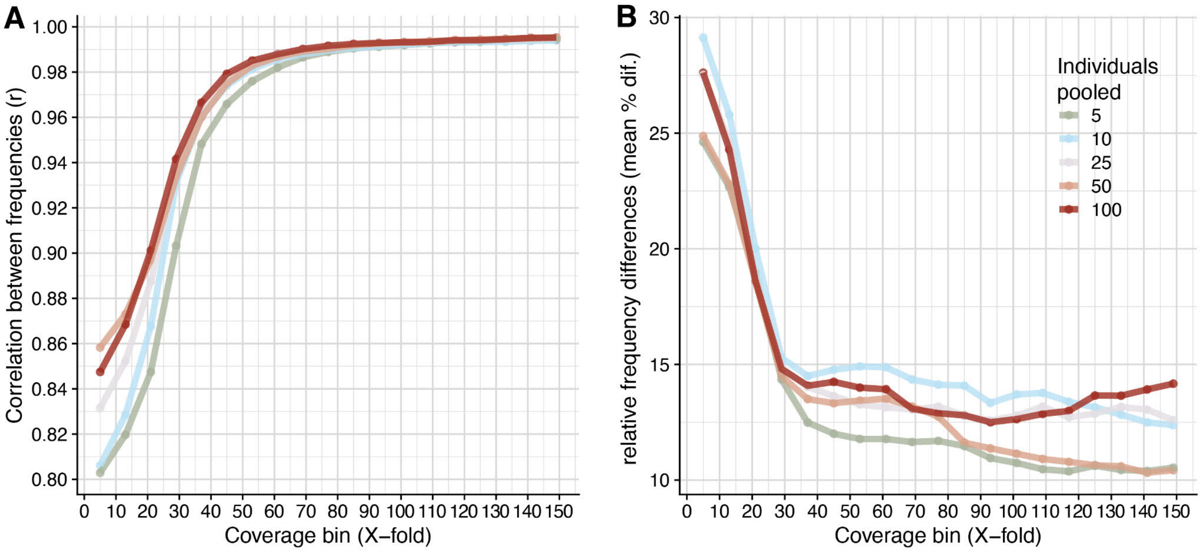

Fig. 4 | Correlation between allele frequencies estimated from direct sequencing of pooled flowers **vs** ***individual DNA extracts pooled at equal concentration.***

*Genome-wide allele frequency comparisons between the same set of 5,10, 25, or 50 individuals as estimated from directly extracting DNA and sequencing from pooled flowers and from sequencing of pooled independent DNA extracts. (**A**) Pearson's correlation between allele frequencies across coverage bins (all alleles with minimum alternative allele count >2 and represented in the* bona fide *11,769,920 set (further subsets based on other quality thresholds did not provide enough data points for coverage breakdown). (**B**) Relative % error of the difference between flower pools and DNA pools across coverage bins.*

Results: We calculated the correlation between allele frequencies recovered from pools of flowers and equal leaf DNA pools both originating from the same sets of plants. Because noise decreases with both increasing numbers of individuals and increasing sequencing coverage, we leveraged the variation in coverage along the genome to compute correlations in increasing coverage bins. Frequencies were highly correlated ($r > 0.98$) for all combinations as long as coverage was over 50X (**Fig. 4**). The small mean relative frequency differences (<15%) of alleles at medium (~40X) coverage for virtually all pairs of flowers or DNA pool libraries suggests that, even when there are small experimental pooling errors (**Fig. 2B-C**), the large number of sequenced individuals dilutes errors (**Fig. 4B**).

Experiment 4: Multi-year field experiment to showcase the power of Pool-Seq to track rapid evolution

Rationale: Ultimately, the cost-effective and scalable Pool-Seq approach is designed to track evolution of populations through time. To showcase its strengths, we conducted an outdoor experiment over two growing seasons starting from a large population of diverse *A. thaliana* genotypes.

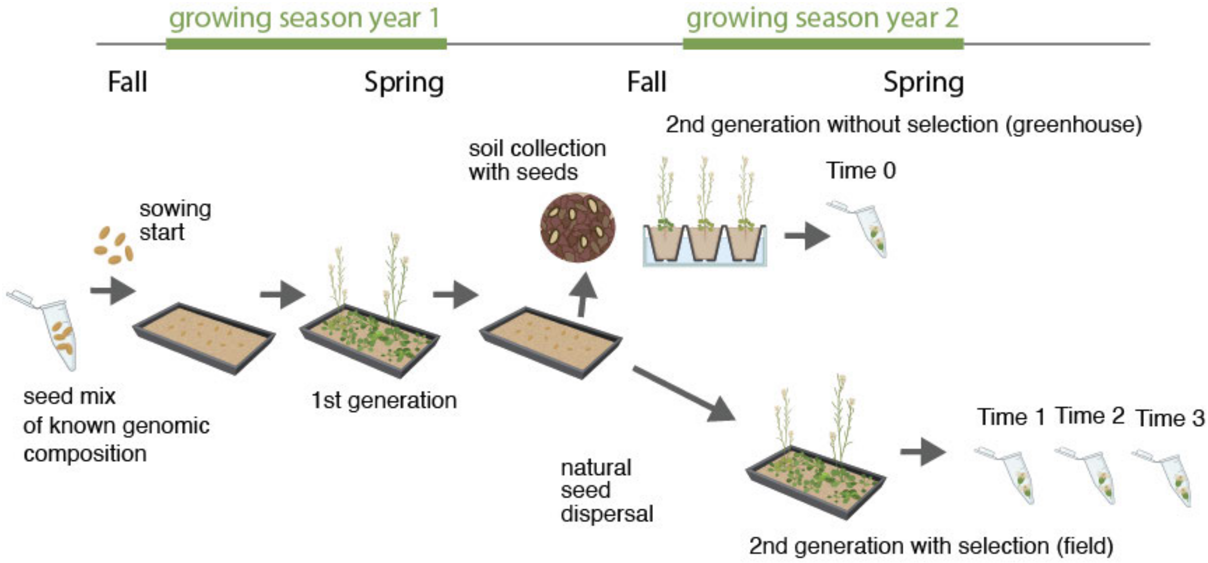
growing season year 1
growing season year 2
Fall
Spring
Fall
Spring
2nd generation without selection (greenhouse)
soil collection with seeds
Time 0
sowing start
seed mix of known genomic composition
1st generation
natural seed dispersal
Time 1
Time 2
Time 3
2nd generation with selection (field)

Fig. 5 | Design of the field experiment (Exp. 4)
A plot was set up containing a mixture of seeds of 451 natural genotypes mixed at equal proportions. After an entire generation of growth in the field with natural seed dispersal, two parallel samplings were conducted. One sampling of soil was conducted in fall prior to natural germination and then planted in a greenhouse and subject to environmental conditions favorable for germination and growth to limit natural selection. The second sampling was conducted in spring after natural germination had occurred and plants were exposed to natural selection that could have led to mortality and survival of different genotypes. Experimental populations outdoors were sampled three times due to longer flowering periods in outdoor conditions.The whole experiment was replicated three times in parallel.

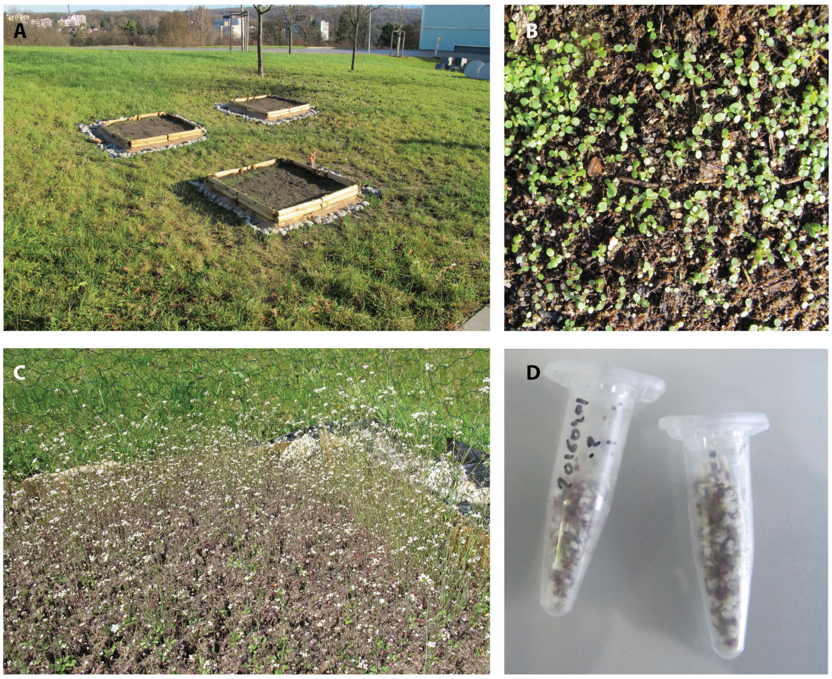

Fig. 6 | Photos of the field experiment (Exp. 4)
*(**A**) Setup of 3 population replicates. (**B**) Close-up of germinating seedlings. (**C**) Abundant flowering shown in one of the replicate plots. (**D**) Sampled flowers for Pool-Sequencing.*

Setup: This experiment was performed in an experimental field at the Max Planck Institute of Biology campus (48.537723, 9.058746, Tübingen, Germany, **Fig. 5-6**), using a seed mix of 451 natural genotypes generated from 2 plants of each genotype and 10 siliques each (ca 90,000 seeds). This set largely overlaps with the 1001 Genomes genotypes, and therefore the starting allele frequencies are known (available in ABRC stocks under accession CS78942, https://abrc.osu.edu/stocks/465820, **Dataset S1**). The seed mixture was split in nine tubes which were sown at three different time points (November 2014, February 2015, and March 2015) in three independent 1×1 m^2 plots (**Fig. 6A**). This design was used to reduce the chance of disturbance events occurrences that would impact germination. After the first generation had dispersed seed during late spring in the field and synchronized with the local climate and photoperiod, soil samples were collected prior to the second season's natural germination (early fall 2015) and transferred to an indoor greenhouse with optimal conditions for the species, to enable germination, survival, and reproductive success for as wide a set of genotypes as possible. From the three plots, 56, 69, and 101 adult plants were sampled from the growth chamber as a baseline (Time 0 in **Table S6 and Fig. 5**). From the field plots, in the following spring, 164, 415, and 593 surviving and reproducing adults of the second generation were sampled for sequencing at 3 different time points to capture the entire temporal window of flowering (see number of adults per time sample in **Table S6 and Fig. 5**). We aimed to sample one flower per individual, paying attention to sample from small to large plants uniformly. The flowers collected at Time 1, 2, and 3 were sequenced separately. A total of 1,398 individuals were sequenced in 12 pools of replicate x time point combinations (**Table S6**). We used the population genetic statistics of PoPoolation2 as re-implemented in grenedalf to determine genome-wide patterns of F_{ST} across all combinations of replicates and time points accounting for pool size (**Table S6**) and genome-wide variation in coverage.

Results: Plants successfully established in dense patches in the experiment (**Fig. 6B-C**). Tens of thousands of seedlings were observed per plot replicate, which theoretically should enable efficient natural selection (Charlesworth and Charlesworth, 2010). We observed genetic differentiation based on F_{ST} between the baseline (*Time 0*, offspring of the first generation read in the greenhouse) and flowers of surviving individuals of generation 2 (*Time 1,2,3*) was higher ($F_{ST} \approx 0.07$, **Fig. S15**) than differentiation between several independent DNA extractions of subsets of the founder seed mixes (Exp. 1, $F_{ST} \approx 0.0208$, **Fig. S14**). Although we think such genome-wide patterns are likely mostly driven by drift in the wild, a scan along the genome identified several F_{ST} peaks between *Time 0* and *Time 1-3*, revealing genomic regions that diverged above the background noise level (**Fig. 7, Fig. S17,19**).

One of the observed peaks is localized in chromosome 5 near the gene *FLOWERING LOCUS C* (*FLC*, AT5G10140), encoding a MADS-box transcription factor and master regulator of flowering time. The region with elevated F_{ST} is located 5' of the transcription start site of *FLC* (ca. -2.5—0.5K, **Fig. 7**), suggesting that variation in the promoter region was under some form of natural selection in these experiments and thus shifted in allele frequency. Average per-SNP F_{ST} from *Time 0* to all other time points was higher within the approximate promoter region compared to the rest of the genome (mean [95% quantile] = 0.0548 [0.285] in promoter vs. 0.0380 [0.160] outside; Wilcoxon signed-rank test $P = 2.07 \times 10^{-6}$, 5.90×10^{-6}, 2.42×10^{-4}, respectively for the three field E&R replicates) (**Fig. 7A**)**.** That the same F_{ST} peak is recovered by comparing two cohorts in the flowering seasons, *Time 1* vs. *Time 3* , further suggests variation in this genomic region may play a role in determining early vs. late flowering (**Fig. 7B**). Not only that, but raw allele frequency changes from the starting mix of 451 natural genotypes to all sampled flowers (1,172) two generations after the start of the experiment (**Fig. 7C**).

We leveraged the fact that Experiment 4 was conducted in parallel to a previous common garden experiment 1.51 km away (48.545809, 9.042449) with similarly rich and highly-overlapping *A. thaliana* genotype sets (Exposito-Alonso et al., 2019). In the common garden, each genotype was (individually) scored for an estimated number of seeds per plant that reached adult reproductive stage. Using an imputed matrix of the 1001 Genomes (http://arapheno.1001genomes.org, https://aragwas.1001genomes.org) and a Linear Mixed Model (Kang et al., 2008), we conducted Genome-Wide Associations (GWA) to identify genetic variants that explained variation in seed set per plant (**Fig. 7D**), specifically in the "thp" condition of that experiment: Tübingen, high rainfall, population replicate, (Exposito-Alonso et al., 2019) (For similar evidence in the "mli": Madrid, low rainfall, individual replicate, see **Fig. S18**). This common-garden-scored fitness and GWA approach is one of the most direct ways to quantify natural selection driven by a specific environment (Exposito-Alonso et al., 2019; Gompert et al., 2017). It is expected that genetically-based fitness differences among plants would lead to genotype and allele frequencies changes over time (although such multi-generational experiments are not often conducted). As expected, we also found an overlap between the above peak of temporal F_{ST} allele frequency differentiation in Experiment's 4 E&R and moderate fitness-associated SNPs in the parallel common garden (**Fig. 7D**), with an average of fitness effect sizes significantly elevated within the same region observed above (Wilcoxon test $P = 0.0313$). The fact that flowering time, manually scored in the parallel common garden, was negatively correlated at the plant level with relative seed production (Spearman's rank correlation $r = -0.404$, S = 31048965, $P < 2.2 \times 10^{-16}$) and survival ($r = -0.187$, S = 26399658, $P = 2.074 \times 10^{-5}$) further supports our finding that natural selection may have driven frequency changes in alleles in the *FLC* locus in our multi-year E&R field experiment. While the signal in the *FLC* locus is more readily interpretable, and is thus a helpful example to illustrate the application of our methods, this region is far from being the only region displaying strong temporal differentiation (**Fig. S19**). Multiple regions had F_{ST} > 0.2 and showed parallel patterns in three or more replicates or temporal samples of flowers (**Dataset S2**). Although some genes involved in disease or dehydration responses are suggestive, most difficult-to-interpret peaks will deserve more attention in future studies. All in all, our experiment fulfills the purpose of testing the ability of a simple and cost-effective Pool-Seq approach to detect rapid evolution of plants subject to strong natural selection pressures at resolutions comparable even to those of time-intensive and costly common garden experiments and Genome-Wide Association studies.

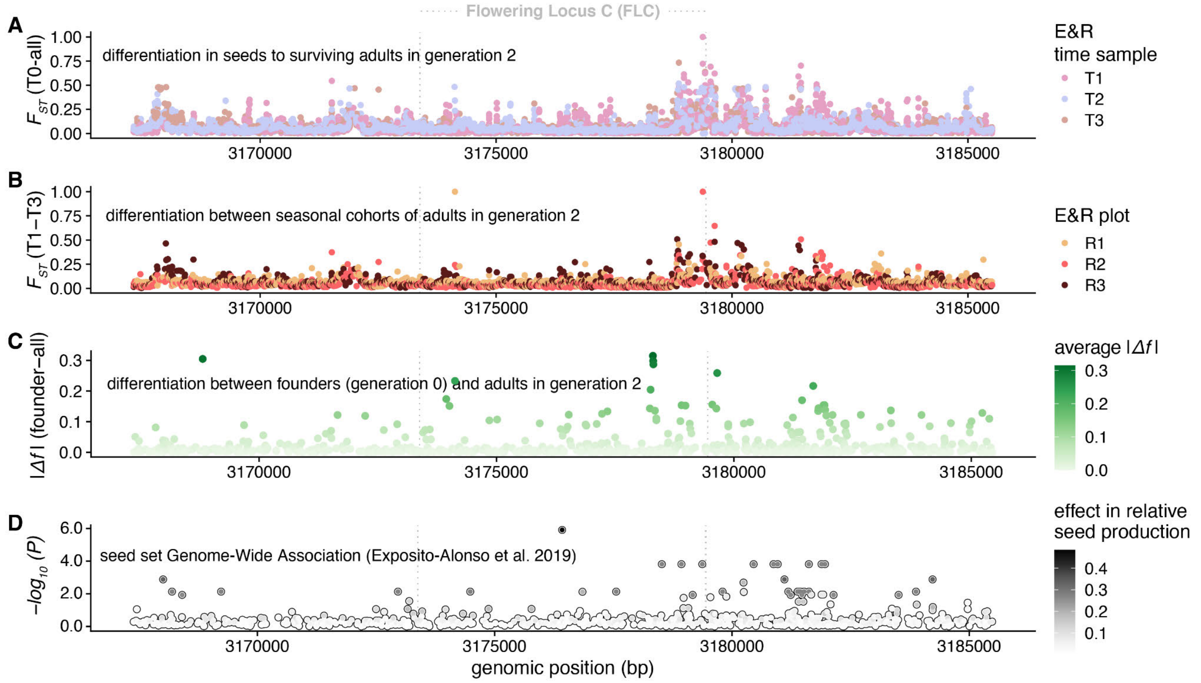

Fig. 7 | Temporal allele frequency change in a multi-year Evolve & Resequence experiment compared to fitness effects in a common garden experiment in the FLC region.

*(**A-B**) Temporal allele frequency differentiation (F_{ST}) in the Flowering Locus C region on chromosome 5 showing peaks of differentiation around the first exon and the upstream promoter region of the gene (positions around 3,180,000; note the protein coding strand is the reverse strand). (**A**) Differentiation between the baseline "without selection" (Time 0) and the flower samples of surviving adults in nature at three time points (Time 1-3). (**B**) Differentiation between the earliest and latest flowering cohort for the three replicate plots. (**C**) Average allele frequency change between the founder seed mix (one generation prior to Time 0) to adults sampled in generation 2 (Time 1-3). (**D**) Genome-Wide Association between genetic variants in the 1001 Genomes and outdoor seed production in a common garden experiment (Exposito-Alonso et al., 2019) 1.51 kilometers away from the Evolve & Resequence experiment in (A-B).*

Discussion and Outlook

The paradigm that evolution is a slow process is being challenged by more and more evidence from experiments with both animals and plants that allele frequencies within populations fluctuate or change in the span of seasons or decades following environmental changes (Bergland et al., 2014; Franks and Weis, 2008). Scalable whole-genome sequencing approaches based on Pool-Seq (Schlötterer et al., 2014) have enabled the generation of population genomic datasets across continental scales such as "*Drosophila* Evolution over Space and Time" (DEST) (Kapun et al., 2021) (https://dest.bio) or large-scale multi-generational Evolve & Resequence experiments with *D. melanogaster* (Rudman et al., 2021). Projects of a similar scale for plants are currently rare, a notable exception being the barley composite cross long-term evolution experiment initially developed by Harlan and then continued by Jain and Allard (Allard and Jain, 1962; Suneson, 1956). Here we present Pool-Seq laboratory protocols and new efficient software implementations which are scalable to high-throughput, longitudinal experimental evolution studies of thousands of plant populations at low cost.

We here presented an open-source and streamlined frequency calling pipeline that automatically downloads, checks, and runs all the required software tools from raw fastq file to a frequency table of Pool-Seq samples (Czech and Exposito-Alonso, 2021). Such a reproducible pipeline also facilitates parallel runs with different pipeline parameters for tool benchmarking and quality controls for tool parameter comparisons. We show that if a set of *bona fide* SNPs is already known for the species, as is the case with *Arabidopsis thaliana*'s 1001 Genomes Project catalog (1001 Genomes Consortium, 2016), estimation of allele frequencies from mapped reads is successful without the need for sophisticated SNP callers to identify new variation, as long as there is sufficient coverage and quality filters are implemented (Guirao-Rico and González, 2021; Tilk et al., 2019). Two cases may benefit from further tool implementation in $\texttt{g}_{\text{r}}\texttt{enepipe}$: In the absence of *bona fide* SNPs, Pool-Seq-specific likelihood or Bayesian SNP callers such as SNAPE are ideal to discover new SNPs while reducing false positives (Guirao-Rico and González, 2021). In the presence of ultra-low coverage sequencing, if individual sequencing of founders is available, allele frequency estimates can be further improved using simulations and linkage disequilibrium information based on the tools HARP and HAFpipe (Kessner et al., 2013;Tilk et al., 2019).

To enable faster and more user-friendly Pool-Seq-based evolutionary analyses at scale, we have developed $\texttt{g}_{\text{r}}\texttt{enedalf}$. This tool re-implements the now-classic PoPoolation software (Kofler et al., 2011a) in C++ from the ground up and expands its functionality and types of compatible input file formats. Speed improvements in the order of ~100X now enable conducting, for instance, pairwise F_{ST} calculations among thousands of samples in hours rather than months (Czech and Exposito-Alonso, 2022).

With these bioinformatic improvements in hand, we show that direct whole-genome sequencing of a mixture of seeds can properly characterize the standing genetic variation of a hypothetical starting pool of founder individuals for an E&R experiment. Further, direct sampling of flower tissues (or similarly-sized organs or leaf punches) also enables efficient genetic tracking of plant populations with tens of thousands of individuals over time—a scale currently not feasible for experiments with separate individual DNA extracts or library preparations (Fracassetti et al., 2015; Gautier et al., 2013; Rellstab et al., 2013; Roda et al., 2017). This sampling method potentially provides an alternative experimental design to common garden experiments, and its simplicity would potentially facilitate citizen-science real-time evolution projects in large organisms.

Finally we showcase that the described Pool-Seq protocols can be applied in large outdoor E&R experiments using *A. thaliana* seed resources. The fact that linkage decays surprisingly fast in *A. thaliana* (**Fig. S9**) (Kim et al., 2007)—probably owing to a ~2-16% outcrossing rate that shuffles enough standing genetic variation (Bomblies et al., 2010; Platt et al., 2010)—may enable identification of narrow mapping regions containing adaptive loci using E&R, perhaps even narrower than what Genome-Wide Associations can currently achieve (**Fig. 7**) (1001 Genomes Consortium, 2016; Atwell et al., 2010). The success of Experiment 4 motivates the use of this approach at a larger scale, and seems to provide a genomic sensitivity similar to labor-intensive common garden experiments that are confined to a few environment (Agren and Schemske, 2012; Exposito-Alonso et al., 2019, 2018; Fournier-Level et al., 2011; Manzano-Piedras et al., 2014) ,

Despite the intriguing and complementary association between fitness effect sizes in common garden experiments and allele frequency changes in our E&R (Experiment 4), the rapid evolutionary signals inferred here are limited to the single environment studied. To comprehensively study rapid evolutionary adaptation across climates using E&R, we have initiated a project called "Genomics of rapid Evolution to Novel Environments" network (GrENE-net), which is a large-scale extension of Experiment 4 presented here. This internationally distributed E&R GrENE-net project involves 45 field sites (https://grenenet.org), was started from the same seed mix of Experiment 1, and has been conducted from 2017 until 2022 (the time of writing)—featuring the largest temporal and spatial scale among known Evolve & Resequence experiments. The accumulating sequencing data, expected to exceed 5Tb and over 2,500 population samples, should enable better temporal and spatial tracking of rapid evolution and understanding of climate × genotype × fitness interactions than any previous large-scale common garden experiment (Exposito-Alonso et al., 2019; Fournier-Level et al., 2011; Lovell et al., 2021). It is our hope that the GrENE-net experiment will enable researchers to establish a direct link between environment and natural selection at the allele frequency level, stimulate theoretical development in evolutionary genetics, and empower plant biologists' search for the genetic basis of adaptation. If biologists wish to forecast plant responses under changing climate conditions, long-term and highly spatially replicated E&Re datasets such as this one will be paramount.

Additional Information

Data and Code availability Reads were deposited at NCBI SRA with accession number: <TBD upon publication>. Genomes of founder populations are available as part of the 1001 Genomes project: http://1001genomes.org/data/GMI-MPI/releases/v3.1/. Scripts of the analyses in this manuscript are available at https://github.com/lczech/grenepilot-paper, which contains all settings used for the runs of our $\mathtt{g_renepipe}$ (Czech and Exposito-Alonso, 2021) workflow for variant calling, as well as all python and R scripts for the figures presented here. Genome frequency manipulations and Pool-Seq-corrected population genetic statistics are implemented in $\mathtt{g_renedalf}$ (Czech and Exposito-Alonso, 2022).

Acknowledgements We thank Robert Kofler for discussing his PoPoolation and PoPoolation2 implementation details. We further thank the members of the Moi Lab for feedback on the manuscript and analyses.

Funding statement This work was funded by ERC-CAPS 1001G+ and the Max Planck Society (D.W.) and by the Carnegie Institution for Science and National Institutes of Health's Early Investigator Award (1DP5OD029506-01) (M.E.-A.). The computing for this project was performed on the Calc, Memex, and Moi Node clusters from the Carnegie Institution for Science.

Disclosure statement D.W. consults for breeding companies and is a co-founder of COMPUTOMICS, which provides service to breeding companies. The other authors declare no competing financial interests. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contribution MEA, FV, JFS, conceived the project after initial discussions with DW. MEA and DW acquired financial support for the project leading to this publication, provided study materials, reagents, materials, laboratory samples, instrumentation, computing resources, or other analysis tools. MEA, FV, managed and coordinated the research activity planning and execution. LC, MEA, JPS, YP, TB, conducted statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data. JH, KF, RS, YP, PL, MEA, and BAR set up and conducted laboratory protocols and plant experiments. MEA, FV, JFS created genotype seed collections. The GrENE-net consortium contributed to experimental design and launched the GrENE-net.org experiments. LC, YP, JSP, PL, TB, MEA wrote the first manuscript draft, and all authors edited and reviewed the latest manuscript version.

References

1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**:481–491. doi: 10.1016/j.cell.2016.05.063

Agren J, Schemske DW. 2012. Reciprocal transplants demonstrate strong adaptive differentiation of the model organism Arabidopsis thaliana in its native range. *New Phytol* **194**:1112–1122. doi:10.1111/j.1469-8137.2012.04112.x

Allard RW, Jain SK. 1962. Population studies in predominantly self-pollinated species. Ii. Analysis of quantitative genetic changes in a bulk-hybrid population of barley. *Evolution* **16**:90–101. doi:10.1111/j.1558-5646.1962.tb03201.x

Anderson JT, Wadgymar SM. 2019. Climate change disrupts local adaptation and favours upslope migration. *Ecol Lett.* doi: 10.1111/ele.13427

Andrews S, Others. 2017. FastQC: a quality control tool for high throughput sequence data. 2010.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**:627–631. doi: 10.1038/nature08800

Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. 2015. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**:e0128036. doi:10.1371/journal.pone.0128036

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**:474–485. doi: 10.1002/dvg.22877

Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. 2014. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLoS Genet* **10**:e1004775. doi: 10.1371/journal.pgen.1004775

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120. doi: 10.1093/bioinformatics/btu170

Bomblies K, Yant L, Laitinen R a., Kim S-T, Hollister JD, Warthmann N, Fitz J, Weigel D. 2010. Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. *PLoS Genet* **6**:e1000890–e1000890. doi:10.1371/journal.pgen.1000890

Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, Keane T, Davies RM. 2021. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**. doi:10.1093/gigascience/giab007

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F. 2010. Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLoS Genet* **6**:e1000940. doi: 10.1371/journal.pgen.1000940

Capblancq T, Fitzpatrick MC, Bay RA, Exposito-Alonso M, Keller SR. 2020. Genomic Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes. *Annu Rev Ecol Evol Syst* **51**:245–269. doi: 10.1146/annurev-ecolsys-020720-042553

Charlesworth B, Charlesworth D. 2010. Elements of Evolutionary Genetics. W. H. Freeman.

Clausen J, Keck DD, Hiesey WM. 1941. Regional Differentiation in Plant Species. *Am Nat* **75**:231–250.

Czech L, Exposito-Alonso M. 2022. grenedalf: population genetic statistics to study rapid evolution with Pool-Seq. https://github.com/lczech/grenedalf

Czech L, Exposito-Alonso M. 2021. grenepipe: A flexible, scalable, and reproducible pipeline to automate variant and frequency calling from sequence reads. *arXiv*.

Endler L, Betancourt AJ, Nolte V, Schlötterer C. 2016. Reconciling Differences in Pool-GWAS Between Populations: A Case Study of Female Abdominal Pigmentation in Drosophila melanogaster. *Genetics* **202**:843–855. doi: 10.1534/genetics.115.183376

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**:3047–3048. doi:10.1093/bioinformatics/btw354

Exposito-Alonso M, 500 Genomes Field Experiment Team, Burbano HA, Bossdorf O, Nielsen R, Weigel D. 2019. Natural selection in the *Arabidopsis thaliana* genome in present and future climates. *Nature* **573**:126–129. doi: 10.1038/s41586-019-1520-9

Exposito-Alonso M, Brennan AC, Alonso-Blanco C, Picó FX. 2018. Spatio-temporal variation in fitness responses to contrasting environments in Arabidopsis thaliana. *Evolution*. doi:10.1111/evo.13508

Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. 2011. A map of local adaptation in Arabidopsis thaliana. *Science* **334**:86–89. doi:10.1126/science.1209271

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

Franks SJ, Weis AE. 2008. A change in climate causes rapid evolution of multiple life-history traits and their interactions in an annual plant. *J Evol Biol* **21**:1321–1334. doi:10.1111/j.1420-9101.2008.01566.x

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

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bioGN]*.

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

Gompert Z, Egan SP, Barrett RDH, Feder JL, Nosil P. 2017. Multilocus approaches for the measurement of selection on correlated genetic loci. *Mol Ecol* **26**:365–382. doi:10.1111/mec.13867

Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. 2017. The dynamics of molecular evolution over 60,000 generations. *Nature*. doi: 10.1038/nature24287

Grant PR, Grant BR. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. *Science* **296**:707–711. doi: 10.1126/science.1070315

Guirao-Rico S, González J. 2021. Benchmarking the performance of Pool-seq SNP callers using simulated and real sequencing data. *Mol Ecol Resour* **21**:1216–1229. doi:10.1111/1755-0998.13343

Günther T, Coop G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics* **195**:205–220. doi: 10.1534/genetics.113.152462

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet* **7**:e1001375. doi: 10.1371/journal.pgen.1001375

Hivert V, Leblois R, Petit EJ, Gautier M, Vitalis R. 2018. Measuring Genetic Differentiation from Pool-seq Data. *Genetics* **210**:315–330. doi: 10.1534/genetics.118.300900

Iranmehr A, Akbari A, Schlötterer C, Bafna V. 2017. CLEAR: Composition of Likelihoods for Evolve And Resequence Experiments. *Genetics*. doi: 10.1534/genetics.116.197566

Juenger T, Purugganan M, Mackay TF. 2000. Quantitative trait loci for floral morphology in Arabidopsis thaliana. *Genetics* **156**:1379–1392. doi: 10.1093/genetics/156.3.1379

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* **178**:1709–1723. doi:10.1534/genetics.107.080101

Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, Goubert C, Rota-Stabelli O, Kankare M, Bogaerts-Márquez M, Haudry A, Waidele L, Kozeretska I, Pasyukova EG, Loeschcke V, Pascual M, Vieira CP, Serga S, Montchamp-Moreau C, Abbott J, Gibert P, Porcelli D, Posnien N, Sánchez-Gracia A, Grath S, Sucena É, Bergland AO, Guerreiro MPG, Onder BS, Argyridou E, Guio L, Schou MF, Deplancke B, Vieira C, Ritchie MG, Zwaan BJ, Tauber E, Orengo DJ, Puerma E, Aguadé M, Schmidt P, Parsch J, Betancourt AJ, Flatt T, González J. 2020. Genomic Analysis of European Drosophila melanogaster Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Mol Biol Evol* **37**:2661–2678. doi:10.1093/molbev/msaa120

Kapun M, Nunez JCB, Bogaerts-Márquez M, Murga-Moreno J, Paris M, Outten J, Coronado-Zamora M, Tern C, Rota-Stabelli O, García Guerreiro MP, Casillas S, Orengo DJ, Puerma E, Kankare M, Ometto L, Loeschcke V, Onder BS, Abbott JK, Schaeffer SW, Rajpurohit S, Behrman EL, Schou MF, Merritt TJS, Lazzaro BP, Glaser-Schmitt A, Argyridou E, Staubach F, Wang Y, Tauber E, Serga SV, Fabian DK, Dyer KA, Wheat CW, Parsch J, Grath S, Veselinovic MS, Stamenkovic-Radak M, Jelic M, Buendía-Ruíz AJ, Josefa Gómez-Julián M, Luisa Espinosa-Jimenez M, Gallardo-Jiménez FD, Patenkovic A, Eric K, Tanaskovic M, Ullastres A, Guio L, Merenciano M, Guirao-Rico S, Horváth V, Obbard DJ, Pasyukova E, Alatortsev VE, Vieira CP, Vieira J, Roberto Torres J, Kozeretska I, Maistrenko OM, Montchamp-Moreau C, Mukha DV, Machado HE, Barbadilla A, Petrov D, Schmidt P, Gonzalez J, Flatt T, Bergland AO. 2021. Drosophila Evolution over Space and Time (DEST) - A New Population Genomics Resource. *bioRxiv*. doi:10.1101/2021.02.01.428994

Kessner D, Turner TL, Novembre J. 2013. Maximum Likelihood Estimation of Frequencies of Known Haplotypes from Pooled Sequence Data. *Molecular Biology and Evolution*. doi:10.1093/molbev/mst016

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in Arabidopsis thaliana. *Nat Genet* **39**:1151–1155. doi: 10.1038/ng2115

Kingsolver JG, Hoekstra HE, Hoekstra JM, Berrigan D, Vignieri SN, Hill CE, Hoang A, Gibert P, Beerli P. 2001. The strength of phenotypic selection in natural populations. *Am Nat* **157**:245–261. doi:10.1086/319193

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

Kofler R, Pandey RV, Schlötterer C. 2011b. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**:3435–3436. doi:10.1093/bioinformatics/btr589

Köster J, Rahmann S. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**:2520–2522. doi: 10.1093/bioinformatics/bts480

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**:D1202–10. doi: 10.1093/nar/gkr1090

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**:2987–2993. doi:10.1093/bioinformatics/btr509

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. doi: 10.1093/bioinformatics/btp352

Lovell JT, MacQueen AH, Mamidi S, Bonnette J, Jenkins J, Napier JD, Sreedasyam A, Healey A, Session A, Shu S, Barry K, Bonos S, Boston L, Daum C, Deshpande S, Ewing A, Grabowski PP, Haque T, Harrison M, Jiang J, Kudrna D, Lipzen A, Pendergast TH 4th, Plott C, Qi P, Saski CA, Shakirov EV, Sims D, Sharma M, Sharma R, Stewart A, Singan VR, Tang Y, Thibivillier S, Webber J, Weng X, Williams M, Wu GA, Yoshinaga Y, Zane M, Zhang L, Zhang J, Behrman KD, Boe AR, Fay PA, Fritschi FB, Jastrow JD, Lloyd-Reilley J, Martínez-Reyna JM, Matamala R, Mitchell RB, Rouquette FM Jr, Ronald P, Saha M, Tobias CM, Udvardi M, Wing RA, Wu Y, Bartley LE, Casler M, Devos KM, Lowry DB, Rokhsar DS, Grimwood J, Juenger TE, Schmutz J. 2021. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature*. doi:10.1038/s41586-020-03127-1

Lowry DB, Hall MC, Salt DE, Willis JH. 2009. Genetic and physiological basis of adaptive salt tolerance divergence between coastal and inland Mimulus guttatus. *New Phytol* **183**:776–788. doi:10.1111/j.1469-8137.2009.02901.x

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

Manzano-Piedras E, Marcer A, Alonso-Blanco C, Picó FX. 2014. Deciphering the adjustment between environment and life history in annuals: lessons from a geographically-explicit approach in Arabidopsis thaliana. *PLoS One* **9**:e87836. doi: 10.1371/journal.pone.0087836

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

Merilä J, Sheldon BC, Kruuk LE. 2001. Explaining stasis: microevolutionary studies in natural populations. *Genetica* **112-113**:199–222.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Others. 2021. Sustainable data analysis with Snakemake. F1000Res 10: 33.

Monnahan PJ, Colicchio J, Fishman L, Macdonald SJ, Kelly JK. 2020. Predicting evolutionary change at the DNA level in a natural Mimulus population. *PLOS Genetics*. doi: 10.1101/2020.06.23.166736

Nosil P, Villoutreix R, de Carvalho CF, Farkas TE, Soria-Carrasco V, Feder JL, Crespi BJ, Gompert Z. 2018. Natural selection and the predictability of evolution in Timema stick insects. *Science* **359**:765–770. doi: 10.1126/science.aap9125

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**:292–294. doi:10.1093/bioinformatics/btv566

Pfenninger M, Reuss F, Kiebler A, Schönnenbeck P, Caliendo C, Gerber S, Cocchiararo B, Reuter S, Blüthgen N, Mody K, Mishra B, Bálint M, Thines M, Feldmeyer B. 2021. Genomic basis for drought resistance in European beech forests threatened by climate change. *Elife* **10**:e65532. doi:10.7554/eLife.65532

Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, Agren J, Bossdorf O, Byers D, Donohue K, Dunning M, Holub EB, Hudson A, Le Corre V, Loudet O, Roux F, Warthmann N, Weigel D, Rivero L, Scholl R, Nordborg M, Bergelson J, Borevitz JO. 2010. The scale of population structure in Arabidopsis thaliana. *PLoS Genet* **6**:e1000843. doi:10.1371/journal.pgen.1000843

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

Robinson DO, Coate JE, Singh A, Hong L, Bush M, Doyle JJ, Roeder AHK. 2018. Ploidy and Size at Multiple Scales in the Arabidopsis Sepal. *Plant Cell* **30**:2308–2329. doi: 10.1105/tpc.18.00344

Roda F, Walter GM, Nipper R, Ortiz-Barrientos D. 2017. Genomic clustering of adaptive loci during parallel evolution of an Australian wildflower. *Mol Ecol* **26**:3687–3699. doi: 10.1111/mec.14150

Rowan BA, Patel V, Weigel D, Schneeberger K. 2015. Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3* **5**:385–398. doi: 10.1534/g3.114.016501

Rudman SM, Greenblum SI, Rajpurohit S, Betancourt NJ, Hanna J, Tilk S, Yokoyama T, Petrov DA, Schmidt P. 2021. Direct observation of adaptive tracking on ecological timescales in Drosophila. *bioRxiv*. doi: 10.1101/2021.04.27.441526

Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation. *Nat Rev Genet* **14**:807–820. doi: 10.1038/nrg3522

Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU. 2015. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity* **114**:431–440. doi: 10.1038/hdy.2014.86

Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals -- mining genome-wide polymorphism data without big funding. *Nat Rev Genet* **15**:749–763. doi:10.1038/nrg3803

Siepielski AM, Morrissey MB, Buoro M, Carlson SM, Caruso CM, Clegg SM, Coulson T, DiBattista J, Gotanda KM, Francis CD, Hereford J, Kingsolver JG, Augustine KE, Kruuk LEB, Martin RA, Sheldon BC, Sletvold N, Svensson EI, Wade MJ, MacColl ADC. 2017. Precipitation drives global variation in natural selection. *Science* **355**:959–962. doi:10.1126/science.aag2773

Suneson CA. 1956. An evolutionary plant breeding method 1. *Agron J* **48**:188–191. doi:10.2134/agronj1956.00021962004800040012x

Thurman TJ, Barrett RDH. 2016. The genetic consequences of selection in natural populations. *Mol Ecol* **25**:1429–1448. doi: 10.1111/mec.13559

Tilk S, Bergland A, Goodman A, Schmidt P, Petrov D, Greenblum S. 2019. Accurate Allele Frequencies from Ultra-low Coverage Pool-Seq Samples in Evolve-and-Resequence Experiments. *G3* **9**:4159–4168. doi: 10.1534/g3.119.400755

Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM. 2011. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS Genet* **7**:e1001336. doi: 10.1371/journal.pgen.1001336

Walsh B, Blows MW. 2009. Abundant Genetic Variation + Strong Selection = Multivariate Genetic Constraints: A Geometric View of Adaptation. *Annu Rev Ecol Evol Syst* **40**:41–59. doi:10.1146/annurev.ecolsys.110308.120232

GrENE
-net.org

Genomics of **r**apid **E**volution in **N**ovel **E**nvironments

A Coordinated Distributed Evolution Experiment with Arabidopsis thaliana

Supplemental Information Guide

Monitoring adaptation and demography of plant experimental populations with Pool-Sequencing

GrENE-net.org consortia authors

Supplemental Materials & Methods: Extended DNA preparation, sequencing methods, and computational analyses.

Supplemental Mathematical Appendix: Population genetic equations adapted for Pool-Seq.

Genomics of rapid Evolution in Novel Environments

A Coordinated Distributed Evolution Experiment with Arabidopsis thaliana

GrENE-net.org

GrENE-net.org consortia authors

Abdelaziz, Mohamed (1), Alexander, Jake (2), Bergelson, Joy (3), Bossdorf, Oliver (4), Cohen, Ofer (5), Colautti, Robert I. (6), Delker, Carolin (7), Dimitrakopoulos, Panayiotis G. (8), Durka, Walter (9), Escribano-Avila, Gema (10), Franks, Steven J. (11), Fritschi, Felix B. (12), Galanidis, Alexandros (13), Garcia-Fernández, Alfredo (14), Hamann, Elena (15), Hanna Nomoto (16), Iriondo, J.M. (17), Keller, Stephen (18), Korte, Arthur (19), Lara-Romero, Carlos (20), Maag, Daniel (21), March-Salas, Martí (22), Morente-López, Javier (23), Pärtel, Meelis (24), Quint, Marcel (25), Seifan Merav (26), Snoek, Basten L. (27), Stam, Remco (28), Sternberg, Marcelo (29), Stift, Marc (30), Stinchcombe, John R. (31), Till-Bottraud, Irène (32), Traveset, Anna (33), Valay, Jean-Gabriel (34), Van Zanten, Martijn (35), Violle, Cyrille (36), Wódkiewicz Maciej (37), Zuzana Münbergová (38)

(1) Department of Genetics, University of Granada, Granada, Spain.
(2) Institute of Integrative Biology, ETH Zurich, Zürich, Switzerland
(3) Department of Biology, NYU, NY, NY USA
(4) Plant Evolutionary Ecology, Institute of Evolution and Ecology, University of Tübingen, Germany
(5) School of Plant Sciences and Food Security, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel
(6) Biology Department, Queen's University, Kingston, ON Canada
(7) Crop Physiology, Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg
(8) Biodiversity Conservation Laboratory, Department of Environment, University of the Aegean, 81100 Mytilene, Lesbos, Greece
(9) Department of Community Ecology, Helmholtz Centre for Environmental Research-UFZ, 06120 Halle, Germany
(10) Arboriculture Department, Tecnigral SL, Madrid, Spain.
(11) Department of Biological Sciences and the Louis Calder Center, Fordham University, Bronx, New York, USA
(12) Division of Plant Science & Technology, University of Missouri, Columbia, MO, USA
(13) Biodiversity Conservation Laboratory, Department of Environment, University of the Aegean, 81100 Mytilene, Lesbos, Greece
(14) ECOEVO group, Rey Juan Carlos University, Móstoles, Spain
(15) Biological Sciences, Fordham University, Bronx, NY, USA
(16) Institute of Integrative Biology, ETH Zürich, Zürich, Switzerland
(17) ECOEVO research group, Biodiversity and Conservation Area, ESCET, Universidad Rey Juan Carlos, Móstoles, Spain
(18) Department of Plant Biology, University of Vermont, Burlington, USA
(19) Center for Computational and Theoretical Biology, University of Wuerzburg, Wuerzburg, Germany
(20) Biodiversity and Conservation Area, Rey Juan Carlos University, Mostoles, Spain
(21) Department of Pharmaceutical Biology, Julius-von-Sachs-Institute, Julius-Maximilians-Universität Würzburg, Würzburg, Germany
(22) Plant Evolutionary Ecology, Faculty of Biological Sciences, Goethe University Frankfurt, Max-von-Laue-Str. 13, 60438 Frankfurt am Main, Germany
(23) Área de Biodiversidad y Conservación, Departamento de Biología y Geología, Universidad Rey Juan Carlos-ESCET, Tulipán s/n. 28933 Móstoles, Madrid (Spain).
(24) Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
(25) Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany + German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany
(26) Mitrani Department of Desert Ecology, Swiss Institute for Dryland Environmental & Energy Research, Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Israel
(27) Theoretical Biology & Bioinformatics, Utrecht University, Padualaan 8, 3584 CH Utrecht the Netherlands
(28) Chair of Phytopathology, Technical University of Munich
(29) School of Plant Sciences and Food Security, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel.
(30) Ecology, Department of Biology, University of Konstanz
(31) Ecology & Evolutionary Biology, University of Toronto, Toronto, Canada
(32) Université Clermont Auvergne, CNRS, GEOLAB, 63000 Clermont_Ferrand, France
(33) Global change research group, Mediterranean Institute of Advanced Studies, Esporles, Mallorca, Balearic Islands, Spain
(34) Lautaret garden, Université Grenoble Alpes, CNRS, Grenoble, France
(35) Molecular Plant Physiology, Department of Environmental Biology, Utrecht University, Utrecht, The Netherlands
(36) CEFE, Univ Montpellier, CNRS, EPHE, IRD
(37) Faculty of Biology, Biological and Chemical Research Centre,University of Warsaw, Warsaw, Poland
(38) 1. Institute of Botany, Czech Academy of Sciences, Pr_honice, Czechia. 2. Department of Botany, Faculty of Science, Charles University, Prague, Czechia.

Supplemental Materials & Methods for:

Monitoring rpaid evolution of plant populations at scale with Pool-Sequencing

Plant growth protocol 3
Experiment 1 3
Experiment 2 3
Experiment 3 3
Experiment 4 3
DNA extraction 3
Experiments 1, 3, and 4 3
Experiment 2 4
Library preparation 5
Experiment 1 5
Experiment 2 5
Experiment 3 6
Experiment 4 6
References 7
Supplemental Datasets and Tables 8
Dataset S1 | Ecotype IDs used for outdoor experiment and GrENE-net seed mixture 8
Table S1 | GrENE-net founder seed mix DNA extraction replicates 9
Table S2 | Input DNA concentration library preparation for validation experiments. 10
Table S3 | Sampling of 50 leaves 11
Table S4 | Combinatorics of flower and leaf pooling 13
Table S5| Pool extraction of flower combinatorics 15
Table S6 | Sampling of pilot field experiment (Exp. 4) 16
Supplemental Figures 17
Fig. S1 | Cartoon of rationale of Pool-Seq 17
Fig. S2 | Hierarchical sampling of flowers for Pool-sequencing of different sizes 18
Fig. S3 | Allele frequencies from the 1001 Arabidopsis Genomes and from seed Pool-seq. 19
Fig. S4 | LD decay in the 1001G, the 470 and 231 and accessions sets 20
Fig. S5 | Coverage of the seed sequencing 21
Fig. S6 | Correlation between raw frequencies and SNP caller fields (all coverages) 22
Fig. S7 | Correlation between raw frequencies and SNP caller fields (50-100X) 23
Fig. S8 | Correlation between raw frequencies and SNP caller fields (100-250X) 24
Fig. S10 | Correlation between raw frequencies and SNP caller fields (250-500X) 25

Fig. S11 | The problem of SNP calling for Pool-Seq using diploid callers 26
Fig. S12 | Allele frequencies in the 2 equal mass DNA pool using different quality filters. 27
Fig. S13 | Example random Binomial draws and recovered allele frequencies 28
Fig. S14 | Fraction of DNA contribution to Pool-seq for a 2-flower pool and a 2-leaf pool 29
Fig. S15 | Fst distributions between all pairs of seed libraries. 30
Fig. S15 | Fst distributions of all replicates of E&R (Exp. 4). 31
Fig. S16 | Fst distributions of Time 0 and all replicates in E&R (Exp. 4). 32
Fig. S17 | Example genome-wide FST scan of E&R (Exp 4). 33
Fig. S18 Example genome-wide FST scan of E&R (Exp 4). 34
Fig. S19 | Example genome-wide FST scan of E&R (Exp 4) with 10Kb window averages. 35

Plant growth protocol

Experiment 1

The 231 ecotypes were bulked in growth chambers at 20°C under the long-day condition (16 hours light / 8 hours dark) in three locations. In the Max Planck Institute for Biology, Germany, the growth chambers at the University of Tübingen's Institute of Evolutionary Ecology, Germany, and the CNRS Centre for Functional and Evolutionary Ecology in Montpellier, France.

Experiment 2

The Col-0 and RUM-20 plants were grown in growth chambers at 22°C under the long-day condition.

Experiment 3

We took one tube containing 0.1 g of the founder seed mix (~5,000 seeds), bleach-sterilized it, washed it (20 min; 500ml solution, 10% bleach, 20% SDS) and submerged the seeds in 1% agar solution for 5 days at 4 °C in the dark. Seeds were planted in trays with soil (CL-P, Einheitserde Werkverband e.V., Sinntal-Altengronau Germany) in 25 trays, 1,000 pots, and germinants were thinned to one plant per pot. We watered abundantly, growing the plants at 16 °C for 13 days (long day conditions) before a 60 day vernalization at 4 °C (short day conditions). The vernalization approach aimed to avoid flowering time differences among our diverse genotypes. Subsequently, the trays were transferred to 20 °C under the long-day condition for flowering. Two weeks later, 50 pots were randomly chosen to sample leaves and flowers whenever the plants bloomed

Experiment 4

The genotypes planted are 451 natural accessions (**Dataset S1**), all mixed together, across three plots (about 2 seeds/cm^2 in 1m^2 plots). 20 siliques from two different parental individuals of all genotypes were pooled and sowed in three batches: one in November 2014, one in February 2015, one in March 2015. On March 2, 2016, before flowering, bulk soil was taken to germinate seeds in the growth chamber and collect flowers for sequencing to avoid any selection of genotypes. 50-101 flowers from the plants growing in the growth chamber were sampled to generate allele frequency data for time point 0 (**Table S5**). In the field, 50-100 flowers, 80-200 flowers, and 60-300 flowers were sampled on April 1 (time point 1), April 22 (time point 2), and May 6 (time point 3) respectively and used for sequencing.The numbers of flowers sampled per plot are outlined in **Table S5**.

DNA extraction

Experiments 1, 3, and 4

The GrENE-net founder seed mix, containing 231 natural accessions (**Dataset S1**, Exp 1), was aliquoted into eight replicates according to the tissue input amount recommended by the Qiagen DNeasy Plant Mini kit (Hilden, Germany) (**Table S1**). Seed aliquots were suspended in 0.1% agar and kept at 4°C in the dark for 9 to 11 days to initiate germination. Then, seed aliquots were centrifuged and the supernatant was removed. 0.5 mL of rock and 800 µL of lysis buffer AP1 from the DNeasy kit were added to the seed tubes. Tissue homogenization was carried out using the Quickprep adapter in a FastPrep-24 (MP Biomedicals, Irvine, CA, USA) with the following setting: 6.0 m/sec for 40 seconds. Each tube was homogenized for a total of 2 rounds. 8 µL of RNase (100 mg/mL) from the DNeasy kit was added to the seed homogenate. After a short vortex and a quick spin, the seed homogenate was incubated at 65°C for 10 minutes. After the incubation, 185 µL of buffer P3 from the DNeasy kit was added to the seed lysate. The tube was inverted and incubated on ice for 5 minutes. The rest of the extraction followed the standard Qiagen DNeasy Plant Mini protocol. DNA was eluted in 100 µL of AE buffer.

The leaf subsamples and flower subsamples (Exp. 3) were extracted similarly with the DNeasy Plant Mini kit (Qiagen, Hilden, Germany) with modifications to the grinding step. Tissue samples and 5 ceramic beads were placed in a screw-cap tube and froze with liquid nitrogen. The homogenization was again carried out using FastPrep-24 (MP Biomedicals, Irvine, CA, USA) with a different setting: 4.0 m/sec for 15 seconds. Each tube was homogenized for a total of 2 rounds. The rest of the extraction followed the standard DNeasy protocol.

The field experiment samples (Exp. 4) of pools of flowers (Table S6) were processed as previously (Exp. 3).

Experiment 2

Due to the high cost of commercial kits such as the Qiagen DNeasy Plant Mini kit ($4.46 per isolation at listed price), we set up a cheaper plate-based DNA extraction protocol based on the widely used 2x CTAB protocol (Doyle and Doyle, 1987). All GrENE-net DNA extracts were isolated using this custom protocol from pooled flower samples collected from the 45 field sites. We partially replicated the leaf and flower comparison (see Experiment 3 in the main text) using our CTAB/chloroform protocol. In the case of flowers, two flowers of similar size were collected in the same tube prior to DNA extraction (n = 3). In the case of leaves, a leaf was independently extracted, but leaf extracts from two distinct ecotypes, Col-0 and RUM-20, were combined at similar DNA mass prior to library preparation (n = 3) (see Experiment 2 in the main text).

2-mercaptoethanol was added to the 2x CTAB buffer (1.4 M NaCl, 100 mM Tris pH 8.0, 20 mM EDTA pH 8.0, 2% w/v CTAB, 1% w/v PVP, ddH$_2$O) to a final concentration (v/v) of 0.3%. The buffer was warmed at 65°C for at least 30 minutes. Using a TissueLyser II (Qiagen, Hilden, Germany), frozen plant tissues were pulverized with 3.2 mm steel beads in 2.0 mL tubes on chilled adapter sets 2 x 24. Homogenization was carried out at 22/s for 35 sec and repeated until the frozen tissues attained the appearance of greenish white powders. 500 µL of pre-warmed 2x CTAB buffer was added to each tube to thoroughly resuspend the pulverized tissue. Samples were incubated at 65°C for 50 minutes and inverted every 10 to 15 minutes to resuspend the precipitates. After incubation, the lysate was transferred to a new 2.0 mL tube. When the lysate was cooled to room temperature, 500 µL of chloroform:isoamyl alcohol (24:1) was added to the lysate. The tube was vigorously shaken until the lysate and chloroform appeared well-mixed. The sample was centrifuged at 20,000 rcf for 14 minutes or until the upper aqueous layer appeared clear. 300 µL of the aqueous layer was transferred to a new tube or a 96-well deep well plate if doing high-throughput processing. 225 µL (0.75 vol) of isopropanol was added to the supernatant and mixed well by pipetting. The sample was incubated at 4°C for at least 30 minutes or at -20°C overnight. After incubation, the sample was centrifuged at max speed for 15 minutes in a tube. Alternatively, the 96-well plate was centrifuged at 6,100 rcf for 45 minutes. After discarding the supernatant, freshly prepared 70% ethanol was added to wash the DNA pellet. The sample was centrifuged at max speed for 5 minutes in a tube. Alternatively, the 96-well plate was centrifuged at 6,100 rcf for 30 minutes. The ethanol was removed and the pellet was left to air dry for 10 to 15 minutes. The DNA pellet was eluted in

Tris buffer containing RNase A (10 mM Tris-HCl pH9.0, ddH_2O, 20 µg/mL RNase A). The eluate was incubated at 37°C for 30 minutes. After a pulse spin, the DNA extract was stored at -20°C.

Library preparation

Experiment 1

Because the total number of samples was small and we had large amounts of DNA (**Table S1**), we conducted library preparations with Illumina's TruSeq PCR free library kit (Ilumina, San Diego, California).

Experiment 2

Because there were only 11 samples, tube-based quantification was performed using the Qubit dsDNA HS assay. The readings for the input DNA concentration are documented in Table S2. The library preparation protocol was based on (Baym et al., 2015) with some modifications. 2 µL of DNA sample was mixed with 2.75 µL TD buffer (Tagment DNA Buffer) and 0.25 µL TD enzyme (Mira Loma, California, USA). The tagmentation reaction mixture was mixed well by gentle pipetting. After a flash spin, the sample was incubated at 55°C for 10 minutes and held at 10°C.

Once equilibrated to room temperature or lower, the samples were flash spinned. Then, the tagmented DNA was mixed with 8 µL 2x KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Boston, MA, USA), 1.5 µL 10 µM P5 indexing primers (final concentration 0.75 µM), 1.5 µL 10 µM P7 indexing primers (equimolar to P5), and 4 µL Tris-Cl buffer (pH 8.0). The PCR reaction mixture was mixed well by gentle pipetting and the liquid was spinned down. The DNA was amplified using the following thermal cycling program:

1. 72°C for 3 minutes
2. 95°C for 3 minutes
3. 98°C for 20 seconds
4. 63°C for 30 seconds
5. 72°C for 30 seconds
6. Repeat from step 3 for 11 additional cycles (i.e. a total of 12 cycles)
7. 72°C for 5 minutes
8. 10°C hold

For post-amplification cleanup and size selection, the 11 libraries were multiplexed in a 1.5 mL tube by mixing 10 µL of each library. The library volume was estimated by aspirating with a P200 pipette (V_{lib}). 0.45 volume (i.e. 0.45 x V_{lib}) of homemade SPRI beads was added to the 11-plex library. The tube was incubated for 5 minutes on a regular rack and then incubated for 5 minutes on a magnet stand until a bead pellet forms. This bead pellet represents the first elution fraction. The supernatant excluding the first pellet was transferred to a new 1.5 mL tube on a regular rack. 0.6 volume (i.e. [0.6 - 0.45] x V_{lib}) of homemade SPRI beads was added to the supernatant. The tube was incubated for 5 minutes on a regular rack and then incubated for 5 minutes on a magnet stand until a bead pellet forms. This bead pellet represents the second elution fraction. The supernatant excluding the second pellet was removed. Each magnetic bead pellet was washed by gently adding 700 µL 70% ethanol and was incubated for 30 seconds before removing the ethanol. The ethanol wash was repeated once. The bead pellet was air dried until they lost the shine and began showing tiny cracks.

The tube containing the bead pellet was taken off the magnet and resuspended in 36 µL of AE buffer (10 mM Tris-Cl, 0.5 mM EDTA; pH 9.0). After incubating on a regular rack for 3 minutes, the tube was put on magnet stands for 5 minutes until the bead pellet formed. 34 µL of the eluate fraction was transferred to a new 1.5 mL tube. Both eluted fractions were quantified with Qubit and analyzed on a TapeStation 4150 (Memphis, Tennessee, USA) using a D1000 ScreenTape (Cedar Creek, Texas, USA).The second fraction was sequenced on a HiSeq 2 x 150 lane (**Fig. S6**).

Experiment 3

To compare seeds with flowers and leaf extracts without library preparation differences, we conducted library preparations with Illumina's TruSeq PCR library kit (Ilumina, San Diego, California) as in Experiment 1.

Experiment 4

The library preparation procedure was similar to what was described in Rowan et al. 2019 *Genetics* with minor volume adjustments. Specifically, the twelve amplified libraries were multiplexed together by mixing 5 µL of each library. The total volume was brought up to 100 µL with 10 mM Tris-Cl (pH 8.5). The rest of the size selection was performed as written in Rowan et al. (2019). The fragment length distribution of bead fraction 3 was verified with a Bioanalyzer before being sent for sequencing on an Illumina HiSeq 3000.

References

1001 Genomes Consortium. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**:481–491.

Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. 2015. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**:e0128036.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue (No. RESEARCH). worldveg.tind.io.

Supplemental Datasets and Tables

Dataset S1 | Ecotype IDs used for outdoor experiment and GrENE-net seed mixture

Metadata of the 451 and 231 ecotypes lists.

<google drive link>

Dataset S2 | Genes within 10Kb regions with F_{ST} > 0.2 in 3+ E&R replicates

TAIR summary of gene annotation.

<google drive link>

Table S1 | GrENE-net founder seed mix DNA extraction replicates

Sample identifier	Tissue amount (mg)	Approximate number of individuals	DNA concentration (ng/µL)
GrENE-net 231 founder seed mix #1	100	5000	7.98
#2	100	5000	6.48
#3	100	5000	8.56
#4	17.4	870	21.8
#5	18	900	24.8
#6	18.3	915	23.6
#7	20	1000	24.4
#8	21.6	1080	25

Table S2 | Input DNA concentration library preparation for validation experiments.

Sample identifier	Input concentration (ng/µL)
Col-0 leaf extract	2.17
RUM-20 leaf extract	2.04
Pooled flower extract #1	1.84
Pooled flower extract #3	1.97
Pooled flower extract #5	1.98
Col-RUM leaf extract pool #1	2 (predicted, 2 µL of 8.33 ng/µL diluted in 6.33 µL Tris buffer)
Col-RUM leaf extract pool #2	2 (predicted, 2 µL of 8.53 ng/µL diluted in 6.53 µL Tris buffer)
Col-RUM leaf extract pool #3	2 (predicted, 2 µL of 8.71 ng/µL diluted in 6.71 µL Tris buffer)

*due to the flexibility (i.e. ≤3 ng/µL) in the acceptable input range for this protocol, the three leaf extract pools were not quantified after dilution

Table S3 | Sampling of 50 leaves

sample id	tray	pos	DNA concentration (ng/µl)	Total DNA ng
1	1	a2	12.4	620
2	1	c3	23.2	1160
3	1	c7	23.2	1160
4	1	b7	13.5	675
5	1	b8	11.8	590
6	2	b2	17.3	865
7	2	c2	10.7	535
8	2	a4	17.4	870
9	2	a5	12.6	630
10	2	e7	14	700
11	3	c2	10.4	520
12	3	e3	7.54	377
13	3	a4	10.4	520
14	3	b7	18.9	945
15	3	c8	8.3	415
16	4	e2	12.1	605
17	4	d3	19.3	965
18	4	a5	17	850
19	4	e7	6.38	319
20	4	e6	7.06	353
21	4	b8	21.2	1060
22	5	e2	14.4	720
23	5	c3	21.4	1070
24	5	a5	19.1	955
25	5	e6	12.8	640
26	5	e8	10.2	510
27	6	b2	7.24	362
28	6	e3	15.5	775
29	6	d2	12.3	615
30	6	a5	7.3	365
31	6	e4	9.82	491
32	6	a6	8.2	410
33	6	e5	12	600
34	6	c5	9.9	495
35	6	e6	7.48	374
36	6	b7	12.3	615
37	7	d4	9.42	471
38	7	e2	9.42	471

39	7	b4	13.5	675
40	7	c3	7.84	392
41	7	d7	7.2	360
42	7	c7	11.7	585
43	8	e1	10.5	525
44	8	e2	7.22	361
45	8	d4	9.96	498
46	8	a4	16.1	805
47	8	d5	11.6	580
48	8	c6	7.02	351
49	8	b7	9.86	493
50	8	a8	15.7	785

Table S4 | Combinatorics of flower and leaf pooling

The 50 randomly selected plants were sampled for the 50 and 100 samples as well as for nested samples of smaller sets of plants. A graphical scheme of this sampling is in **Fig. S3**.

sample id	tray	pos	x100	x50a	x50b	x25a	x25b	x10b1	x10b2	x5
1	1	a2	yes	yes	yes		yes			yes
2	1	c3	yes	yes	yes		yes	yes		
3	1	c7	yes	yes	yes	yes				
4	1	b7	yes	yes	yes		yes		yes	
5	1	b8	yes	yes	yes	yes				
6	2	b2	yes	yes	yes		yes	yes		
7	2	c2	yes	yes	yes		yes	yes		
8	2	a4	yes	yes	yes		yes	yes		
9	2	a5	yes	yes	yes	yes				
10	2	e7	yes	yes	yes		yes		yes	
11	3	c2	yes	yes	yes		yes	yes		
12	3	e3	yes	yes	yes		yes			yes
13	3	a4	yes	yes	yes		yes	yes		
14	3	b7	yes	yes	yes	yes				
15	3	c8	yes	yes	yes		yes	yes		
16	4	e2	yes	yes	yes	yes				
17	4	d3	yes	yes	yes	yes				
18	4	a5	yes	yes	yes		yes		yes	
19	4	e7	yes	yes	yes	yes				
20	4	e6	yes	yes	yes		yes		yes	
21	4	b8	yes	yes	yes	yes				
22	5	e2	yes	yes	yes		yes	yes		
23	5	c3	yes	yes	yes	yes				
24	5	a5	yes	yes	yes		yes		yes	
25	5	e6	yes	yes	yes	yes				
26	5	e8	yes	yes	yes	yes				
27	6	b2	yes	yes	yes	yes				
28	6	e3	yes	yes	yes	yes				
29	6	d2	yes	yes	yes		yes			yes
30	6	a5	yes	yes	yes	yes				
31	6	e4	yes	yes	yes	yes				
32	6	a6	yes	yes	yes	yes				
33	6	e5	yes	yes	yes		yes			yes
34	6	c5	yes	yes	yes	yes				
35	6	e6	yes	yes	yes	yes				
36	6	b7	yes	yes	yes	yes				

37	7	d4	yes	yes	yes	yes				
38	7	e2	yes	yes	yes		yes		yes	
39	7	b4	yes	yes	yes		yes	yes		
40	7	c3	yes	yes	yes		yes		yes	
41	7	d7	yes	yes	yes		yes		yes	
42	7	c7	yes	yes	yes	yes				
43	8	e1	yes	yes	yes		yes		yes	
44	8	e2	yes	yes	yes	yes				
45	8	d4	yes	yes	yes		yes		yes	
46	8	a4	yes	yes	yes	yes				
47	8	d5	yes	yes	yes	yes				
48	8	c6	yes	yes	yes	yes				
49	8	b7	yes	yes	yes		yes	yes		
50	8	a8	yes	yes	yes		yes			yes

Table S5| Pool extraction of flower combinatorics

Tube ID	DNA concentration (ng/ul)
100 1	20.8
100 2	20.8
50 A	31.8
50 B	33.2
25 A	23.8
25 B	19.5
10 B2	5.48
5 B	4.68

Table S6 | Sampling of pilot field experiment (Exp. 4)

A total of 12 samples were sequenced either from seed banks or flowers of surviving plants in the outdoor field. Sequencing metrics are provided for each sample and the final working in silico pool.

						Input DNA			coverag	pool
S	**Time**	**rep**	**ID**	**# flowers**	**origin**	**(ng/µL)**	**read count**	**bp**	**e**	**coverage**
1	0	1	1_0	56	Seed bank	0.376	40214556	6032183400	48.3	199.6
2	0	2	2_0	69	Seed bank	0.366	62882228	9432334200	75.5	
3	0	3	3_0	101	Seed bank	0.424	63206340	9480951000	75.8	
4	1	1	1_1	80	Flowers from field	0.404	51991084	7798662600	62.4	219.3
5	1	2	2_1	160	Flowers from field	0.474	65024068	9753610200	78	
6	1	3	3_1	200	Flowers from field	0.454	65731096	9859664400	78.9	
7	2	1	1_2	65	Flowers from field	0.374	71196816	1.068E+10	85.4	206.3
8	2	2	2_2	205	Flowers from field	0.452	47974450	7196167500	57.6	
9	2	3	3_2	296	Flowers from field	0.33	52761222	7914183300	63.3	
10	3	1	1_3	19	Flowers from field	0.25	60332690	9049903500	72.4	206.3
11	3	2	2_3	50	Flowers from field	0.434	102265250	15339787500	122.7	
12	3	3	3_3	97	Flowers from field	0.452	64052368	9607855200	76.9	

Supplemental Figures

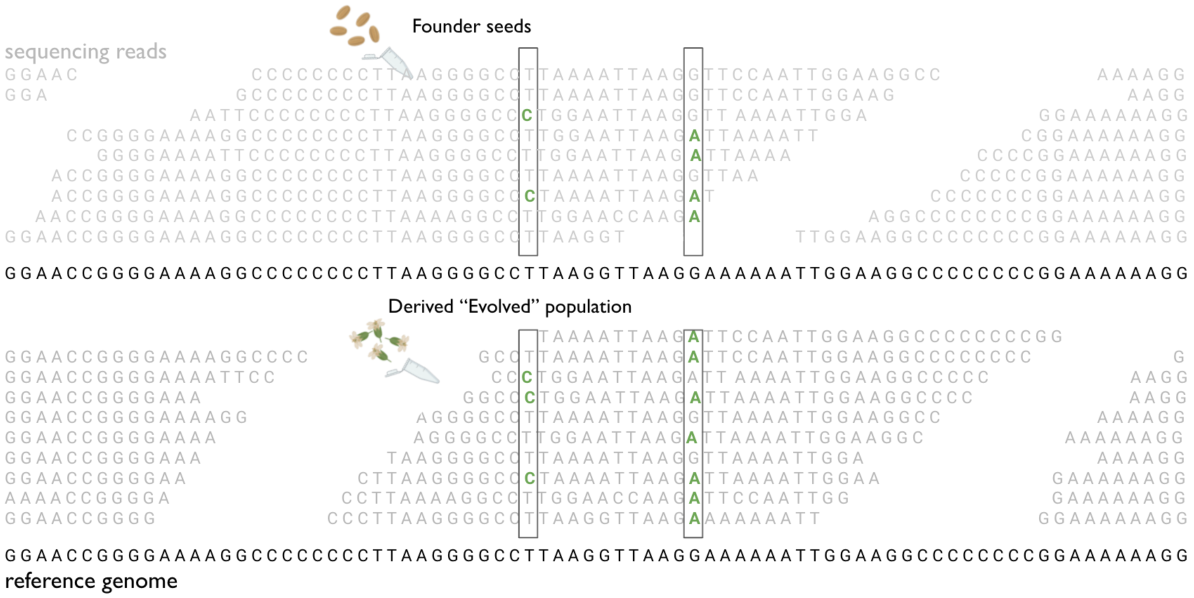
Founder seeds
sequencing reads
GGAACCGGGGAAAAGGCCCCCCCCTTAAGGGGCCTTAAGGTTAAGGAAAAAATTGGAAGGCCCCCCCCGGAAAAAAGG
Derived "Evolved" population
GGAACCGGGGAAAAGGCCCCCCCCTTAAGGGGCCTTAAGGTTAAGGAAAAAATTGGAAGGCCCCCCCCGGAAAAAAGG
reference genome

Fig. S1 | Cartoon of rationale of Pool-Seq

Reads from Illumina sequencing (typically ~150 bp, not at scale) "piled" against the region of the genome where they map to. The rationale is that if founder allele frequencies, or a reference sample that did not experience natural selection, we can extract meaningful evolutionary insights from comparing those with "evolved" populations, i.e. those that have grown in outdoor environments.

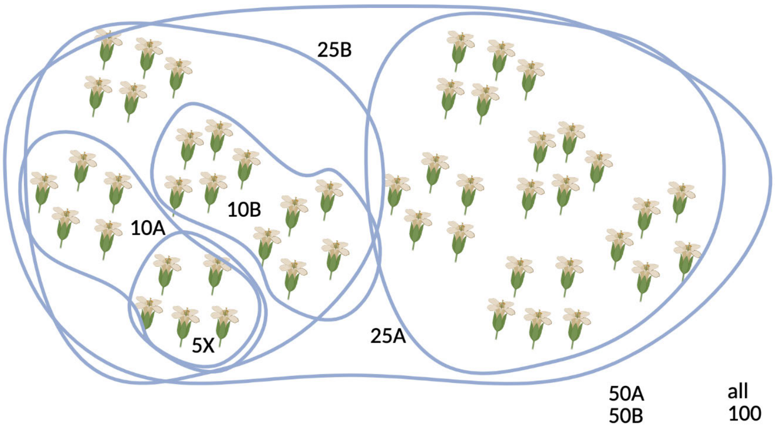

Fig. S2 | Hierarchical sampling of flowers for Pool-sequencing of different sizes
The selection of 50 individuals of Experiment 3 was conducted randomly from ~2,500 plants, but smaller sets of individuals were conducted in a nested fashion (e.g. the 25B samples were the same individuals as the 10A, 10B, and 5X sample. This may enable downstream allele frequency comparisons).

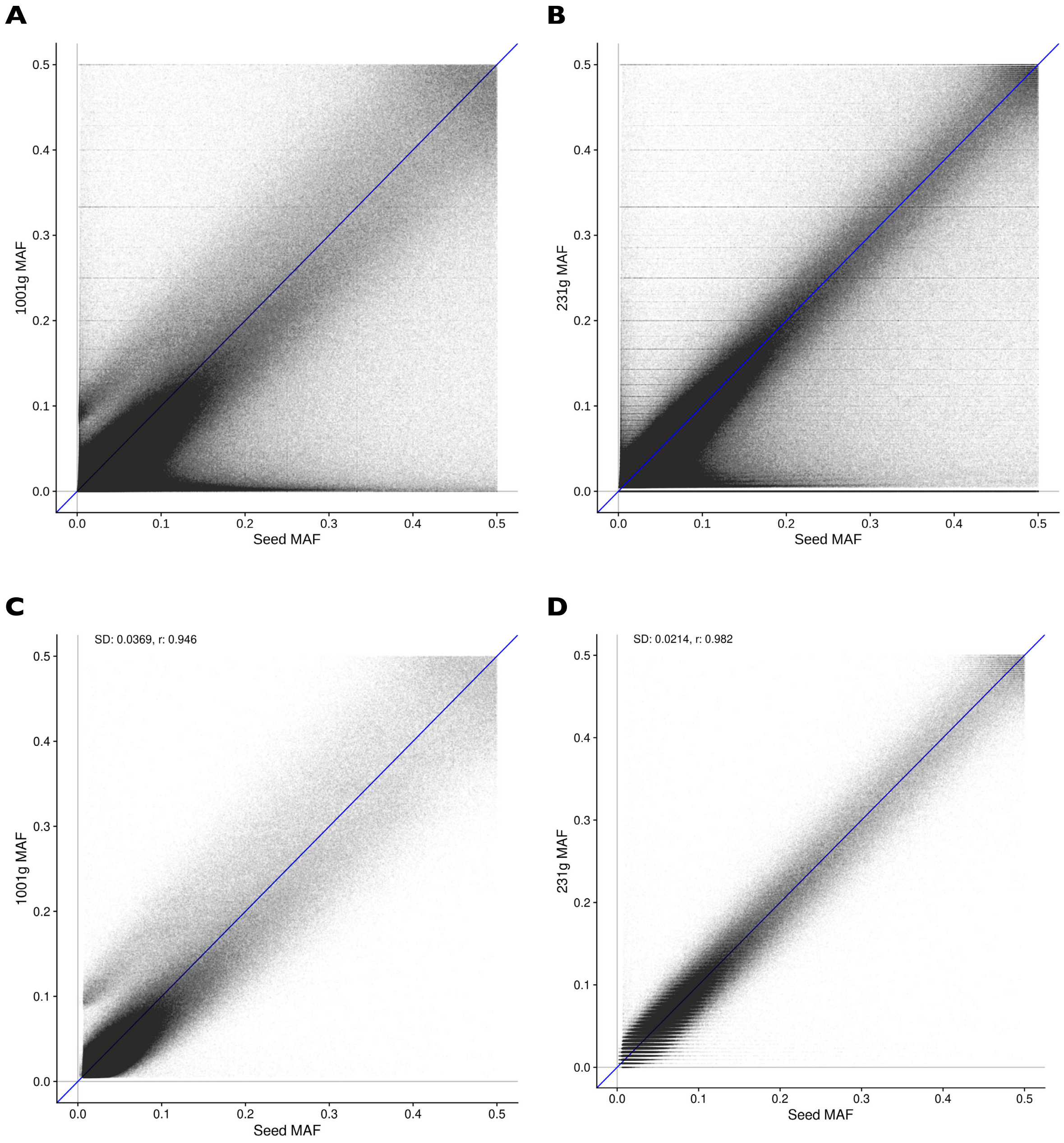

Fig. S3 | Allele frequencies from the 1001 Arabidopsis Genomes and from seed Pool-seq.

The x-axis is the folded seed founder frequency used to source outdoor experiments of GrENE-net.org, based on counting nucleotides at each locus in a bam/pileup file of the mapped reads (we converted bam to pileup for easier file parsing; same in all bam-based plots below). The y-axis is the folded frequency characterized from (**A**) the 1001 Genomes VCF and (**B**) the GrENE-net founder VCF. (**C**) and (**D**) are the same comparisons as (A) and (B) but for only *bona fide* of 1,353,386 biallelic SNPs from the 515g subset of the 1001 Genomes.

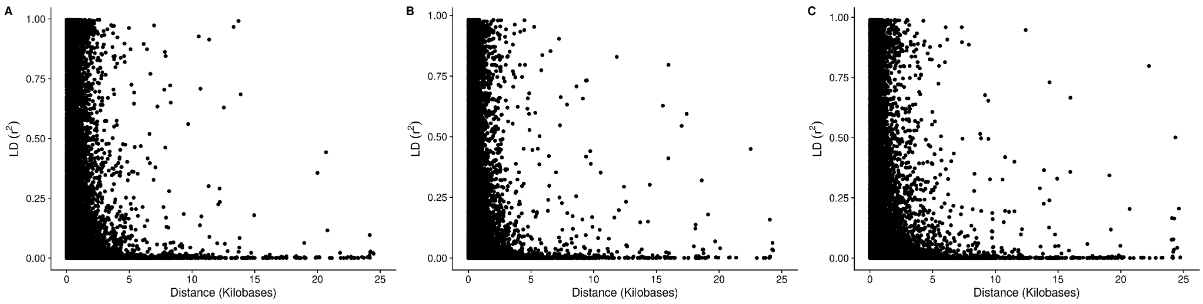

Fig. S4 | LD decay in the 1001G, the 470 and 231 and accessions sets

Linkage Disequilibrium LD decay using r^2 for the genome collection of (**A**) the 1001 Genomes Project, (**B**) a subset of 231 used in Experiment 1 and 3, and (**C**) a subset of 451 of the 1001 Genomes used in Experiment 4.

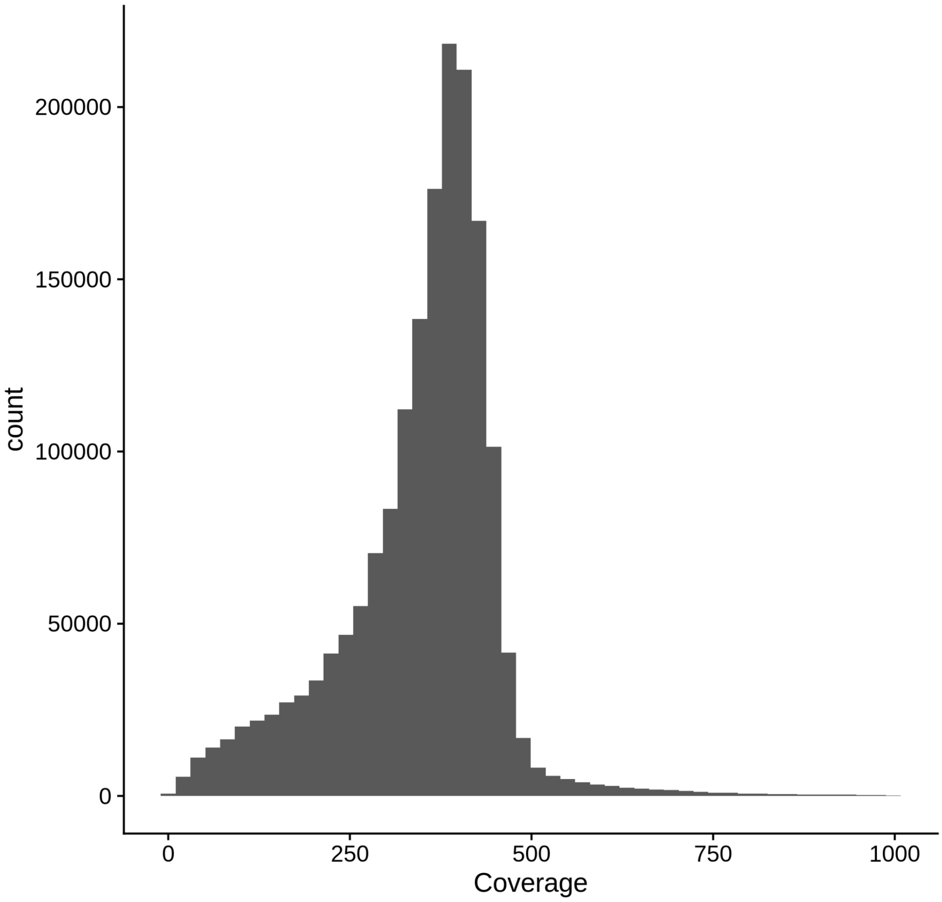

Fig. S5 | Coverage of the seed sequencing
Example of the coverage of the seeds as a *in silico* merge of 8 library samples (see **Table S1**) used in Fig. S3.

coverage-all

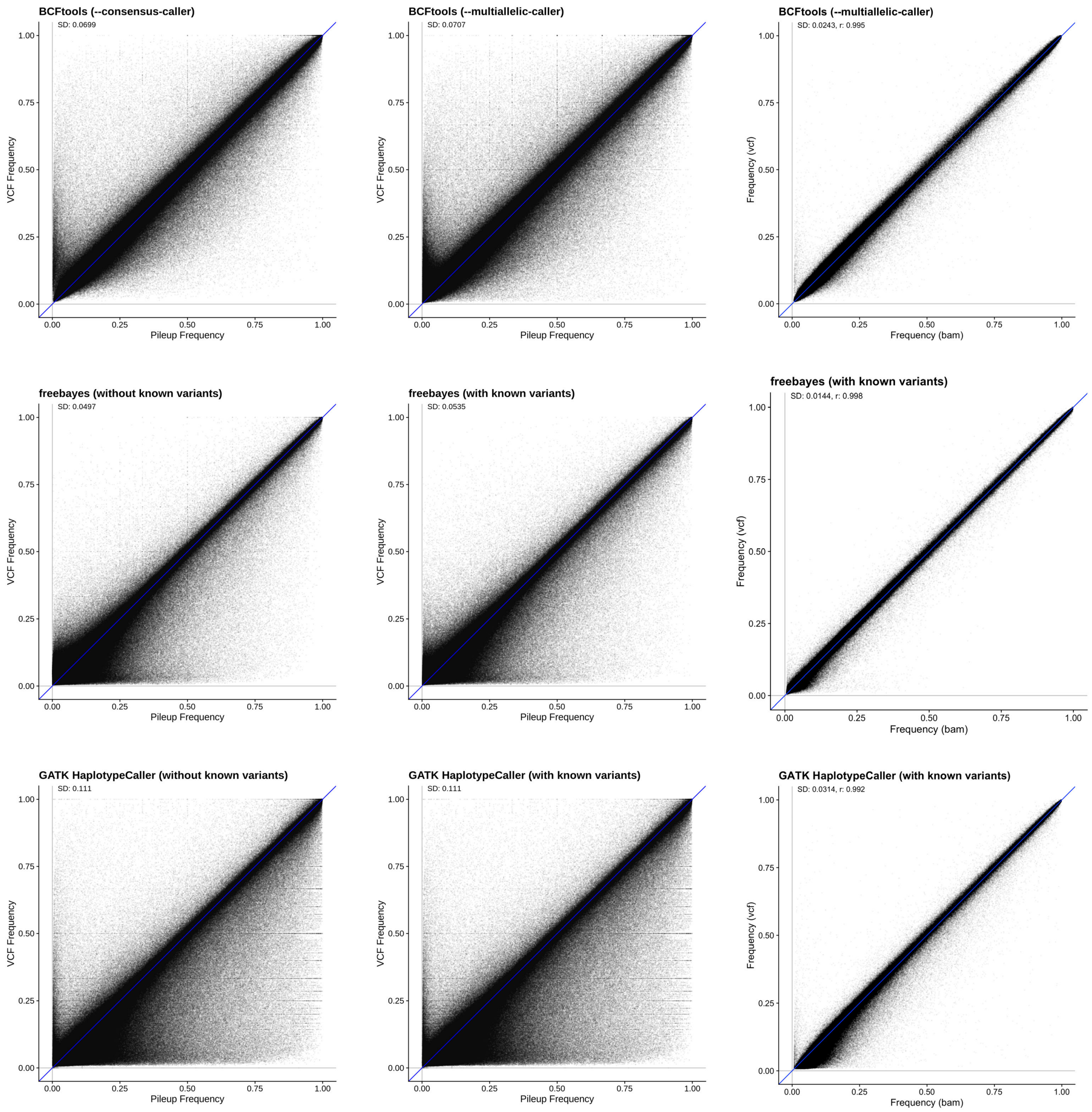

Fig. S6 | Correlation between raw frequencies and SNP caller fields (all coverages)
The x-axis shows the allele frequency of a biallelic SNP based on bam/pileup format where raw ratios of alternative and reference bases are computed, and is shared across all comparisons. The y-axis shows the allele frequency of the same biallelic SNPs as inferred from the allelic depth ("AD") VCF field from SNP calling outputs. Deviations from the y=x axis likely reveal artifacts created by SNP calling softwares. Each row presents three typical callers: BCFtools, freebayes, and GATK. Each column represents different calling mode or filterings: 'free' or discovery SNP calling, guided SNP calling at known variable positions from the 1001 Genomes, and subset of SNPs to *bona fide* of 1,353,386 biallelic SNPs from the 515g subset of the 1001 Genomes.

Coverage 50-100X

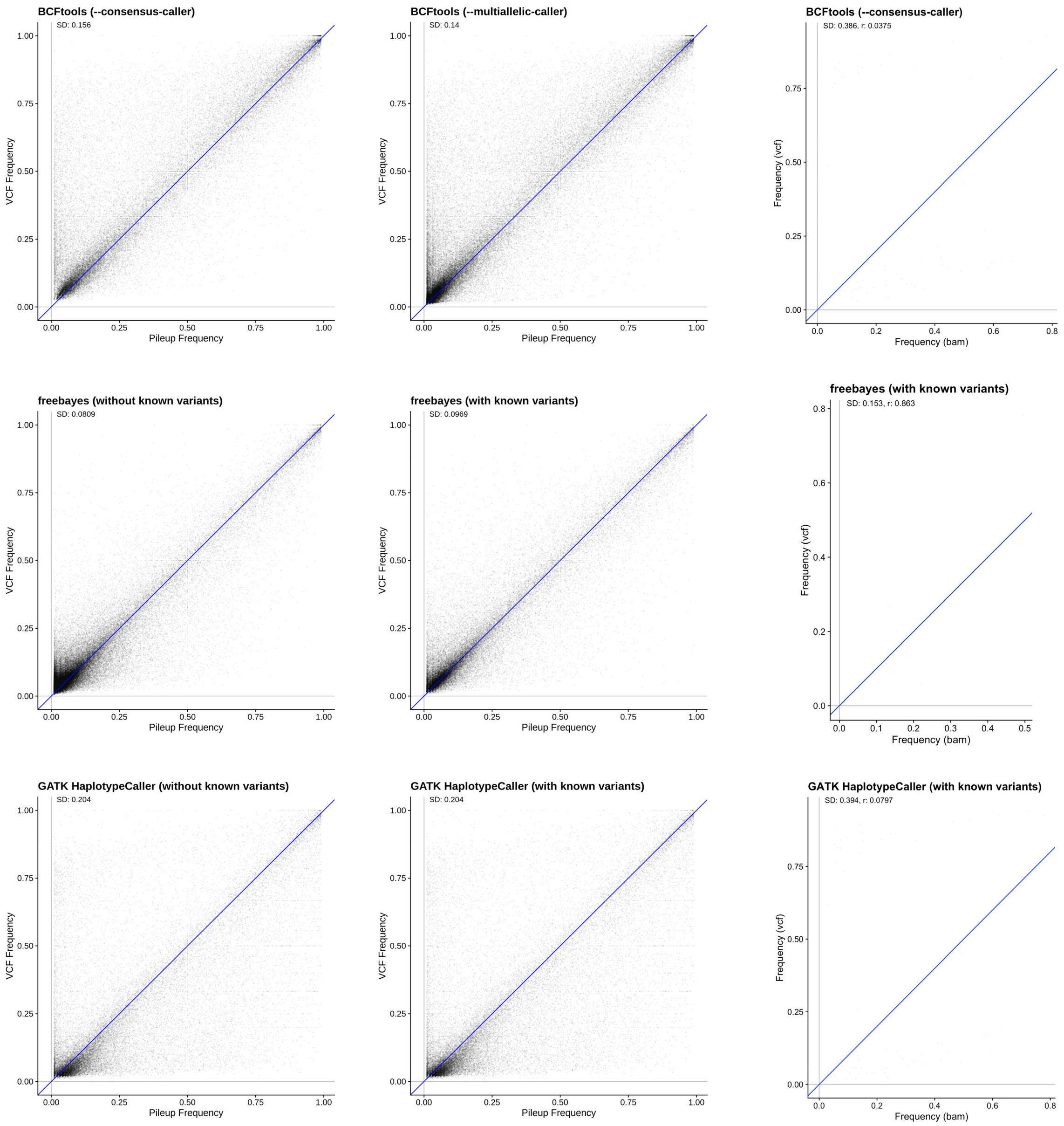

Fig. S7 | Correlation between raw frequencies and SNP caller fields (50-100X)

Same as Fig. S5 for coverage 50-100X.

Coverage-100-250

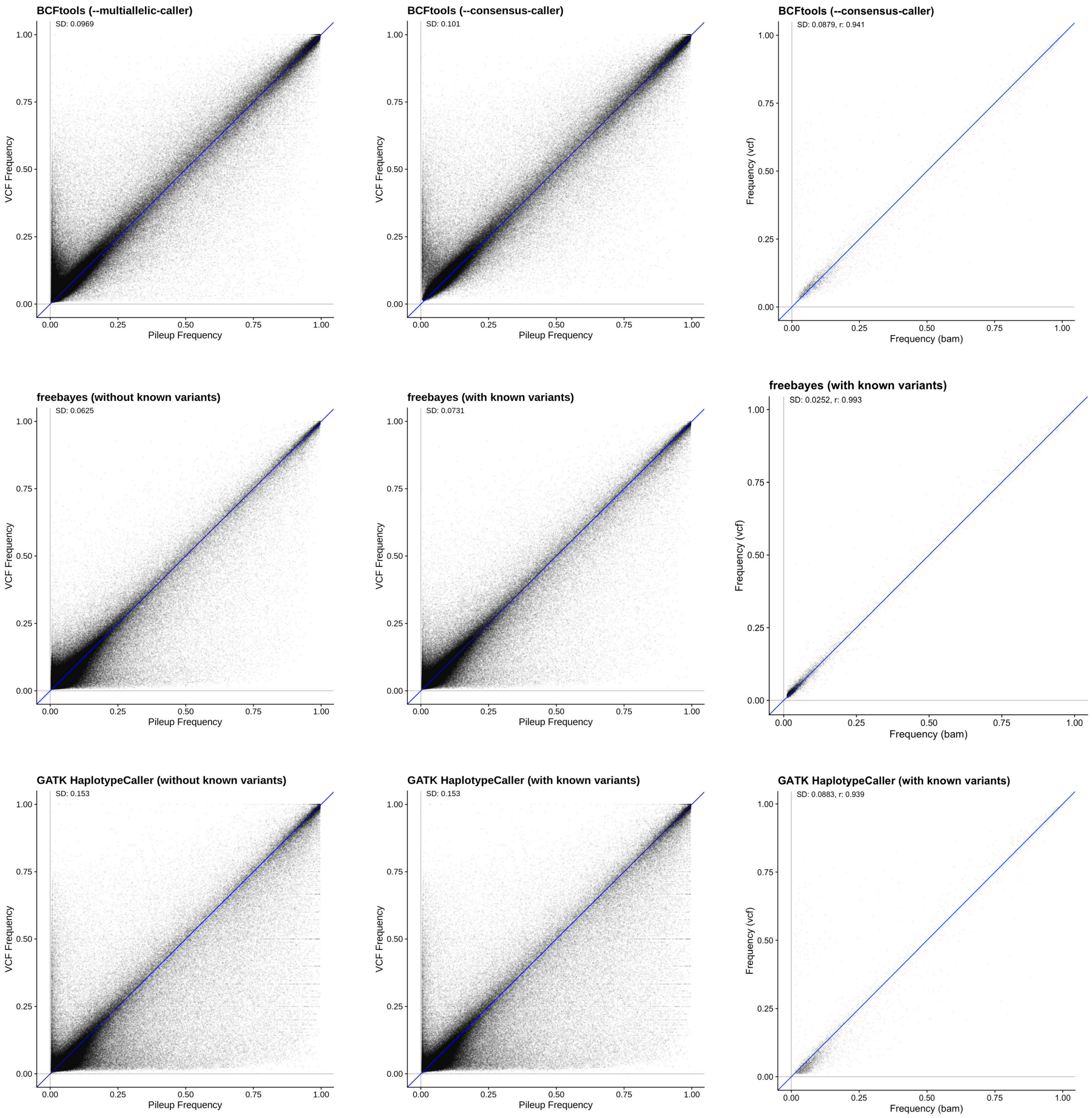

Fig. S8 | Correlation between raw frequencies and SNP caller fields (100-250X)

Same as Fig. S5 for coverage 100-250X.

Coverage-250-500

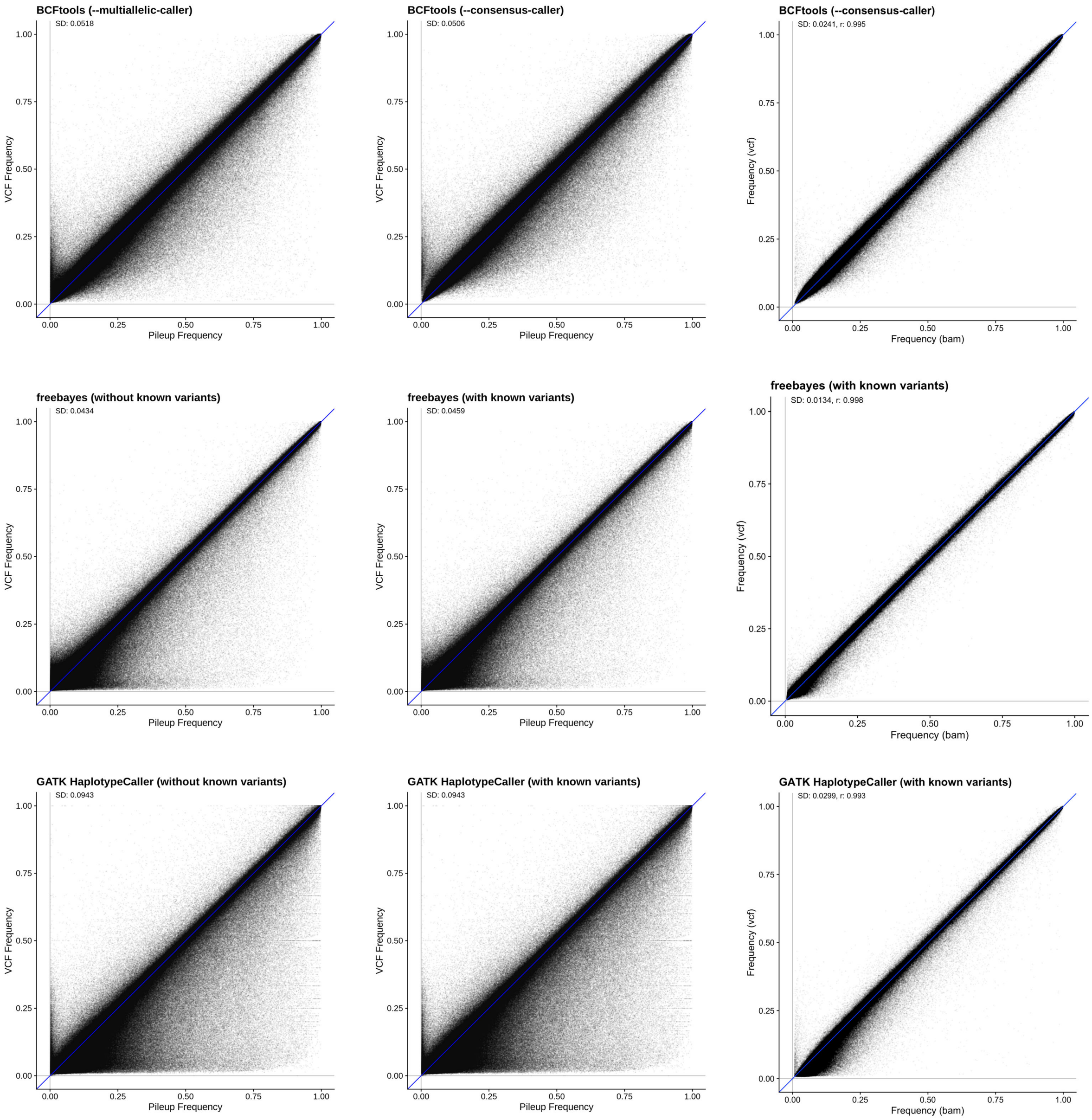

Fig. S10 | Correlation between raw frequencies and SNP caller fields (250-500X)

Same as Fig. S5 for coverage 250-500X.

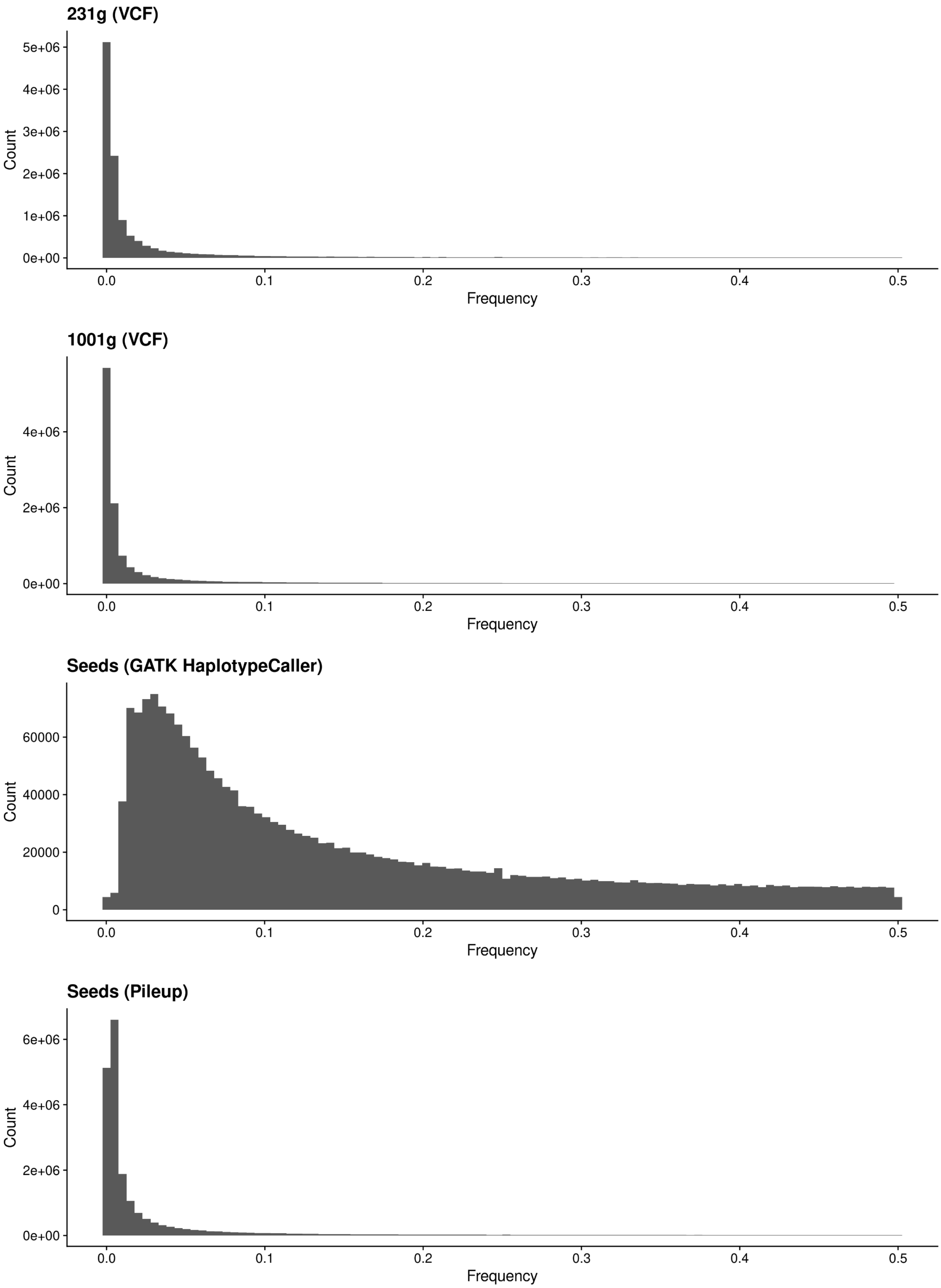

Fig. S11 | The problem of SNP calling for Pool-Seq using diploid callers

The figures show that the frequency of variants in the 1001 genomes or the 231 genomes subset is negative exponential, as expected from the Site Frequency Spectrum. The frequency calling of seeds, with high coverage, appears to be biased for intermediate frequencies in GATK HaplotypeCaller but retains the exponential decay using ratios of bases based on pileup using g$_r$enedalf in the last panel.

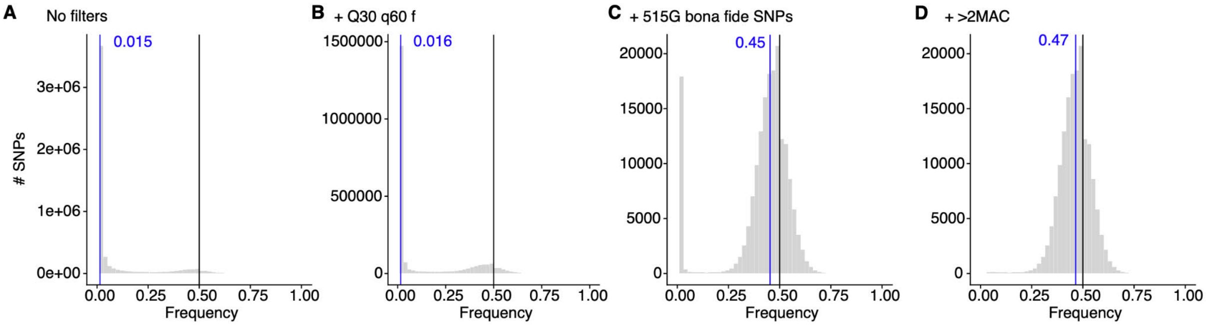

Fig. S12 | Allele frequencies in the 2 equal mass DNA pool using different quality filters. Histograms of allele frequencies from a pooled library of 2 distinct ecotypes and the expectation of 50% (black line). (**A**) Allele frequencies without any filter show the great majority of alleles must be artifacts, as there is a high point mass close to 0 frequency. (**B**) Reduction of likely artifacts, yet still high noise, using quality filters of bases with PHRED score above 30 (Q30) and from reads with mapping quality over 60 (q60) and for reads where forward and reverse map to the same region (f). (**C**) Subsetting allele frequencies to only those SNP found in the 1001 Genomes (1001 Genomes Consortium, 2016) with the highest quality 1.3 Million from (Exposito-Alonso et al. 2019) mostly removes all the noise signal with the exception of some rare variants. (**D**) Final removal of SNPs with only 1 or 2 bases supporting the alternative allele (minimum allele count >2) finally leaves a clean Binomial distribution of allele frequencies (owed to limited coverage) around the expected 50%.

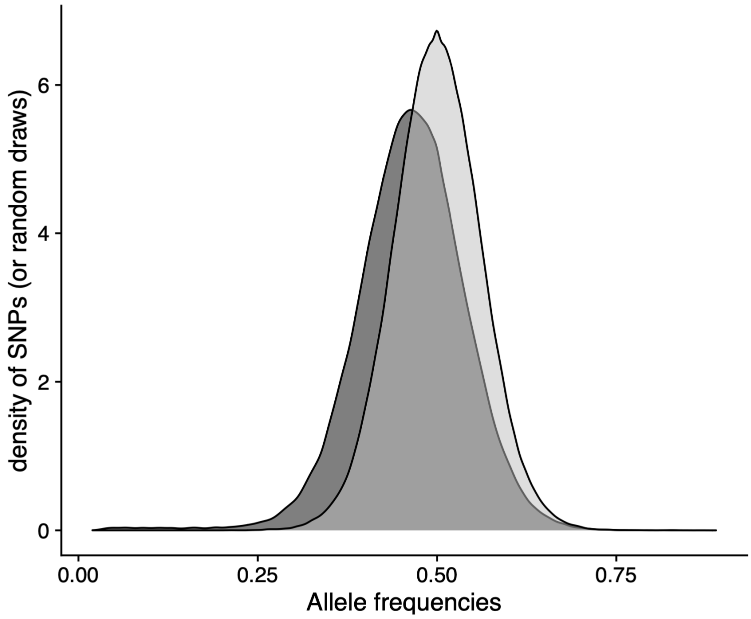

Fig. S13 | Example random Binomial draws and recovered allele frequencies

The distribution from **Fig. S12D** is compared to a random Binomial distribution with expected average frequency 50% (gray) and the same coverage distribution as the empirical sample.

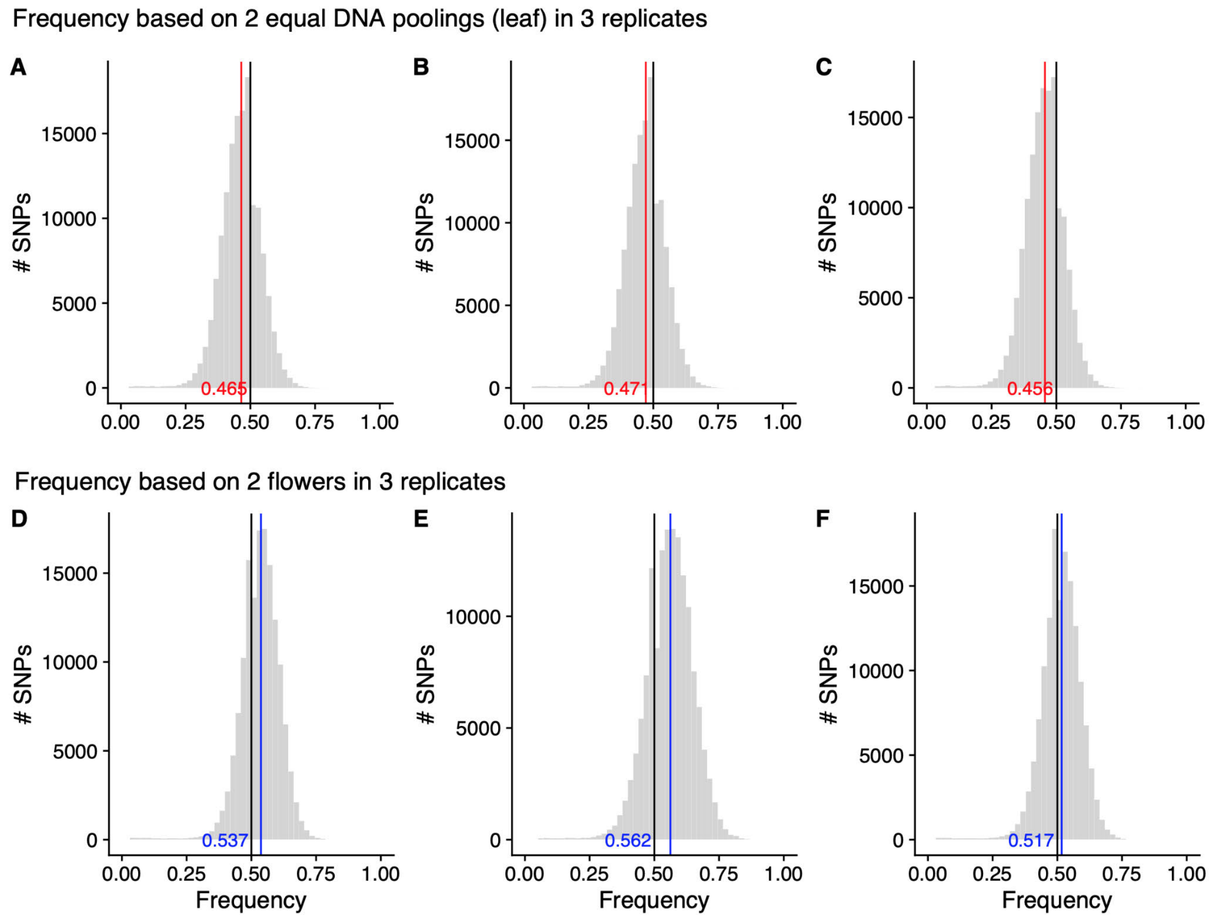

Fig. S14 | Fraction of DNA contribution to Pool-seq for a 2-flower pool and a 2-leaf pool
Dispersal of allele frequencies from the expected 50% (black vertical lines) for a pool of 2 DNA sources at equal concentration (red, **A-C**) and two flower pools (**D-F**). Both replicated three times.

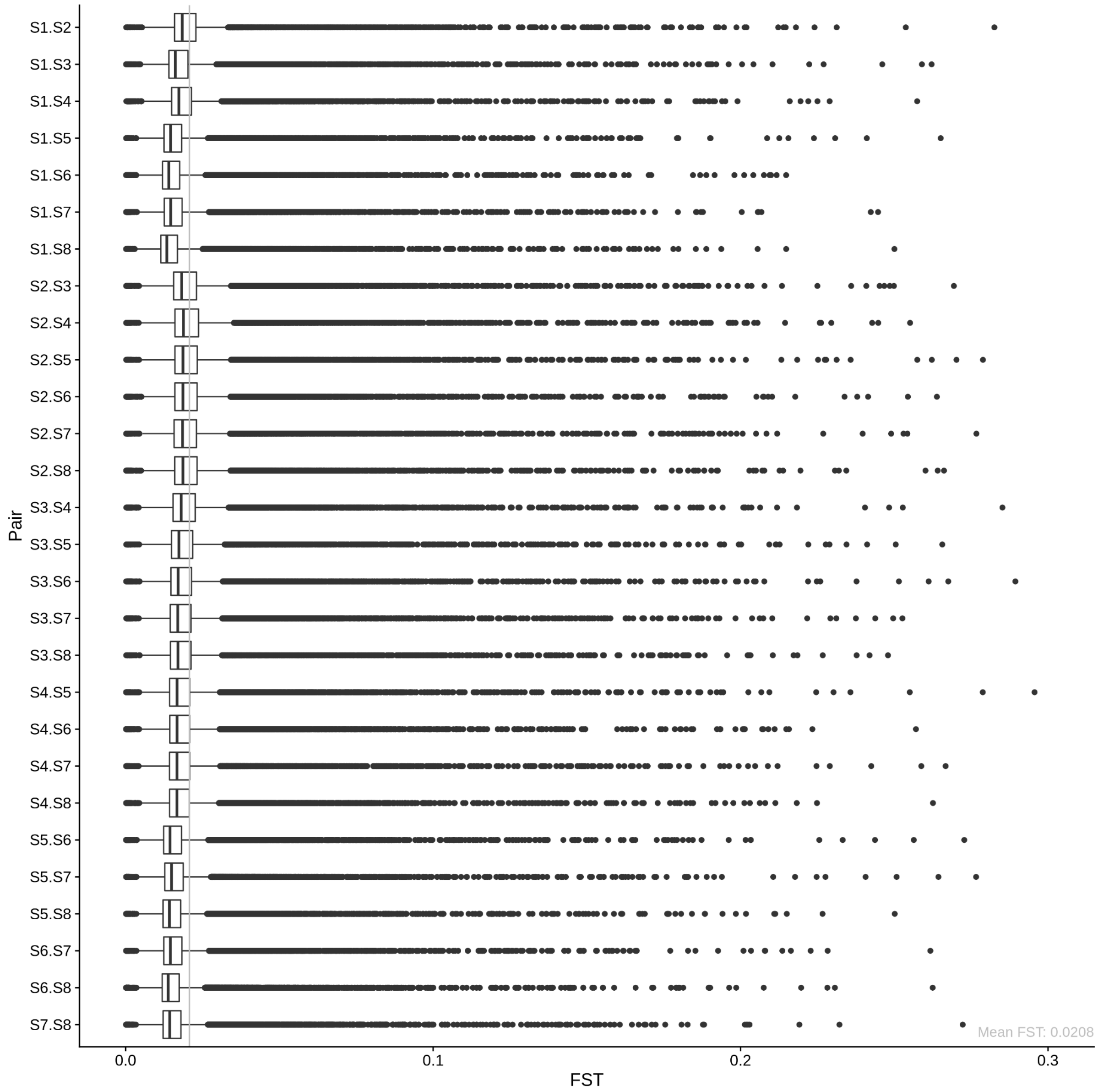

Fig. S15 | *Fst* distributions between all pairs of seed libraries.
Here, we computed Fst between all pairs of seeds (Experiment 1) in windows of 10k base pairs across all five chromosomes, based on frequencies from the mapped data (bam/pileup files). The window size was chosen to roughly fit the expected LD decay in *A. thaliana*. The boxes show the 25th, 50th, and 75th percentiles (i.e., quartiles), with whiskers extending to 1.5 times the interquartile range (distance between the first and third quartiles). Data points outside this range are plotted as individual points. The overall average Fst between all pairs is 0.0208, shown here as a gray vertical line, which represents the biological and statistical noise in population structure between replicates, and hence is the lower bound and baseline that we expect in other comparisons of Fst, as shown below.

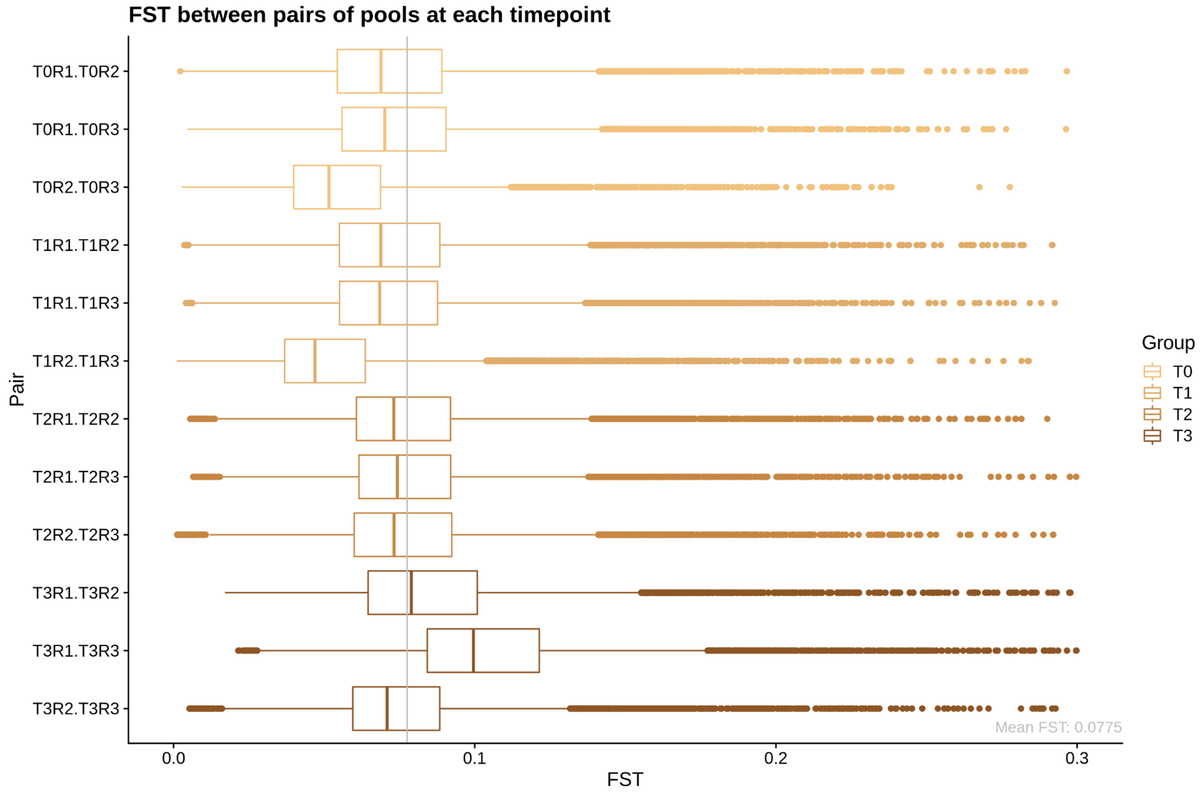

Fig. S15 | *Fst* distributions of all replicates of E&R (Exp. 4).

Here, we used the data from Experiment 4, which are three technical replicates (R1-R3) grown from the same seed mix (here encoded as "time point" T0), where flowers were collected at three different time points (T1-T3) during flowering time in the spring of 2016. We here show Fst between all pairs of replicates, across all time points, in windows of 10k base pairs across all five chromosomes. The window size was again chosen to fit the expected LD decay in *A. thaliana*. Properties of the box plots are as above in Figure S11. The mean Fst, again represented as a gray vertical line here, is 0.0775, which is more than three times the value of the seed baseline of 0.0208 in Figure 11. Note that the Fst of pairs that involve Replicate 1 is higher than that of the R2 vs R3 pairs. This is likely because R1 suffered a disturbance in the soil that could have created a bottleneck in this population and hence increased differentiation.

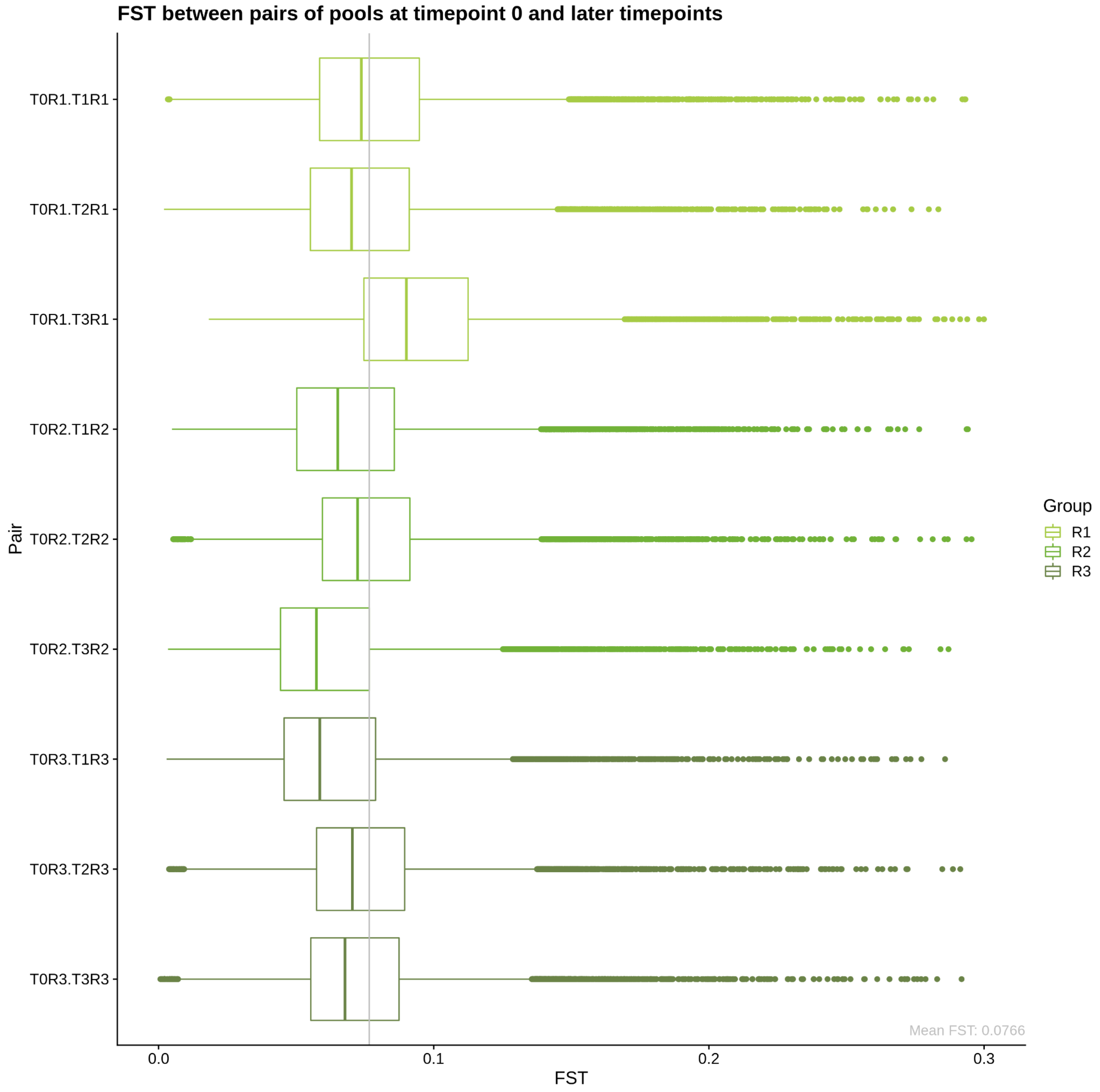

Fig. S16 | *Fst* distributions of Time 0 and all replicates in E&R (Exp. 4).
Here, we used the data from Experiment 4, which are three technical replicates (R1-R3) grown from the same seed mix (here encoded as "time point" T0), where flowers were collected at three different time points (T1-T3) during flowering time in the spring of 2016. We here show Fst between the seeds and the flowering time points for each replicate, in windows of 10k base pairs across all five chromosomes. The window size was again chosen to fit the expected LD decay in *A. thaliana*. Properties of the box plots are as above in Figure X. The mean Fst across all replicates and timepoints, again represented as a gray vertical line here, is 0.0766, which is more than three times the value of the seed baseline of 0.0208 in Figure X. This indicates that even within one generation (from seeds to flowers), there is some differentiation happening, which suggests that rapid adaptation to the local environment of the field site has taken place.

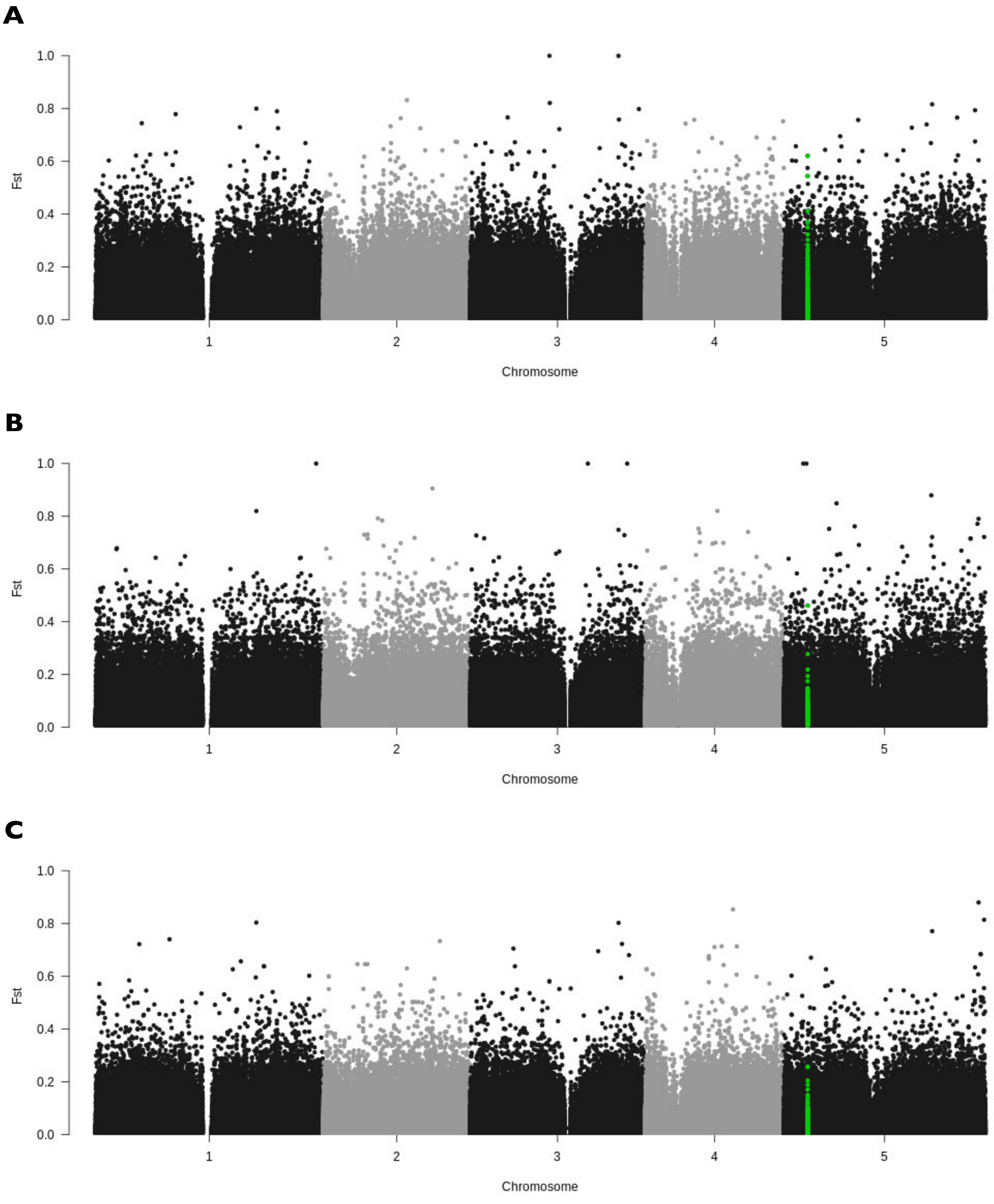

Fig. S17 | Example genome-wide F_{ST} scan of E&R (Exp 4).

Genome-wide F_{ST} calculated using `grenedalf` for replicate 2 between Timepoint 0 and Timepoint 1 (**A**), Timepoint 2 (**B**), and Timepoint 3 (**C**). SNPs overlapping with the *Flowering Locus C* (*FLC*) gene are highlighted in green (zoom in of the region in main Fig. 7). Only SNPs part of the *bona fide* 11,769,920 biallelic SNPs are shown.

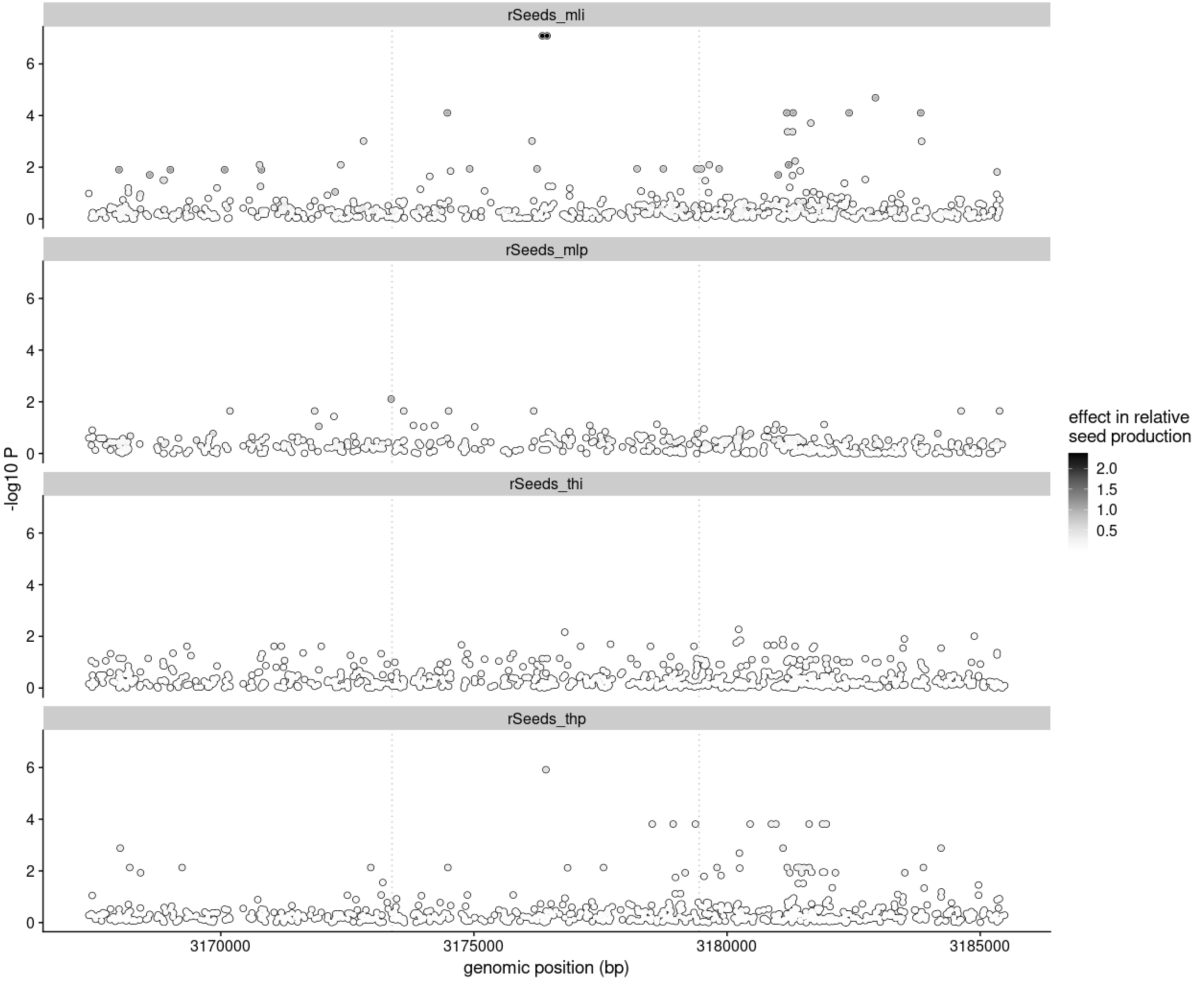

Fig. S18 Example genome-wide F_{ST} scan of E&R (Exp 4).

Genome-Wide association of seed set in four key conditions from Exposito-Alonso et al. 2019. Along with condition Tübingen-high precipitation-population replicate ("thp", Fig. 7), condition Madrid-low precipitation-individual replicate ("mli") also show signs of SNP association to seed set within as well as the putative promoter region of the FLC gene (dotted box).

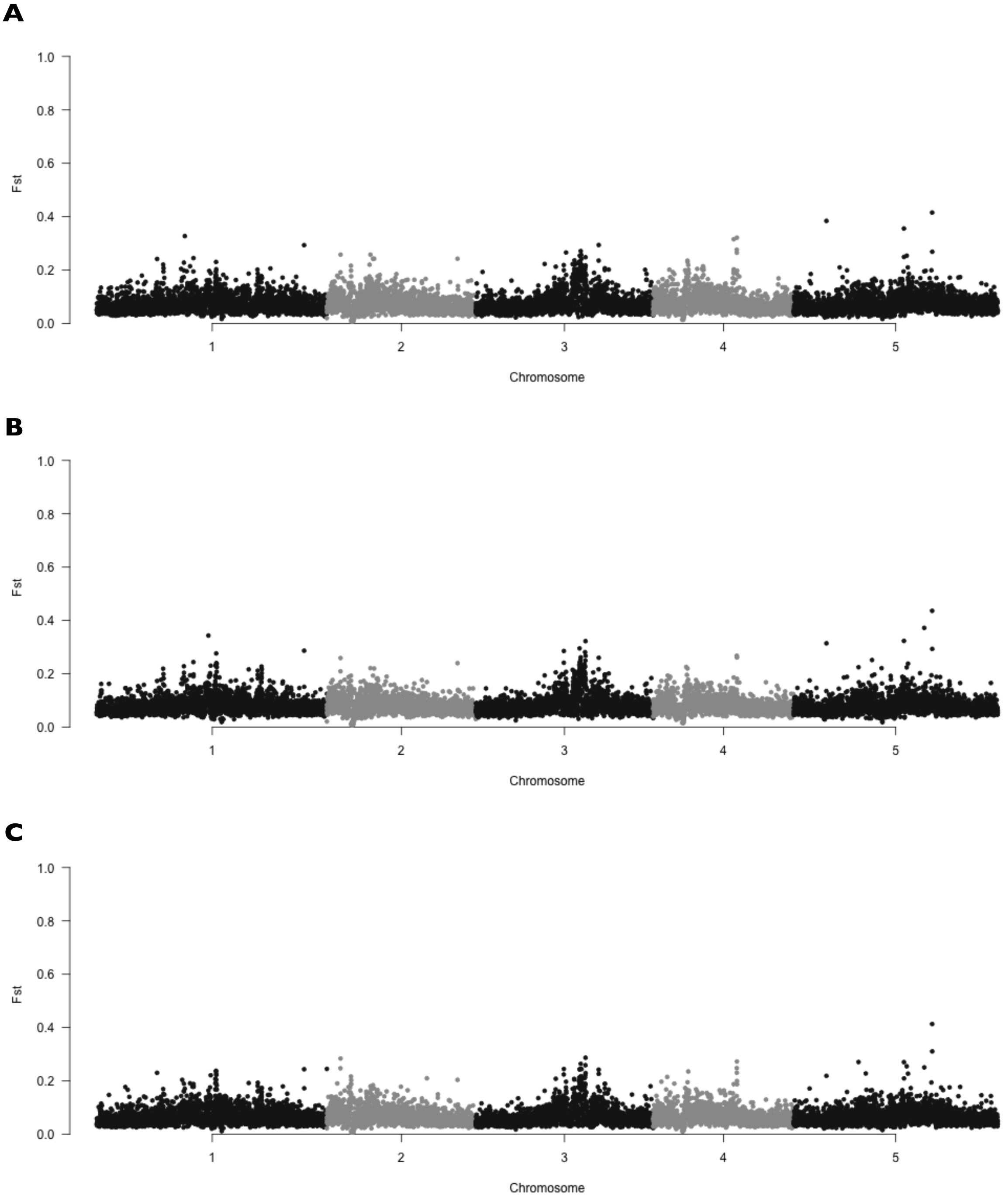

Fig. S19 | Example genome-wide F_{ST} scan of E&R (Exp 4) with 10Kb window averages. Genome-wide F_{ST} averages over 10Kb regions for the same samples as Fig. S17. Multiple areas in the genome show high differentiation.

Supplementary Text:
Pool-Sequencing corrections for population genetic statistics

Lucas Czech, Jeffrey P. Spence, and Moisés Expósito-Alonso
Correspondence: moisesexpositoalonso@gmail.com, luc@s-cze.ch

This document describes our assessment of pool-sequencing-specific equations for population genetic measures of diversity (such as θ_π, $\theta_{\text{Watterson}}$, Tajima's D), and differentiation (such as F_{ST}). We re-render some approaches originally presented and implemented in PoPoolation [10] and PoPoolation2 [11]. The aim of these equations is to correct for biases of pool sequencing, such as limited sample size (number of individuals pooled, or pool size n) and limited read size (number of reads obtained from those individuals, or coverage C).

This document is largely based on two sources:

- The reverse-engineered code of PoPoolation and PoPoolation2. We want this document to represent the equations that are actually computed when running these programs, as we feel that they need a more thorough assessment than what is available in the current literature.
- The PoPoolation equations document `correction_equations.pdf` as found in their code repository; we provide a copy at https://github.com/lczech/popoolation/blob/master/files/correction_equations.pdf. This document derives some of the equations implemented, but also contains some more that might be interesting for a deeper understanding of the topic.

Lastly, we here introduce novel estimators for F_{ST} for pool sequencing data, that correct for both biases described above.

DISCLAIMER: This is a draft document (last updated 2022-02-02) that describes the equations as implemented in PoPoolation, to the best of our knowledge. We aim to make this public to discuss our new implementation and novel methods improving on PoPoolation with its developers and the research community. Please reach out to us if you have comments or feedback.

1 Definitions

1.1 Pool Sequencing Data

We first define the input that we assume to be given for all subsequent equations. In the software implementation of the equations, these are be based on the input data, or set by the user as parameters.

n : Pool size, provided by the user. This is the number of individuals that were pooled together for sequencing.

C : Observed coverage. This is the number of reads sequenced from the pool that span the given position in the genome.

b : Minimum allele count, provided by the user. We do not want to consider SNPs with fewer than b alternative reads in the data, as they might be sequencing errors. Note that we assume b to be a user-provided constant, and hence leave it out of (most) function arguments for simplicity.

1.2 Notation

τ : Nucleotides, with $\tau \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$.

c_τ : Nucleotide counts, i. e., how many reads have a certain nucleotide τ at a given genomic position. Hence, $C = \sum_\tau c_\tau$.

$\boldsymbol{c}$: Vector of nucleotide counts (for convenience), i. e., $\boldsymbol{c} = (\ c_{\text{A}}, c_{\text{C}}, c_{\text{G}}, c_{\text{T}}\ )$.

f_τ : Nucleotide frequencies, i. e., $f_\tau = c_\tau / C$.

$\boldsymbol{f}$: Vector of nucleotide frequencies (for convenience), i. e., $\boldsymbol{f} = (\ f_\mathrm{A}, f_\mathrm{C}, f_\mathrm{G}, f_\mathrm{T}\ )$.

u, v : For biallelic SNP positions, we simplify, and instead of the four c_τ values, just use u for the count of the reference (major, or "first") allele, and v for the count of the alternative (minor, or "second") allele. We here leave it open whether the reference allele is defined by some reference genome, or simply the major (highest count) allele. In the PoPoolation notation, this means that $u \ \widehat{=}\ m$, or sometimes $u \ \widehat{=}\ i$; PoPoolation uses both, depending on context, and $v \ \widehat{=}\ C - m$, or $v \ \widehat{=}\ C - i$.

m : Index of summation over potential levels of coverage C.

k : Index of summation over potential pool sizes n.

1.3 Harmonic Numbers

We define a_1 and a_2 based on (generalized) harmonic numbers, as the sum of (squared) reciprocals of the first $n - 1$ positive integers:

$$a_1(n) = \sum_{k=1}^{n-1} \frac{1}{k} \tag{1}$$

$$a_2(n) = \sum_{k=1}^{n-1} \frac{1}{k^2} \tag{2}$$

These will be needed in several of the below equations. We use this notation as a compromise between Equation (3.6) of Hahn (2018) [5] and the notation of a_n and b_n used in Achaz (2008) [1] for these quantities. Note that the standard definition of harmonic numbers $H(n)$ includes the n-th element, which the above definition does not.

2 Theta Pi

First, we derive equations for θ_π, also called Tajima's π, based on its original (classic) definition, but correcting for biases introduced by pool sequencing, following PoPoolation.

We have two aims. First, to produce an unbiased estimator of θ_π based on pool size and coverage-induced noise, and second, to derive an expectation of the "population mutation rate" θ from the former estimator.

Unbiased Pool-seq estimator of θ_π for biallelic SNPs

We first define θ_π, as usual, as the heterozygosity based on the allele frequencies at a given locus:

$$\theta_\pi(\boldsymbol{f}) = \frac{n}{n-1} \left(1 - \sum_\tau f_\tau^2 \right) \tag{3}$$

using the possible nucleotide frequencies $\boldsymbol{f} = (\ f_\mathrm{A}, f_\mathrm{C}, f_\mathrm{G}, f_\mathrm{T}\ )$, with $\tau \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, with Bessel's correction for the number of sequences n (which is the equivalent of individual sequencing for our pool size n). See for example Equation (3.1) of Hahn (2018) [5] for the original definition for individuals.

In the pool-sequencing case, however, we have a coverage of C reads at a given position in the genome. We hence can use nucleotide counts c_τ instead, that is, $f_\tau = c_\tau / C$. We furthermore use Bessel's correction $\frac{C}{C-1}$ based on the read coverage to obtain an unbiased estimate, and reformulate the above as:

$$\theta_\pi(\boldsymbol{c}, C) = \frac{C}{C-1} \left(1 - \sum_\tau \frac{c_\tau^2}{C^2} \right) \tag{4}$$

At this point, the PoPoolation equations document begins to simplify the above equation, and then breaks it down for biallelic SNPs. However, their (and our) implementation differ from this, and use the above equation that can work with any (not just biallelic) SNPs. We hence do not introduce these simplifications here. Note however that the computation is still only conducted on biallelic sites, as the correction term introduced below assumes this.

Expected value of population mutation rate θ from nucleotide diversity

We now use the expectation of the θ_π estimator to infer the population mutation rate $\hat{\theta}$, accounting for noise generated by the total coverage C and pools of n individuals. We assume a biallelic site, and use expectations from a neutral site frequency spectrum under constant population size:

$$\mathbb{E}(\theta_\pi | C, n) = P(\mathrm{SNP}|n) \cdot \sum_{m=b}^{C-b} \theta_\pi(m, C) \cdot P(m|C, n) \tag{5}$$

In words, the expected value is computed by summing all possible SNP counts (that exceed the minimum count b) that can occur in a pool with coverage C (using the first/major allele count m here, with second/minor allele count $C - m$ implicit), weighted by the probability to have each of those counts, and scaled by the probability to have a SNP in the first place.

Here, we are using the minimum allele count b that we want to consider (as provided by the user), meaning that we only consider SNPs that have at least b reads for either the first or second allele. As we are only using the first allele count m in the equation above, and do not know which of the two counts is the larger one, we “sandwhich” our potential values for the coverage between b and $C - b$.

The two probabilities used above are computed as follows.

$P(\mathrm{SNP}|n)$ is the probability of observing a SNP in a pool of n individuals:

$$P(\mathrm{SNP}|n) = \theta \sum_{k=1}^{n-1} \frac{1}{k} = \theta a_1(n) \tag{6}$$

Assuming that all variation is neutral, and that the population is of constant size and in mutation-drift equilibrium, by definition, $\theta = \mathbb{E}(S/a_1(n))$ with S segregating sites. Then, the a_1 terms cancel out, meaning that Eq. (6) yields the proportion of variable sites.

$P(m|C, n)$ is the probability of observing m as first allele count in a SNP with C reads from a pool of dimension n:

$$P(m|C, n) = \frac{1}{a_1(n)} \sum_{k=1}^{n-1} \frac{1}{k} P(m|C, n, k) \tag{7}$$

$P(m|C, n, k)$ is the probability of having a first allele count of m observed in C reads that were taken from a pool of n individuals with first allele count of k. That is, m is the allele count in the reads, and k is the allele count in the pool:

$$P(m|C, n, k) = \binom{C}{m} \left(\frac{k}{n}\right)^m \left(\frac{n-k}{n}\right)^{C-m} \tag{8}$$

In words, $P(m|C, n, k)$ follows a binomial distribution, with m successes out of C trials with a success probability of k/n for each trial. That is, we compute how likely it is to observe m as the first/major allele count in C reads, given the frequency k/n of the major allele in the pool. Again, the count of the second/minor allele is implicitly used here as $C - m$.

Starting from Eq. (5), we can now put this together:

$$\begin{aligned}\mathbb{E}(\theta_\pi | C, n) &= P(\mathrm{SNP}|n) \cdot \sum_{m=b}^{C-b} \theta_\pi(m, C) \cdot P(m|C, n) \\ &= \theta a_1(n) \cdot \sum_{m=b}^{C-b} \frac{2u(C-m)}{C(C-1)} \cdot \frac{1}{a_1(n)} \sum_{k=1}^{n-1} \frac{1}{k} \binom{C}{m} \left(\frac{k}{n}\right)^m \left(\frac{n-k}{n}\right)^{C-m} \end{aligned} \tag{9}$$

Final approximation for population mutation rate Theta

We can now solve this for θ to define our final corrected estimate $\theta_{\pi,\mathrm{pool}}$.

$$\theta \approx \frac{\mathbb{E}(\theta_\pi | C, n)}{a_1(n) \cdot \sum_{m=b}^{C-b} \theta_\pi(m, C) \cdot P(m|C, n)} \tag{10}$$

This only leaves the $\mathbb{E}(\theta_\pi|C,n)$ term unresolved, which we however can estimate from our data using the classic estimator as shown in Eq. (4); note however that this is only evaluated on biallelic SNPs that have at least a count of b. In total, this yields:

$$\theta_{\pi,\text{pool}}(\boldsymbol{c},C,n) := \frac{\frac{C}{C-1}\left(1-\sum_\tau \frac{c_\tau^2}{C^2}\right)}{\sum_{m=b}^{C-b}\frac{2m(C-m)}{C(C-1)} \cdot \sum_{k=1}^{n-1}\frac{1}{k}\binom{C}{m}\left(\frac{k}{n}\right)^m\left(\frac{n-k}{n}\right)^{C-m}} \tag{11}$$

Note that the a_1 terms cancel out, and that the denominator only depends on the total coverage C and the pool size n (and not on any allele counts c_τ), and hence only needs to be computed once per coverage level, yielding a significant computational speedup.

The above computation is for a single variant site. For a whole region or window, the values of $\theta_{\pi,\text{pool}}(\boldsymbol{c},C,n)$ are simply summed up. This is the equation as implemented in PoPoolation as the measure called `pi`, and implemented in our grenedalf as well.

3 Theta Watterson

For Watteron's estimator θ_w, we follow the same approach as above. In order to derive the pool-sequencing corrected equations, we first define θ_w as usual:

$$\theta_w(u,C) = \frac{S(u)}{\sum_{k=1}^{C-1} 1/k} \tag{12}$$

where classically, S is the number of segregating sites, see for example Equation (3.5) of Hahn (2018) [5]. We are here working with a biallelic SNP at a single site, which as before we only want to consider if its count is within the limits of the minimum allele count b, and so we define:

$$S(u) = \begin{cases} 1 & \text{if } b \le u \le C-b \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Reasoning the same as above, we get the expected value of θ_w as:

$$\mathbb{E}(\theta_w|C,n) = P(\text{SNP}|n) \cdot \frac{\sum_{m=b}^{C-b} P(m|C,n)}{\sum_{k=1}^{C-1} 1/k}$$

with the two probability terms again as in Eq. (6) and Eq. (7). For conciseness, we here only resolve $P(\text{SNP}|n)$:

$$= \theta a_1(n) \cdot \frac{\sum_{m=b}^{C-b} P(m|C,n)}{\sum_{k=1}^{C-1} 1/k} \tag{14}$$

We can again solve this for θ, to get our corrected estimate:

$$\theta \approx \mathbb{E}(\theta_w|C,n) \cdot \frac{\sum_{k=1}^{C-1} 1/k}{a_1(n) \cdot \sum_{m=b}^{C-b} P(m|C,n)} \tag{15}$$

Again we can use the classic value $\theta_w(u,C)$ of Eq. (12) for the expected value $\mathbb{E}(\theta_w|C,n)$, so that the summation over $1/k$ in the numerator and in $\theta_w(u,C)$ cancel out here. We can now define our estimate:

$$\theta_{w,\text{pool}}(u,C,n) := \frac{S(u)}{a_1(n) \cdot \sum_{m=b}^{C-b} P(m|C,n)}$$

We can now replace $P(m|C,n)$ according to Eq. (7) and Eq. (8). The $a_1(n)$ terms in the denominator and in $P(m|C,n)$ cancel out, leading to the final equation:

$$\begin{aligned} &= \frac{S(u)}{\sum_{m=b}^{C-b}\sum_{k=1}^{n-1}\frac{1}{k}P(m|C,n,k)} \\ &= \frac{S(u)}{\sum_{m=b}^{C-b}\sum_{k=1}^{n-1}\frac{1}{k}\binom{C}{m}\left(\frac{k}{n}\right)^m\left(\frac{n-k}{n}\right)^{C-m}} \end{aligned} \tag{16}$$

As before, the denominator only depends on the coverage C, and hence only needs to be computed once per coverage level that is present in the data.

Again, the approach to compute this for a window is to sum up all values across the SNPs in the window. This is the equation as implemented in PoPoolation as the measure called `theta`, and implemented in our grenedalf as well.

4 Tajima's D

Above, we have defined pool-sequencing corrected estimators θ_π and θ_w. Now, we want to use them to define a test akin to Tajima's D for pool sequencing. We are here again following the PoPoolation approach, and re-derive their equations.

4.1 Pool-Sequencing Correction

The PoPoolation equations document derives the following estimator. To the best of our knowledge, this is however not implemented in PoPoolation; instead, they compute Tajima's D as presented in the following Section (4.2). We still introduce the approach here, for reference, and in the hope that it might be helpful.

First, we define:

$$d_{\text{pool}}(u, C) \ := \ \theta_{\pi,\text{pool}}(u, C) \ - \ \theta_{w,\text{pool}}(u, C) \tag{17}$$

using the major allele count u at a site with coverage C, and use this to define our statistic:

$$D_{\text{pool}}(u, C) \ := \ \frac{d_{\text{pool}}(u, C)}{\sqrt{\text{Var}(d_{\text{pool}}(u, C))}} \tag{18}$$

In order to compute the variance of d_{pool} (leaving out function arguments for simplicty), we start with the standard expansion of the variance:

$$\text{Var}(d_{\text{pool}}) = \mathbb{E}(d_{\text{pool}}^2) - \mathbb{E}(d_{\text{pool}})^2$$

At this point, we use that d_{pool} is unbiased (for populations at equilibrium) , and hence has an expected value of 0, that is, $\mathbb{E}(d_{\text{pool}})^2 = 0$. The PoPoolation equations document notes that this is only true if they did their previous calculations correctly, but we trust they did .

Then, we can compute the variance as:

$$\begin{aligned} \text{Var}(d_{\text{pool}}) &= \mathbb{E}(d_{\text{pool}}^2) \\ &= P(\text{SNP}|n) \cdot \sum_{m=b}^{C-b} d_{\text{pool}}^2(m, C) \cdot P(m|C, n) \end{aligned}$$

which can be resolved using equations Eq. (6) and Eq. (7) from previous sections:

$$= \theta \cdot \sum_{m=b}^{C-b} \left(\theta_{\pi,\text{pool}}(m, C) - \theta_{w,\text{pool}}(m, C)\right)^2 \cdot \sum_{k=1}^{n-1} \frac{1}{k} \binom{C}{m} \left(\frac{k}{n}\right)^m \left(\frac{n-k}{n}\right)^{C-m} \tag{19}$$

This leaves θ to be estimated. PoPoolation suggests to estimate it as $\theta_{\pi,\text{pool}}$ on the same window on which we are computing D_{pool}. This assumes that all individuals contribute the same number of reads to the pool.

The first summation in Eq. (19) involves computing $\theta_{\pi,\text{pool}}$ and $\theta_{w,\text{pool}}$ repeatedly $C - 2b$ many times, with each of these computations involving to compute their respective denominators, as shown in Eq. (11) and Eq. (16). However, as C remains constant throughout this computation, these denominators (the correction terms) are identical, so that we only need to compute them once, to gain a $\approx C$-fold speedup.

At this point, the PoPoolation equation document also introduces an approach to compute Tajima's D based on the above in windows. We here skip this part for brevity.

4.2 Integration with Classic Tajima's D

On large windows, the classic Tajima's D is not a measure of significance (in number of standard deviations away from the null hypothesis), but instead is a measure of the magnitude of the divergence from neutrality. This is because all loci are considered completely linked, even if they are not in reality.

However, the above pool-sequencing Tajima's D instead consideres all loci as completely unlinked, and thus represents the number of standard deviations away from neutrality. Therefore, it gives a different numerical result that has a much higher absolute value compared to classic Tajima's D.

Now, we want to obtain a correction term for the pool-sequence Tajima's D to obtain values that are comparable to classic Tajima's D in non-small windows, that is, we want a measure of the magnitude of the divergence from neutrality. We again follow the PoPoolation approach, and here derive the equations that are actually implemented.

Approach by Achaz

To this end, PoPoolation2 uses a modified version of the Y^* test of Achaz (2008) [1], which was originally developed as a test for neutrality despite the presence of sequencing errors. This test only works when excluding singletons, that is, we set $b := 2$ for this part.

Following PoPoolation and Achaz (2008) [1], we first define:

$$f^*(n) = \frac{n-3}{a_1(n) \cdot (n-1) - n} \tag{20}$$

which is then used to define:

$$\alpha^*(n) = f^{*2} \cdot \left(a_1(n) - \frac{n}{n-1} \right) + f^* \cdot \left(a_1(n) \cdot \frac{4(n+1)}{(n-1)^2} - 2 \cdot \frac{n+3}{n-1} \right) - a_1(n) \cdot \frac{8(n+1)}{n(n-1)^2} + \frac{n^2+n+60}{3n(n-1)} \tag{21}$$

and:

$$\begin{aligned} \beta^*(n) = f^{*2} \cdot \left(a_2(n) - \frac{2n-1}{(n-1)^2} \right) + f^* \cdot \left(a_1(n) \cdot \frac{8}{n-1} - a_1(n) \cdot \frac{4}{n(n-1)} - \frac{n^3+12n^2-35n+18}{n(n-1)^2} \right) \\ - a_1(n) \cdot \frac{16}{n(n-1)} + a_1(n) \cdot \frac{8}{n^2(n-1)} + \frac{2(n^4+110n^2-255n+126)}{9n^2(n-1)^2} \end{aligned} \tag{22}$$

Note that these equations were originally developed for data from individuals, and hence here, n denotes the number of individuals *as if* we were doing individual sequencing.

NB: The PoPoolation document recommends to counter-check the correctness of their equation with the original of Achaz (2008) [1]. In fact, PoPoolation introduced a slight mistake in the last term of β^*, which we have fixed here. Above is the (hopefully) correct one, following Achaz (2008) [1]. Note that the mistake only concerns the PoPoolation equations document, but not their implementation.

The number of individuals sequenced

The only unresolved parameter is n, which corresponds to the number of individuals sequenced – if we were to do individual sequencing. In our case of pool sequencing, according to PoPoolation, we can reasonably substitute this with the expected number of distinct individuals sequenced.

To this end, we use the coverage C, as well as the pool size n, which we here use as our substitute for the number of individuals sequenced. Then, we define $\tilde{n}$ (called $\tilde{n}_{base}$ in the PoPoolation equations document) as the expected number of individuals from our pool that have been sequenced:

$$\tilde{n} = \sum_{k=1}^{T} \sum_{j=1}^{k} (-1)^{k-j} \cdot k \binom{n}{k} \binom{k}{j} \left(\frac{j}{n} \right)^C \tag{23}$$

where $T = \max(C, n)$; if n is much larger than C, we can assume $\tilde{n} \approx C$. Our substitute $\tilde{n}$ is then obtained by averaging $\tilde{n}$ over the window W.

Computing the expected number of distinct individuals sequenced corresponds to the following statistical question: Given a set of integers $A = \{1, \ldots, n\}$ (corresponding to individuals), pick a set B of C elements from set A with replacement (corresponding to reads); what is the expected number of distinct values (individuals) that have been picked in B (that we have reads from)?

PoPoolation computes this value by brute force using Eq. (23), that is, by trying all possible ways to pick numbers from the set. However, there exists a closed form solution to this question, which yields massive speedups for larger coverages, which we have implemented.

One way to arrive at the closed form expression is as follows: Define an indicator random variable I_i for $1 \leq i \leq n$ as 1 if individual i is present in the set B (that is, if individual i has been sequenced), and as 0 if not. Then, the size of set B is simply $\sum_{i=1}^{n} I_i$.

The probability that I_i equals 1 (that is, that individual i has been sequenced) for any i is given by:

$$P(I_i = 1) = 1 - \left(\frac{n-1}{n}\right)^C \tag{24}$$

In words, this is the complement of *not* picking i in all of the C picks from set A.

The expected size of the set B can then be computed by linearity of expectation for all i, yielding our closed form expression:

$$\tilde{n} = n\left(1 - \left(\frac{n-1}{n}\right)^C\right) \tag{25}$$

This is equation that we compute in our implementation to arrive at $\tilde{n}$ for a given coverage C and poolsize n.

Final estimator for D

Now that we have a way of computing a reasonable value for the number of individuals sequenced, we can finally define the estimator:

$$\tilde{D}_{\text{pool}} = \frac{\theta_\pi - \theta_w}{\sqrt{|W|^{-1} \cdot \alpha^*(\tilde{n}) \cdot \theta \;+\; \beta^*(\tilde{n}) \cdot \theta^2}} \tag{26}$$

following PoPoolation and Achaz (2008) [1]. This requires $b = 2$; furthermore, PoPoolation suggests to use “not too small” windows. We are using the size $|W|$ of the window here, that is, the total length along the genome, which is typically much larger than the number of SNPs in the window . The θ used in the denominator is simply θ_w again.

The above is the estimator as implemented in PoPoolation and in our implementation.

4.3 Assumptions and Biases

In the above computation of the correction term for Tajima’s D for pool sequencing, several assumptions are made that lead to the resulting estimator being conservative, i. e., yielding smaller values that what would be expected from individual sequencing of samples. Based on the explanation in the PoPoolation manual (most of the text in this section is copied from there), we here explore the underlying assumptions and biases.

The locally fluctuating coverage is replaced by the minimum coverage. This makes the variance estimator larger, and therefore leads to conservative estimates of Tajima’s D.

The random number of different individuals sequenced under a given coverage C is replaced by its expected value $\tilde{n}$. This assumption should not affect the results much: If the pool size is large compared to the coverage, sequencing the same individual more than once is uncommon.

Furthermore the number of different individuals sequenced will have a low variance. As we are working with the minimum coverage, $\tilde{n}$ will be biased downwards, tending to give a conservative estimate of the variance.

At different positions, the subsets from the pool are sequenced might be different. Their coalescent histories will be correlated but not identical. As the classical equations for Tajima’s D are for single samples sharing a common coalescent history, there is more independence in the data than assumed with the classical formula. This again should make the variance approximation more conservative.

Summing up, the approximate variance in the above equations provides a conservative approximation, and the values for Tajima's D will tend to be smaller than those that would be expected for an experiment based on individual sequencing of single samples.

Lastly, the PoPoolation code repository contains a plot showing the correlation between the classical Tajima's D and the corrected Tajima's D using the equations described above; please see here, where the x-axis corresponds to the classical value, and the y-axis the the corrected one. This plot has been made with real-world data from Drosophila with a coverage of 12, a window size of 500 and a minimum count of 1.

4.4 PoPoolation Bugs

From our assessment of the PoPoolation code, and from personal communication with Robert Kofler, we suspect that the implementation of the above $\tilde{D}_{\text{pool}}$ in PoPoolation $\leq$ v1.2.2 contains several bugs, which alter the numerical results of the computation of Tajima's D. At the moment, we are in contact with Robert Kofler, are still verifying these bugs, and are investigating their consequences. We here want to thank Robert for his positive reply and his support regarding our questions.

5 Fixation Index F_{ST} for Pool-Seq

In this section, we will derive unbiased estimators of various measures of heterozygosity in two populations for Pool-sequencing data. These will then be combined to obtain "sample-size" and "pool-size" corrected estimators of two definitions of F_{ST}. On top of these two novel estimators for F_{ST} in the pool-sequencing context, we also walk through the two existing estimators as suggested by Kofler *et al.* (2011b) [11] and Karlsson *et al.* (2007) [9]. Both are implemented in PoPoolation2, and are called the "classical" or "conventional pool sequencing" approach, and the "Karlsson approach adapted to digital data", respectively, in Kofler *et al.* (2011b) [11]. We compare all four approaches to each other, and show that the "classical" approach is biased for lower coverages or small pool sizes, and the Karlsson approach is biased for small pool sizes (bias on the order of 1/pool size). See also Hivert *et al.* (2018) [7] for an assessment of F_{ST} in the pool-sequencing context.

There are several non-equivalent *definitions* of F_{ST}. The overall goal is to measure some degree of differentiation between two populations, which can be represented as a proportion of variation that cannot be explained by variation within populations. What is unclear is a proportion of *what* variation? There are two natural candidates leading to two related, but distinct definitions of F_{ST}. The first definition, which we will call F_{ST}^{Nei} following Nei (1973) [12], considers the proportion of the total variation in the two populations. This statistic is also called G_{ST}, see for example Equation (5.5) of Hahn (2018) [5]. The second definition, which we will call F_{ST}^{Hudson} follwing Hudson *et al.* (1992) [8], considers the proportion of the variation between populations, see also Cockerham (1969) [4] and Weir and Hill (2002) [13]. This second definition is also considered in Karlsson *et al.* (2007) [9], which we examine below in Section (5.4).

To make this more formal, we can consider the probability that two haploids carry different alleles. We could consider drawing the two haploids from the same population (with the population chosen at random), which we call π_{within}; or we could consider drawing the two haploids from *different* populations, which we call π_{between}; or finally we could consider drawing the two haploids totally at random from either population (potentially the same populations, potentially different populations) which we call π_{total}. See Bhatia *et al.* (2013) [2] for more background information on this.

Our two definitions of F_{ST} are then

$$F_{ST}^{\text{Nei}} = 1 - \frac{\pi_{\text{within}}}{\pi_{\text{total}}} \tag{27}$$

$$F_{ST}^{\text{Hudson}} = 1 - \frac{\pi_{\text{within}}}{\pi_{\text{between}}} \tag{28}$$

If we consider a single locus with up 4 alleles, with frequencies $f_{\tau p}$ (possibly zero) with τ denoting the allele with $\tau \in \{A, C, G, T\}$ and p denoting the population with subscripts (1) and (2), we can calculate the various πs as follows

$$\pi_{\text{within}} = \frac{1}{2}\left[\left(1 - \sum_{\tau} f^2_{\tau(1)}\right) + \left(1 - \sum_{\tau} f^2_{\tau(2)}\right)\right] \tag{29}$$

$$\pi_{\text{between}} = 1 - \sum_{\tau} f_{\tau(1)} f_{\tau(2)} \tag{30}$$

$$\pi_{\text{total}} = \frac{1}{2}\pi_{\text{within}} + \frac{1}{2}\pi_{\text{between}} \tag{31}$$

which are then used in our above definitions of $\mathrm{F_{ST}}$.

5.1 Unbiased estimators of the πs

Since both definitions of $\mathrm{F_{ST}}$ rely on these πs, we will need to derive unbiased estimates for them. We will show below that the following are unbiased estimators of the corresponding quantities without hats:

$$\begin{aligned}\widehat{\pi}_{\text{within}} = \frac{1}{2}\Bigg[&\left(\frac{n_{(1)}}{n_{(1)}-1}\right)\left(\frac{C_{(1)}}{C_{(1)}-1}\right)\left(1 - \sum_{\tau}\left(\frac{c_{\tau(1)}}{C_{(1)}}\right)^2\right) \\ + \quad &\left(\frac{n_{(2)}}{n_{(2)}-1}\right)\left(\frac{C_{(2)}}{C_{(2)}-1}\right)\left(1 - \sum_{\tau}\left(\frac{c_{\tau(2)}}{C_{(2)}}\right)^2\right)\Bigg]\end{aligned} \tag{32}$$

$$\widehat{\pi}_{\text{between}} = 1 - \sum_{\tau}\left(\frac{C_{\tau,1}}{C_{(1)}}\right)\left(\frac{C_{\tau,2}}{C_{(2)}}\right) \tag{33}$$

$$\widehat{\pi}_{\text{total}} = \frac{1}{2}\widehat{\pi}_{\text{within}} + \frac{1}{2}\widehat{\pi}_{\text{between}} \tag{34}$$

In the following, we derive these estimators.

Unbiased estimator of $\widehat{\pi}_{\text{within}}$

We have derived previously that

$$\mathbb{E}\left[\left(\frac{n_{(1)}}{n_{(1)}-1}\right)\left(\frac{C_{(1)}}{C_{(1)}-1}\right)\left(1 - \sum_{\tau}\left(\frac{C_{\tau,1}}{C_{(1)}}\right)^2\right)\right] = \left(1 - \sum_{\tau} f^2_{\tau,1}\right),$$

within a single population. It follows immediately that averaging these estimators across the two populations is unbiased for π_{within}.

Unbiased estimator of $\widehat{\pi}_{\text{between}}$

Since the two pools are independent, we have that

$$\mathbb{E}\left[\widehat{\pi}_{\text{between}}\right] = 1 - \sum_{\tau}\mathbb{E}\left[\left(\frac{C_{\tau,1}}{C_{(1)}}\right)\right]\mathbb{E}\left[\left(\frac{C_{\tau,2}}{C_{(2)}}\right)\right].$$

The frequency of alleles within a pool is an unbiased estimate for the frequency in the population, so

$$\mathbb{E}\left[\left(\frac{C_{\tau,p}}{C_{(p)}}\right)\right] = f_{\tau,p},$$

showing that $\widehat{\pi}_{\text{between}}$ is unbiased for π_{between}.

Unbiased estimator of $\widehat{\pi}_{\text{total}}$

That $\widehat{\pi}_{\text{total}}$ is unbiased for π_{total} follows immediately from the definition of π_{total} in Eq. (31) and the unbiasedness of $\widehat{\pi}_{\text{within}}$ and $\widehat{\pi}_{\text{between}}$.

5.2 Final unbiased estimators of $\mathrm{F_{ST}}$ per SNP and per window

These estimators then immediately suggest the following ratio estimators for the different definitions of $\mathrm{F_{ST}}$:

$$\widehat{\mathrm{F}}_{\mathrm{ST}}^{\mathrm{Nei}} = 1 - \frac{\widehat{\pi}_{\text{within}}}{\widehat{\pi}_{\text{total}}} \tag{35}$$

$$\widehat{\mathrm{F}}_{\mathrm{ST}}^{\mathrm{Hudson}} = 1 - \frac{\widehat{\pi}_{\text{within}}}{\widehat{\pi}_{\text{between}}} \tag{36}$$

All of this has been for a single site, but we are often interested in combining information across SNPs within a window W (or possibly genome wide). In such a case, define $\widehat{\pi}^{\ell}_{\text{within}}$ to be $\widehat{\pi}_{\text{within}}$ as above but for SNP $\ell \in W$. Define $\widehat{\pi}^{\ell}_{\text{between}}$ and $\widehat{\pi}^{\ell}_{\text{total}}$ analogously. We then combine information across the SNPs in the window W as

$$\widehat{\mathrm{F}}_{\mathrm{ST}}^{\mathrm{Nei}} = 1 - \frac{\sum_{\ell \in W} \widehat{\pi}^{\ell}_{\text{within}}}{\sum_{\ell \in W} \widehat{\pi}^{\ell}_{\text{total}}} \tag{37}$$

$$\widehat{\mathrm{F}}_{\mathrm{ST}}^{\mathrm{Hudson}} = 1 - \frac{\sum_{\ell \in W} \widehat{\pi}^{\ell}_{\text{within}}}{\sum_{\ell \in W} \widehat{\pi}^{\ell}_{\text{between}}} \tag{38}$$

See Bhatia *et al.* (2013) [2] for a practical and theoretical justification for using this “ratio of averages” instead of using an “average of ratios”. These are our asymptotically unbiased estimators for $\mathrm{F_{ST}}$ for Pool-seq data, which take the finite sampling of individuals from the population, and the finite sampling of reads from each individual in the pool, into account.

At the time of writing, we have only theoretically derived these estimators, but have not yet implemented them in our software.

5.3 Estimator of $\mathrm{F_{ST}}$ as implemented in PoPoolation2

The implementation in PoPoolation2 [11] offers two ways to estimate $\mathrm{F_{ST}}$: What they call the “classical” or “conventional” approach by Hartl and Clark (2007) [6], and an approach adapted to digital data following Karlsson *et al.* (2007) [9]. In this and the next section, we discuss these estimators. For comparability and historical backwards compatibility, we however still offer both these estimators in our implementation.

Furthermore, the PoPoolation equations document explains derivations of equations for Pool-seq corrected estimators of $\mathrm{F_{ST}}$, which however to the best of our knowledge are not actually implemented in either PoPoolation nor PoPoolation2. We here still walk through these, see Section (6).

First, we present the “classical” approach as implemented in PoPoolation2, labelled with superscript “PoPool” here. We compute $\mathrm{F_{ST}}$ for two subpopulations, which we here again denote with subscripts (1) and (2), and the total population with (T). We expect poolsizes $n >= 2$.

For each SNP in a given window, PoPoolation2 computes:

$$\widehat{\pi}_{(1)}^{\text{PoPool}} = \frac{C_{(1)}}{C_{(1)} - 1} \cdot \left(1 - \sum_{\tau} f_{\tau(1)}^2 \right) \tag{39}$$

$$\widehat{\pi}_{(2)}^{\text{PoPool}} = \frac{C_{(2)}}{C_{(2)} - 1} \cdot \left(1 - \sum_{\tau} f_{\tau(2)}^2 \right) \tag{40}$$

$$\widehat{\pi}_{(T)}^{\text{PoPool}} = \frac{C_{(T)}}{C_{(T)} - 1} \cdot \left(1 - \sum_{\tau} f_{\tau(T)}^2 \right) \tag{41}$$

with

$$C_{(T)} = \min\left(C_{(1)}, C_{(2)}\right)$$

$$f_{\tau(T)} = \frac{1}{2} \cdot \left(f_{\tau(1)} + f_{\tau(2)}\right)$$

These quantities are accumulated over the window W:

$$\widehat{\pi}_{W(1)}^{\text{PoPool}} = \frac{n_{(1)}}{n_{(1)} - 1} \cdot \sum_{W} \pi_{(1)} \tag{42}$$

$$\widehat{\pi}_{W(2)}^{\text{PoPool}} = \frac{n_{(2)}}{n_{(2)} - 1} \cdot \sum_{W} \pi_{(2)} \tag{43}$$

$$\widehat{\pi}_{W(T)}^{\text{PoPool}} = \frac{n_{(T)}}{n_{(T)} - 1} \cdot \sum_{W} \pi_{(T)} \tag{44}$$

with

$$n_{(T)} = \min\left(n_{(1)}, n_{(2)}\right)$$

Finally, the estimate of $\mathrm{F_{ST}}$is computed as:

$$\widehat{\mathrm{F}}_{\mathrm{FST}}^{\text{PoPool}} = \frac{\pi_{W(T)} - \frac{1}{2}\left(\pi_{W(1)} + \pi_{W(2)}\right)}{\pi_{W(T)}} \tag{45}$$

Because this estimator uses the minimum coverage and minimum pool size for either of the populations to calculate $\pi_{W(T)}$, it produces biased F_{ST} values for small pool sizes or coverages. This was also pointed out by Hivert *et al.* (2018) [7]. It is therefore recommended to use the unbiased estimator presented above.

5.4 Asymptotically Unbiased Estimator of $\mathrm{F_{ST}}$ by Karlsson *et al.*

Another estimator for $\mathrm{F_{ST}}$ that is offered in PoPoolation2 is based on the equations used in Karlsson *et al.* (2007) [9], see the last page of the Supplemental Information of Karlsson *et al.* for their derivation. We here briefly also go through the derivation.

We here call this estimator using the superscript "Karlsson", which is again defined for two subpopulations denoted with subscripts (1) and (2). We are here only looking at biallelic SNPs. Instead of τ for the four nucleotides, we hence use u for the major and v for the minor allele again, where u is the allele with the higher average frequency in the two subpopulations (as opposed to the allele with the highest total count).

We start with the definition of $\mathrm{F}_{\mathrm{FST}}^{\text{Karlsson}}$ from Karlsson *et al.* for the SNPs in a window W:

$$\mathrm{F}_{\mathrm{FST}}^{\mathrm{Karlsson}} = \frac{\sum_W N_k}{\sum_W D_k} \tag{46}$$

where the the numerator N_k and denominator D_k for a single site k in W are:

$$N_k = v_{(1)} \cdot \left(u_{(2)} - u_{(1)}\right) \; + \; v_{(2)} \cdot \left(u_{(1)} - u_{(2)}\right) \tag{47}$$

$$\begin{aligned} D_k &= v_{(1)} u_{(2)} + u_{(1)} v_{(2)} \\ &= N_k + v_{(1)} u_{(1)} + v_{(2)} + u_{(2)} \end{aligned} \tag{48}$$

These are estimated as follows, using the numerator $\hat{N}_k$ and denominator $\hat{D}_k$ at a single site:

$$\hat{N}_k = \left(\frac{u_{(1)}}{C_{(1)}} - \frac{u_{(2)}}{C_{(2)}}\right)^2 - \left(\frac{h_{(1)}}{C_{(1)}} + \frac{h_{(2)}}{C_{(2)}}\right) \tag{49}$$

$$\hat{D}_k = \hat{N}_k + h_{(1)} + h_{(2)} \tag{50}$$

with two additional helpers:

$$h_{(1)} = \frac{u_{(1)} \cdot v_{(1)}}{C_{(1)} \cdot \left(C_{(1)} - 1\right)}$$

$$h_{(2)} = \frac{u_{(2)} \cdot v_{(2)}}{C_{(2)} \cdot \left(C_{(2)} - 1\right)}$$

And finally, these are used to compute the asymptotically unbiased estimator $\widehat{\mathrm{F}}_{\mathrm{FST}}^{\mathrm{Karlsson}}$ for a window W:

$$\widehat{\mathrm{F}}_{\mathrm{FST}}^{\mathrm{Karlsson}} = \frac{\sum_W \hat{N}_k}{\sum_W \hat{D}_k} \tag{51}$$

According to Karlsson *et al.*, when the coverages $C_{(1)}$ and $C_{(2)}$ (called “sample sizes” there) are equal, the estimator reduces to the estimator of F_{ST} given by Weir and Hill (2002) [13]. Karlsson *et al.* further state that by the Lehmann-Scheffé theorem [3, Theorem 4.2.2], it follows that $\hat{N}_k$ and $\hat{D}_k$ are uniformly minimum variance unbiased estimators of N_k and D_k, respectively, and hence conclude that their estimator $\hat{\mathrm{F}}_{\mathrm{ST,K}}$ is also asymptotically unbiased.

The estimator above hence follows what we called the $\mathrm{F}_{\mathrm{ST}}^{\mathrm{Hudson}}$ definition. It however assumes the pool size to be infinite, that is, it is missing Bessel’s correction for pool size. Apart from that, it is identical to our estimator $\widehat{\mathrm{F}}_{\mathrm{ST}}^{\mathrm{Hudson}}$ as explained in Section (5.2).

6 PoPoolation2 Equations Document

The PoPoolations equation document also presents some simplifications and related equations that to the best of our knowledge are not implemented in their software. We hence do not go through them in detail here, but still want to mention them, in case they might be useful for others.

- They present simplified versions of θ_π, θ_w, and Tajima’s D, which assume that allele frequency distribution in the reads is about the same as in the real population, and hence arrives at a simpler computation at the cost of some error. These are also useful for individual sequencing.

- As mentioned above in Section (4.1), the document presents an approach to computing Tajima's D based on its variance, and extends this to windows, but (to the best of our knowledge) does not implement this, and instead implement their approach based on Achaz (2008) [1].
- They present an approach for computing F_{ST} for J pool-sequenced populations (instead of just two as presented above), extend this approach to large regions as well as single SNPs, and introduce weights that take the number of sequenced individuals in each population into account. More work is needed to compare this approach to their implementation and to our novel estimators.

These alternative approaches however need further assessment and comparison to the other approaches presented here.

References

[1] Achaz, G. (2008). Testing for neutrality in samples with sequencing errors. *Genetics*, **179**(3), 1409–1424.

[2] Bhatia, G. *et al.* (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, **23**(9), 1514–1521.

[3] Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day Series in Probability and Statistics. Prentice Hall.

[4] Cockerham, C. C. (1969). Variance of Gene Frequencies. *Evolution*, **23**(1), 72.

[5] Hahn, M. W. (2018). *Molecular Population Genetics*. Sinauer Associates, Oxford University Press.

[6] Hartl, D. L. and Clark, A. G. (2007). *Principles of Population Genetics*. Sinauer.

[7] Hivert, V. *et al.* (2018). Measuring Genetic Differentiation from Pool-seq Data. *Genetics*, **210**(1), 315–330.

[8] Hudson, R. R. *et al.* (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, **132**(2), 583–589.

[9] Karlsson, E. K. *et al.* (2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genetics*, **39**(11), 1321–1328.

[10] Kofler, R. *et al.* (2011a). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS ONE*, **6**(1), e15925.

[11] Kofler, R. *et al.* (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, **27**(24), 3435–3436.

[12] Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences*, **70**(12), 3321–3323.

[13] Weir, B. S. and Hill, W. G. (2002). Estimating F-Statistics. *Annual Review of Genetics*, **36**, 721–750.