

Inference of the distribution of fitness effects of mutations is affected by SNP filtering methods, sample size and population structure

Bea Andersson¹, Wei Zhao², Benjamin Haller³, Ake Brännström¹, and Xiao-Ru Wang²

¹Umeå Universitet

²Umeå University

³Cornell University

May 10, 2023

1 **Inference of the distribution of fitness effects of mutations is affected by SNP filtering methods,**
2 **sample size and population structure**

3 Bea Angelica Andersson^{1*}, Wei Zhao^{1*}, Benjamin C. Haller², Åke Brännström^{3,4,5}, Xiao-Ru Wang¹

4
5 ¹Department of Ecology and Environmental Sciences, Umeå University, Sweden

6 ²Department of Computational Biology, Cornell University, USA

7 ³Department of Mathematics and Mathematical Statistics, Umeå University, Sweden

8 ⁴Advancing Systems Analysis Program, International Institute for Applied Systems Analysis,
9 Laxenburg, Austria

10 ⁵Complexity Science and Evolution Unit, Okinawa Institute of Science and Technology
11 Graduate University, Japan

12

13 *These authors contributed equally to this study.

14 Correspondence to Xiao-Ru Wang, xiao-ru.wang@umu.se, phone +46 907869955

15

16

17

18 Total word counts for the main body of the text: 7367 (excluding references)

19 Word counts for Introduction: 924

20 Word counts for Materials and Methods: 1988

21 Word counts for Results: 2297

22 Word counts for Discussion: 1547

23 Word counts for Conclusion: 258

24 Number of Figures: 5 (all in color)

25 Number of Tables: 1

26 Supporting information: 2 table, 1 figure

27

28 **Abstract**

29 The distribution of fitness effects (DFE) of new mutations has been of interest to evolutionary
30 biologists since the concept of mutations arose. Modern population genomic data enable us to
31 quantify the DFE empirically, but few studies have examined how data processing, sample size
32 and cryptic population structure might affect the accuracy of DFE inference. We used simulated
33 and empirical data (from *Arabidopsis lyrata*) to show the effects of missing data filtering,
34 sample size, number of SNPs and population structure on the accuracy and variance of DFE
35 estimates. Our analyses focus on three filtering methods – downsampling, imputation and
36 subsampling – with sample sizes of 4 ~ 100 individuals. We show that (1) the choice of missing-
37 data treatment directly affects the estimated DFE, with downsampling performing better than
38 imputation and subsampling; (2) the estimated DFE is less reliable in small samples (<8
39 individuals), and becomes unpredictable with too few SNPs (<5000); and (3) population
40 structure may skew the inferred DFE toward more strongly deleterious mutations. We suggest
41 that future studies should consider downsampling for small datasets, and use samples larger
42 than 4 (ideally larger than 8) individuals, with more than 5000 SNPs in order to improve the
43 robustness of DFE inference and enable comparative analyses.

44

45 **Key words:** DFE, missing-data treatment, population structure, sample size, SLiM simulation.

46

47 **1 INTRODUCTION**

48 The *distribution of fitness effects* (DFE) of new mutations can be described as the probability
49 that a new mutation will have a specific effect on the fitness of an individual. This probability
50 distribution affects the accumulation of genetic variation and can thus directly impact the
51 evolutionary trajectory of organisms (Bataillon & Bailey, 2014; Keightley & Eyre-Walker,
52 2007; Ohta, 1992). Understanding the DFE is integral to understanding molecular evolution
53 and remains an important focus in modern evolutionary theory (Chen et al., 2020; Halligan &
54 Keightley, 2009; Kimura, 1968; Ohta, 1973). To date, the arguably most popular methods of
55 inferring the DFE are based on contrasting frequencies of putatively neutral and selected
56 polymorphisms presented as a site frequency spectrum (SFS), describing how commonly
57 mutations of different frequencies occur in a population (Gutenkunst et al., 2009; Keightley &
58 Eyre-Walker, 2007; Kim et al., 2017; Tataru & Bataillon, 2019). Since the SFS can be affected
59 by both neutral and selective processes, most methods use the SFS of synonymous mutations
60 to estimate a demographic model representing the effects of population size changes and genetic
61 drift. Meanwhile, the SFS of non-synonymous mutations are assumed to be shaped by both
62 neutral and selective processes, and can therefore be used to estimate the DFE of non-neutral
63 mutations after demography and drift have been accounted for (Boyko et al., 2008; Huang et
64 al., 2021; Keightley & Eyre-Walker, 2007; Kim et al., 2017; Schneider et al., 2011; Tataru &
65 Bataillon, 2019). However, factors other than demography and selection may also affect the
66 shape of the SFS and thus the estimated DFE.

67 First, SFS-based DFE inferences require that datasets contain no missing sites – all
68 individuals must have complete data for all loci that are to be analysed. Since sequencing
69 techniques are imperfect, such datasets are uncommon (probably non-existent) in empirical
70 population genomics. As a result, missing-data treatment is an essential first step of data
71 processing. To obtain a complete dataset, the data are treated either by filtering out some portion
72 of the data (sub- or downsampling), or filling in the “gaps” using an algorithm such as
73 imputation, see section 2.2 *Missing-data treatment methods*). Depending on how the treatment
74 is performed, there is a risk of altering the relative allele frequencies in the dataset, yielding
75 misleading results (Johri et al., 2021; Larson et al., 2021). Recent studies on DFE have applied
76 different data processing methods; for example see Hämälä & Tiffin (2020) for imputation, and
77 Gossmann et al. (2010) for downsampling. However, it is unknown whether and how the
78 different methods influence DFE estimates.

79 Second, the sizes of datasets used in published DFE studies vary enormously, from as
80 few as two to several hundred individuals (Chen et al., 2017; Hämälä & Tiffin, 2020). The SFS

81 is highly sensitive to sample size, but the minimum number required to achieve stable DFE
82 estimates remains undetermined (but see Kutschera et al. 2020). Similarly, the number of
83 polymorphic sites necessary for reliable DFE estimation is largely unknown. While some
84 studies of model species use whole genome sequencing with millions of single nucleotide
85 polymorphisms (SNPs) available for analysis (Hämälä & Tiffin, 2020), others may only include
86 a few hundred SNPs (Eyre-Walker & Keightley, 2009; Gossmann et al., 2010). Therefore,
87 investigating the impact of sample size (both the number of individuals and sites/SNPs) on DFE
88 estimates is crucial for reliable and accurate DFE estimation.

89 Finally, most methods of SFS-based DFE estimation first estimate a Wright-Fisher
90 demographic model from the neutral variation in order to control for neutral factors affecting
91 the SFS (Keightley & Eyre-Walker, 2007; Tataru & Bataillon, 2019). Such models assume that
92 mating occurs at random in panmictic populations, even though complete absence of population
93 structure is likely rare in wild samples. For example, sampling from a large area is preferred for
94 drawing general conclusions about population genetic dynamics, but it increases the likelihood
95 of including genetic structure in the sample (Perez et al., 2018; Zhao et al., 2020). If cryptic
96 genetic clusters are unwittingly included, the demographic model estimated from the data
97 would not fulfil the assumptions underlying the Wright-Fisher model, and subsequent DFE
98 estimates might be biased. However, population stratification has not to our knowledge been
99 examined as a potential factor affecting the accuracy of DFE inference.

100 In this study we test whether and how data processing methods, sample size, SNP
101 number and population structure influence the results of DFE inference, to raise awareness of
102 their potential confounding effects. We used whole genome re-sequencing data from two
103 populations of *Arabidopsis lyrata* (subsp. *petraea*) to create multiple datasets (Fig. 1) with (1)
104 three different methods of missing-data treatment – downsampling, imputation and
105 subsampling – under different filtering thresholds; (2) different numbers of randomly sampled
106 individuals and sites; and (3) samples with induced population stratification, to be contrasted
107 with uniform, single populations. Then, we conducted forward simulation in SLiM 4.0 (Haller
108 & Messer, 2023) to create a population with a known DFE that matches DFEs estimated in *A.*
109 *lyrata*. Using this known DFE, we evaluate the accuracy of DFE estimates resulting from the
110 different data manipulations. By contrasting the results obtained from the different procedures,
111 we aim to answer the following questions: (1) Do data processing methods and missing-data
112 filtering thresholds affect DFE estimation, and if so, how? (2) How many individuals and SNPs
113 are needed to reach an accurate DFE estimate? and (3) Does population structure affect the
114 DFE, and if so, how? Our results illustrate the importance of careful consideration of all steps

115 in genomic data processing and analysis, both when performing DFE inference and when
116 interpreting its results.

117

118 **2 MATERIALS AND METHODS**

119 **2.1 Genomic dataset and basic quality control**

120 We downloaded the whole genome re-sequencing data for two populations of the perennial,
121 diploid obligately outbreeding *Arabidopsis lyrata* subsp. *petraea*, 29 individuals from Austria
122 and 16 individuals from Norway, from the NCBI SRA database (Table S1). The quality of the
123 sequence reads was first assessed with FastQC
124 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Adapter sequences and low-
125 quality bases were removed using fastp v0.23.0 (Chen et al., 2018) with the parameters “-q 20
126 -l 36 --cut_front --cut_tail -c”. Clean reads were mapped to the *A. lyrata* v.1.0 genome
127 (<https://plants.ensembl.org/>) using the BWA-MEM algorithm with default parameters (Li,
128 2013). PCR duplicates were removed using Picard MarkDuplicates ([http://broadinstitute.
129 github.io/picard/](http://broadinstitute.github.io/picard/)). Reads around putative insertions and deletions were locally realigned using
130 RealignerTargetCreator and IndelRealigner in the Genome Analysis Toolkit (GATK v.3.7-0;
131 (Van der Auwera et al., 2013). Variants were called using the SAMtools and BCFtools pipeline
132 as described in (Li, 2011). Several filtering steps were performed to minimize genotyping
133 errors: indels and SNPs with mapping quality (MQ) <30 were removed, genotypes with
134 genotype quality (GQ) <20 or read depth (DP) <5 were masked as missing, and all SNPs with
135 a missing rate above 50% or allele number above 2 were removed. After these basic filtering
136 steps, a total of 122,432,856 sites (including invariant sites) were retained in the 45 samples for
137 the following analyses.

138

139 **2.2 Missing-data treatment methods**

140 Missing genotypes are common in genomic datasets and should be eliminated before generating
141 an SFS. We tested three methods to treat missing values on the same original datasets –
142 *downsampling*, *imputation* and *subsampling* (Fig. 1a, Fig. 2), and then compared the DFE
143 inferred from each resulting dataset using bootstrapped 95% confidence intervals (CIs).

144 *Downsampling* is performed by randomly selecting n genotypes at each site without
145 replacement (Keightley & Eyre-Walker, 2007); sites with fewer than n genotypes available are
146 removed. A 75% downsampling threshold in a sample size of 100 individuals means that 75
147 random genotypes are sampled at each site (Fig. 2). Sites that contain < 75 genotypes are
148 removed. In this study, we applied downsampling at thresholds 75%, 66%, and 50% on both

149 Austrian and Norwegian datasets. The same set of sites were kept and analyzed in both
150 populations, making direct comparisons of the DFE between populations possible.
151 Downsampling was performed using a Python script available on Dryad (Papadopoulou &
152 Knowles, 2015) with minor modification ([https://github.com/hui-liu/Bioinformatics-
153 Scripts/blob/master/Scripts/Python/sampleDownMSFS_Hui_final.py](https://github.com/hui-liu/Bioinformatics-Scripts/blob/master/Scripts/Python/sampleDownMSFS_Hui_final.py)).

154 *Imputation* refers to the statistical inference (“filling in”) of missing genotypes using
155 the available linkage information from successfully genotyped samples (Fig. 2). We tested
156 threshold 70%, 80% and 90% on the *A. lyrata* datasets (i.e. excluding sites with less than 70%,
157 80% and 90% genotype information available), and filled in the missing genotypes at all other
158 sites using Beagle v5.1 (Browning et al., 2018) with default parameters. We performed
159 imputation using all individuals from both populations, as imputation accuracy tends to increase
160 with sample size, as shown by previous studies (Pook et al., 2020).

161 *Subsampling* works in two steps: 1) Individuals that are missing more than a prescribed
162 fraction of their genotype information are excluded, and 2) for the individuals remaining, any
163 site with a missing genotype is removed (Fig. 2). This means that the size of a subsampled
164 dataset is highly dependent on the individual missing rates and the distribution of missing data
165 across the genome. We first calculated the missingness on a per-individual basis using the
166 parameter “--missing-indv” in VCFtools (Danecek et al., 2011). We then extracted the
167 individuals that had missing rates below the threshold value using “--keep”, and finally, we
168 removed all sites containing missing genotypes by setting the parameter “--max-missing 1” in
169 VCFtools. In the *A. lyrata* dataset, we tested four maximum missing rates per individual – 10%,
170 15%, 20% and 25% (Note: no individual had more than 25% missing data). Note that with
171 higher subsampling thresholds, more individuals but fewer sites are retained (Fig. 1a).

172

173 **2.3 Sample size and SNPs number**

174 To decouple the potential effects of the number of individuals and/or sites on DFE estimation,
175 we randomly sampled 4, 8, 12, 16, 20, 24 or 29 (all) individuals and/or 1K, 10K, 100K, 1M,
176 10M or 55.0M sites from the Austrian population subsampled at a maximum missing rate of
177 25% per individual (Fig. 1). To investigate the effect of sample size, we kept all sites and
178 compared samples with different numbers of individuals (4 - 29). Conversely, to investigate the
179 effect of the number of SNPs, all 29 individuals were kept and a randomly chosen subset of 1K
180 to 10M sites were extracted. Finally, the same subsets of 1K to 10M sites were extracted from
181 a dataset with only 4 individuals. By comparing the DFEs from 4 vs. 29 individuals for each

182 set of sites, we could see the combined effects of the number of individuals and sites on the
183 estimated DFE and confidence intervals (Fig. 3).

184

185 **2.4 Manipulating population structure**

186 To gain an overview of the genetic differentiation between the Austrian and Norwegian
187 populations, we performed a principal component analysis (PCA) on the 45 sampled individuals
188 using Eigensoft v.6.1.4 (Price et al., 2006). The dataset was filtered at a maximum missing rate
189 of 20% per site and a minor allele frequency (MAF) ≥ 0.05 , retaining 3,921,575 SNPs for the
190 PCA. To investigate whether population structure affects DFE estimates, we randomly selected
191 three different subsets (labelled a, b and c) of 10 and 15 individuals from each of the Austrian
192 and Norwegian populations, imputed at an 80% threshold. Single sets from each population
193 were then combined to form 12 new merged populations with four different configurations (Fig.
194 1c): 10 Austrian + 10 Norwegian individuals, 10 Austrian + 15 Norwegian individuals, 15
195 Austrian + 10 Norwegian individuals, and 15 Austrian + 15 Norwegian individuals, each with
196 three replicates. We then estimated the DFE for each subset as well as all merged samples.

197 Using the single and merged datasets we investigated 1) the effect of sample choice
198 within a geographic population on DFE, by comparing the three replicate subsets from a single
199 population (e.g. replicates *a* vs. *b* vs. *c* of subset Aus10), 2) the effect of each geographic
200 population on the merged population, by comparing the DFE of the merged population to each
201 of the contributing populations (e.g. replicate *c* of merged population Aus10+Nor15 vs.
202 replicate *c* of subsets Aus10 and Nor15), and 3) the effect of population differentiation (F_{ST}) on
203 DFE in the merged population. The weighted F_{ST} between the two contributing subsets in each
204 merged population was calculated using VCFtools.

205

206 **2.5 DFE analyses**

207 We used DFE-alpha (Eyre-Walker & Keightley, 2009), a software that uses a maximum-
208 likelihood approach to determine the shape of the DFE of nonsynonymous mutations. In the
209 simplest model, DFE-alpha assumes that mutations at synonymous sites are selectively neutral
210 and that all non-synonymous mutations are deleterious. DFE-alpha first estimates a simple
211 demographic model using the SFS of neutral mutations to represent the effect of drift. We
212 modelled the effect of recent demographic change on neutral SFS by assuming one step
213 population size change and inferred the fitness of new deleterious mutations at the selected sites
214 from a gamma distribution while simultaneously fitting the estimated parameters for the
215 demographic model. The estimated fitness effects of new mutations are scaled by effective

216 population size N_e and selection coefficient s as $N_e s$, and divided into four categories:
217 *effectively neutral* ($0 < -N_e s \leq 1$), *slightly deleterious* ($1 < -N_e s \leq 10$), *moderately deleterious*
218 ($10 < -N_e s \leq 100$) and *strongly deleterious* ($-N_e s > 100$). The DFE is presented as the proportion
219 of nonsynonymous mutations that is expected to fall into each of these categories.

220 We generated a folded SFS for a class of putatively neutral reference sites (4-fold
221 degenerate sites) and a class of selected sites (0-fold degenerate sites) for each dataset. We
222 modelled the effects of recent demographic change on the 4-fold sites SFS by assuming a single
223 population size change event and inferred the fitness of new deleterious mutations at the 0-fold
224 sites from a gamma distribution. The 95% CIs for all DFE estimates were calculated by
225 bootstrapping 0-fold and 4-fold sites with replacement for 99 iterations. We performed
226 bootstraps using 999 and 99 iterations in 9 samples and found no discernible difference in CI
227 size; all reported CIs are thus based on 99 iterations.

228

229 **2.6 Simulations in SLiM**

230 To validate the effects of filtering methods and sample size on DFE estimates, we used SLiM
231 4.0 to simulate a population with a known DFE, represented by a gamma distribution with shape
232 (β) and mean (Es) parameter values matching the DFE estimated in *A. lyrata*. The simulation
233 consisted of a population of 10,000 outcrossing individuals with a genome size of 5 million
234 sites on one contiguous chromosome, and a uniform recombination rate of 4×10^{-8} (Hämälä &
235 Tiffin, 2020). New mutations occurred at a mutation rate of 5.6×10^{-8} and were drawn from a
236 deleterious DFE with a gamma distribution with $\beta=0.1$ and $Es=-100$. The population state at
237 60,000 generations was saved as a .trees file, at which point the effective population size N_e had
238 stabilized around 100 individuals with around 60,000 segregating deleterious mutations. A
239 neutral burn-in and segregating neutral mutations were then added with recapitation and
240 overlaid mutations according to SLiM 4.0 (Haller et al., 2019). After adding neutral mutations,
241 non-segregating sites (selected or neutral) were added between SNP positions and randomly
242 assigned as either selected (20%) or neutral (5%) to approximate the 0-fold and 4-fold ratios in
243 the empirical *A. lyrata* dataset. A VCF file with 1000 randomly sampled individuals was created
244 from the dataset and used in subsequent analyses with DFE-alpha.

245 To get a baseline accuracy for DFE-alpha, 10 replicates of 100 individuals (the
246 maximum size supported by DFE-alpha) from the simulated dataset were analysed, and the
247 estimation error compared to the known DFE was in each case assessed as the Earth Mover's
248 Distance (see below). To investigate the effects of filtering methods, 15% of the sites in each

249 individual in one set of 100 individuals were masked as missing. This dataset was filtered with
250 a) downsampling at a threshold of 85%, b) imputation at a threshold of 85%, or c) subsampling
251 at a threshold of 15%. However, the subsampled dataset retained no 4-fold SNPs in the SFS
252 after filtering, making DFE estimation impossible. We thus instead sampled four replicates of
253 4, 8, 12, 16, 20, 24 and 50 individuals from the 15% missing dataset, and applied subsampling
254 at 100% (i.e. all sites with missing data were excluded). The same sets of sample sizes were
255 then extracted from the downsampled and imputed datasets to compare the accuracy of the
256 different methods while controlling for the effect of sample size. To directly investigate the
257 effect of sample size and SNP number, 10 replicates of 4, 8, 12, 16, 20, 24, 50 and 100
258 individuals were extracted from the datasets with no missing data and analysed with DFE-alpha
259 (Fig. 4a-d).

260 With the DFE associated with the simulated datasets being known, the accuracy of
261 estimated DFE was assessed by comparing them to the known DFE using Earth Mover's
262 Distance (EMD) implemented in the *transport* package in R (Schuhmacher et al., 2019). EMD
263 quantifies the dissimilarity between two distributions as the "work" required to change one
264 distribution to the other, thus taking into account the amount of overlap. In contrast to the widely
265 used Kolmogorov-Smirnov (KS) distance, EMD is not limited by an upper bound, enabling it
266 to more accurately capture substantial differences between distributions. Additionally, EMD is
267 better suited for gauging distances between distributions with long tails. The EMD was
268 evaluated within the range $-10^5 < s < -10^{-3}$ where s represents the selection coefficient for
269 each mutation, in increments of 10^{-3} . Higher EMD values signify a poorer fit between the
270 estimated and true distribution, thus indicating a less accurate result. The EMD values of each
271 dataset was plotted against the number of individuals and SNPs with a regression line to
272 illustrate the relationship.

273

274 **3 RESULTS**

275 **3.1 The effect of missing-data treatments on DFE in *A. lyrata***

276 *Downsampling.* The datasets downsampled to 50%, 66% and 75% of the genotypes per site
277 retained 105.7M, 99.5M, and 95.0M sites, respectively, for both *A. lyrata* populations (Table
278 1). The Austrian datasets contained 15, 19 and 22 "individuals" and 1.39M, 1.46M and 1.47M
279 SNPs for the three thresholds, while the Norwegian population kept 8, 11, and 12 "individuals"
280 and 374K, 366K and 341K SNPs, respectively. The DFE in the Norwegian datasets differed
281 significantly from that of the Austrian population in that neutral mutations were more frequent
282 (31~33%), while slightly (8~9%) and moderately (10~12%) deleterious mutations were less

283 frequent, but the proportion of strongly deleterious mutations was similar (45~51%) (Table 1).
284 Additionally, the impact of filter thresholds from 50% to 75% on the three deleterious groups
285 of mutations in the two populations showed inverse patterns, e.g. strongly deleterious mutations
286 increased with the threshold in the Norwegian population but decreased in the Austrian
287 population. While the estimated DFE varied between populations by 1~10 percentage points
288 under the same method and threshold, it also varied by up to 5 percentage points among the
289 downsampling thresholds within each population.

290

291 *Imputation.* The imputed datasets retained all individuals (i.e. 29 Austrian and 16 Norwegian
292 individuals), and 103.4M, 97.9M and 86.3M sites at the 70%, 80% and 90% thresholds,
293 respectively. In the Austrian population, 1.69M, 1.63M and 1.44M SNPs were included, while
294 399K, 365K and 341K SNPs in the Norwegian population, at the three thresholds, respectively.
295 Increasing the threshold from 70% to 90% only caused 2~4 percentage points of variation in
296 each category of mutations (Table 1). Across both populations, the DFE were stable among
297 imputation thresholds, with the Austrian population displaying slightly larger variance.

298

299 *Subsampling.* In the subsampling trial, we applied four different thresholds, allowing a
300 maximum of 10%, 15%, 20% and 25% missing genotypes per individual. In the Austrian
301 population, a strict threshold of 10% missing data left 8 individuals, 97.4M sites and 844K
302 SNPs in the dataset, while a relaxed 25% threshold preserved all 29 individuals with 55.0M
303 sites and 609K SNPs (Note: increasing the missing rate from 20% to 25% only added one more
304 individual) (Table 1). Increasing the missing threshold from 10% to 25% decreased the
305 estimated neutral mutations from 23% to 20%, and the strongly deleterious mutations from 49%
306 to 32%, while the slightly and moderately deleterious mutations increased from 11% to 17%
307 and from 17% to 30%, respectively. Overall, change the threshold from 10% to 15% induced
308 the largest difference in the DFE of all stepwise increases (3~8 percentage points of difference
309 in all categories).

310 In the Norwegian population, the dataset filtered with a missing rate of 10% included
311 only 2 individuals with 109.4M sites and 249K SNPs. At this level, the DFE was estimated to
312 7% neutral, 86% slightly deleterious, 6% moderately deleterious and no strongly deleterious
313 mutations. Increasing the threshold to 15% increased the number of individuals to 15, retaining
314 80.0M sites and 172K SNPs, and shifted the DFE to 28% neutral, 8% slightly deleterious, 10%
315 moderately deleterious and 53% strongly deleterious mutation. Further relaxing the missing
316 rate to 20% and 25% included one more individual (16 total) and had little effect on the DFE

317 compared to the dataset filtered at 15% (Table 1). Overall, the Austrian population displayed
318 up to 17 percentage points of difference between thresholds, while the Norwegian population
319 displayed up to 79 percentage points of difference when including the dataset filtered at 10%
320 missing data.

321

322 **3.2 The effect of sample size and sites on DFE**

323 We subsampled the Austrian population of *A. lyrata* into 4, 8, 12, 16, 20 and 24 individual sets,
324 each containing 211K, 320K, 357K, 426K, 512K and 557K SNPs, respectively, from the
325 complete dataset of 29 individuals containing 609K SNPs (Fig 3b). We found that decreasing
326 the sample size from 29 to 4 substantially increased the proportion of strongly deleterious
327 mutations from 32% to 45%, while it decreased the proportion of slightly deleterious mutations
328 from 17% to 13% and moderately deleterious mutations from 30% to 20%. Neutral mutations
329 changed only slightly (from 20% to 22%) (Fig. 3a). The partition of DFE remained stable with
330 sample sizes of 8 and upward (≤ 1 percentage points fluctuation). The 95% CIs remained similar
331 and narrow (0.5~4%) in all samples.

332 In the second trial, we randomly sampled 1K, 10K, 100K, 1M and 10M sites in the 29
333 individuals (with 55.0M sites, 609K SNPs), resulting in 10, 109, 1115, 11.1K, and 111K SNPs,
334 in each dataset, respectively. We found that the DFE estimates became increasingly unstable
335 with decreasing the number of sites: the datasets with fewer than 1M sites (11.1K SNPs) showed
336 a large variation in DFE values (8–50 percentage points; Fig. 3b). Notably, a decrease in the
337 number of sites brought a simultaneous increase of the width of the 95% CIs, in a manner not
338 seen when decreasing the numbers of individuals (Fig. 3a vs. 3b). At 1K sites (10 SNPs), the
339 95% CIs for the three deleterious categories covered 98~100% of the entire range of possible
340 values, indicating low confidence in where the true values lie. At 10K sites (109 SNPs) the CIs
341 shrunk but were still large, covering between 34~71% of the possible values. On average, each
342 tenfold decrease in the number of sites increased the size of the bootstrapped 95% CIs 2.5 times.

343 In the third trial, we examined the effect of sites in a small sample of 4 individuals.
344 The sites chosen were the same as those in the second trial, although the set of 1K sites included
345 too few SNPs to be evaluated and was not shown in Fig. 3c. The datasets with 10K, 100K, 1M,
346 10M and all 55.0M sites had 43, 391, 3821, 38.6K and 211K SNPs, respectively. At 10K sites,
347 the DFE in the 4-individual set was drastically different from the 29 individuals. Furthermore,
348 the 95% CIs of neutral, and slightly and moderately deleterious mutations increased by 18~81%
349 in the 4-individual relative to the 29-individual dataset. The CIs for strongly deleterious
350 mutations shrank somewhat in the 4-individual dataset but was still large and spanned 66% of

351 the range of possible values. The DFE estimates at 100K sites and above in 4-individual datasets
352 were very similar (≤ 1 percentage points of difference) to the second trial using 29 individuals
353 (Fig. 3c vs. 3b), but the 95% CIs approximately doubled for the three classes of deleterious
354 mutations.

355

356 **3.3 Accuracy of DFE-alpha in SLiM simulated data**

357 To determine which missing-data treatment and sample sizes produced the least error and thus
358 approximated the true DFE most accurately, we conducted SLiM simulations with a known
359 DFE. The simulation produced a dataset with 1000 individuals and 29,944 SNPs. Using 10
360 replicate samples of 100 individuals, each containing $\sim 15,500$ SNPs, the DFE was estimated to
361 between 29~31% neutral, 8-10% slightly deleterious, 10~13% moderately deleterious and
362 48~52% strongly deleterious mutations; the true DFE should be approximately 30% neutral,
363 9% slightly deleterious, 11% moderately deleterious and 50% strongly deleterious mutations,
364 meaning an error of $\pm 1\sim 2\%$ can be expected with this dataset in optimal conditions. The β and
365 E_s parameters of the gamma distributions ranged between $0.097 \sim 0.128$ and $-276 \sim -33$,
366 respectively, yielding error values ($EMD \times 10^7$) between $3.5 \sim 20.5$ (Fig. 4e). These values are
367 used as reference for the “maximum” accuracy of DFE-alpha for the simulated dataset.

368 To evaluate the effect of filtering methods, we used four replicates of 4, 8, 12, 16, 20,
369 24 and 50 individuals and excluded all missing sites in each sample, which mimics the effect
370 of subsampling at different thresholds. In order to compare these results to downsampling and
371 imputation, the same sample sizes were extracted from the downsampled and imputed datasets
372 created at 85% threshold from the full dataset. At a sample size of 4 individuals, all three
373 methods performed roughly equally well (average EMD was 33.7, 36.4 and 36.5 for
374 downsampling, imputation and subsampling, respectively. Fig. 4b-d,e, Table S2), but
375 subsampling tended to slightly underestimate the proportion of slightly and moderately
376 deleterious mutations (by up to 5% and 7%, respectively), and overestimate strongly deleterious
377 mutations (by up to 11%). Downsampling gave the most accurate results based on the average
378 EMD across all sample sizes above 8 individuals (Fig. 4b,e). Imputation performed slightly
379 worse in all samples except 8 individuals (Fig. 4c,e). Both downsampling and imputation
380 produced results within 1~3% of the range of the reference set at all sample sizes above 4
381 individuals. Subsampling, however, produced highly variable and noticeably less accurate
382 results even at higher sample sizes (Fig. 4d,e). For example, the 4 replicates of 24 individuals
383 produced EMD values between $3.7 \sim 18.8$ for downsampling, $12.3 \sim 47.4$ for imputation and
384 $31.2 \sim 112.8$ for subsampling (Table S2). We found that subsampling produced the most

385 accurate results at an intermediate sample size (e.g. 16 individuals; EMD from 1.7 to 56.1) and
386 became less accurate at sample sizes where fewer SNPs were retained (e.g. 50 individuals with
387 5 SNPs remaining; Fig. 4b, Table S2).

388 Our simulated data verified the trends observed in the empirical data, showing that
389 increased sample size correlated with lower error in DFE estimates when the number of SNPs
390 is not a limiting factor. In the datasets of 4, 8, 12, 16, 20, 24 and 50 individuals (10 replicates
391 of each) with no missing genotypes, the EMD values were the largest in samples of 4 and 8
392 individuals, stabilized around 12 ~ 24 individuals, and then decreased further in 50 individuals
393 to a level similar to that in the 100 individuals (Fig. 4e). Linear regression in these datasets
394 showed that DFE estimation error (EMD) was negatively correlated with number of individuals
395 ($p = 0.00179$, $R^2 = 0.1182$), and even more strongly correlated with the number of SNPs in the
396 dataset ($p = 6.38 \cdot 10^{-6}$, $R^2 = 0.2311$) (Fig 4f). An even stronger negative correlation between
397 EMD and SNP number was seen when the four replicates of 4 ~ 50 individuals from the
398 downsampled, imputed and subsampled datasets were analysed with a joint linear regression (p
399 $= 1.11 \cdot 10^{-9}$, $R^2 = 0.3658$) (Fig 4b). Datasets with few SNPs also displayed larger 95% CIs while
400 the number of individuals had a minor effect on CI size (Fig. 4a-d, Table S2), similar to what
401 was observed in the empirical datasets.

402 In summary, applying different filtering methods and thresholds affected the final data
403 matrix size (number of individuals and SNPs) and subsequent DFE estimates. Imputation and
404 downsampling produced similar and less variable DFE results than subsampling, and
405 downsampling appeared more accurate than imputation for the simulated samples used. Further,
406 higher numbers of individuals and SNPs both increased accuracy of the results, especially at
407 very low sample sizes (4 ~ 8 individuals, <5000 SNPs).

408

409 **3.4 The effect of population structure on DFE**

410 The PCA of the 45 samples from Austria and Norway showed a distinct separation of the two
411 populations along PC1 (which explained 24.7% of the total genetic variance), and separation
412 of the Austrian population into four visible clusters along PC2 (which explained 7.3% of the
413 total genetic variance) (Fig. S1). The weighted F_{ST} between the two populations was 0.228,
414 while the F_{ST} among the four Austrian clusters was relatively small as 0.073. To understand the
415 effect of merging genetically distinct populations on the estimated DFE, we created 12 merged
416 populations with contributions of 10 or 15 individuals from Austria and Norway, with three
417 subsets of each population (Fig. 1c). We then calculated the F_{ST} between the contributing
418 subsets to evaluate how the degree of population stratification in a sample affects the joint DFE

419 estimate. We first examined the DFE in the unmerged replicate samples of 10 and 15 individuals
420 from the two populations. Among the replicates of 10 individuals from the Austrian population,
421 a maximum difference of 2, 3, 7 and 6 percentage points were observed in the neutral, slightly,
422 moderately and strongly deleterious mutations. By comparison, no mutation category varied by
423 more than 2 percentage points in the samples of 15 individuals. Comparably stable DFE
424 estimates were observed in the Norwegian samples, with variation in the range of 0, 2, 3 and 4
425 percentage points for the four categories of mutations in samples of 10 individuals, and less
426 than 1 percentage point of a difference among replicates of 15 individuals (Fig. 4a). However,
427 the DFE estimates were markedly different between the two geographical populations, e.g.
428 neutral mutations shifted up by an average of 9 percentage points while the slight and moderate
429 mutations shifted downwards in Norway compared to Austria. The estimated proportions of
430 strongly deleterious mutations were similar in the two populations.

431 With this population-specific DFE in mind, we then examined the differences between
432 the merged samples and their respective contributing single population subsets. In most cases,
433 the estimated DFE values for the merged samples were in-between the DFE estimates of the
434 contributing subsets, but not always perfectly intermediate (Fig. 5a). The estimated weighted
435 F_{ST} values between the pairs of contributing subsets ranged from 0.218 to 0.263 (mean F_{ST}
436 between 0.085 and 0.131). These estimates are largely in line with previous studies, where
437 mean F_{ST} across European populations of *A. lyrata* ranges between 0.06-0.09 (Marburger et al.,
438 2019). Plotting the weighted F_{ST} against the estimated DFE in the merged populations showed
439 an apparent relationship (Fig. 5b). Using linear regression, F_{ST} was correlated with the
440 proportion of slightly ($R = -0.61$, $p = 0.037$), moderately ($R = -0.60$, $p = 0.038$) and strongly
441 deleterious mutations ($R = 0.66$, $p = 0.02$), but not with that of neutral mutations ($p = 0.17$).
442 These results show that population structure had a significant effect on the deleterious portion
443 of the DFE, with higher F_{ST} potentially driving up the estimated proportion of strongly
444 deleterious mutations and reducing the estimates of the less deleterious classes.

445

446 **4 DISCUSSION**

447 **4.1 Methods of missing-data treatment affect DFE results**

448 Missing-data treatment is the first step in any genomics analyses. Using simulated data with
449 known DFE we were able to evaluate the accuracy of different filtering methods in recovering
450 the true DFE. We found the dataset with no missing data produced the most accurate result,
451 followed by downsampling, then imputation, and then subsampling. The number of SNPs in
452 the downsampled and imputed datasets were similar in all samples, suggesting that any

453 difference in performance between the two methods is likely due to imputation affecting the
454 shape of the SFS in a non-random manner. The assumption that deleterious mutations appear
455 as low-frequency alleles in the SFS, in combination with the relatively small sample sizes used
456 in the tests, makes an SFS-based analysis highly reliant on those low-frequency categories,
457 especially singleton SNPs. Low frequency alleles thus display much higher error rates than
458 higher-frequency alleles in imputation procedures (Pook et al., 2020).

459 Filtering with subsampling produced the least accurate estimates on average. Since
460 increasing the number of individuals in the subsampled dataset decreases the number of sites,
461 this filtering method's performance is thus affected by sample size in two ways, both the number
462 of individuals and the number of SNPs available. This effect is expected to be especially strong
463 in datasets where the distribution of missing data is random (as was the case in our simulated
464 datasets), where a highly dissimilar pattern of missing data across individuals excludes a large
465 number of sites by subsampling. This pattern was not as strong in the empirical datasets where
466 the missing data across individuals was more similar. Thus intermediate sample sizes of
467 individuals are preferable for this method.

468 The array of tested filtering thresholds on the empirical datasets corroborated the trend
469 and conclusions drawn from the simulated datasets. The empirical datasets proved to be more
470 sensitive to minor changes in filtering thresholds as even slight adjustments resulted in
471 significantly different outcomes in some cases. The DFE estimates in the subsampled datasets
472 were unpredictable, both within and among populations. This is most likely a result of
473 substantial downsizing of the data matrix, since the total number of sites and SNPs were reduced
474 by 50–90% in the subsampled datasets compared to the other two methods. Downsampling and
475 imputation produced results with similar levels of variation across the different thresholds. With
476 the simulation results in mind, it could be argued that both methods are be equally valid in this
477 case, and the choice between them might depend on other conditions and computational
478 resources. As a general rule, we recommend filtering data with several thresholds to obtain an
479 overview of the variability produced by each method. This is especially important because the
480 95% CIs do not provide information about whether the filtered and subsampled dataset is
481 representative of the initial population and, as we show in this study, the differences among
482 subsets of samples from the same population can be significant

483 A cursory review of recently published DFE estimation studies shows that
484 downsampling is the most frequently used of the three methods tested here (see Castellano et
485 al., 2019; Chen et al., 2020; Gossmann et al., 2010; Liang et al., 2022; Takou et al., 2021). This
486 is not surprising, since downsampling is considerably faster than imputation, yet retains more

487 data than subsampling. Imputation methods require high quality datasets from the outset to be
488 able to make reliable predictions; datasets with high rates of missing sites and low levels of
489 genome-wide linkage disequilibrium are not ideal for this treatment. With low levels of
490 genome-wide linkage disequilibrium, the presence/absence of any given SNP is mostly
491 uncorrelated with the presence/absence of any other SNP, meaning that there are no patterns of
492 linkage disequilibrium among sites from which imputation can accurately predict the state of a
493 missing site. In such cases, downsampling might be a better choice. With the current rate of
494 improvement in both genome-wide sequence data and computing power, however, we predict
495 an increasing popularity of imputation as a data processing method in DFE estimation and other
496 population genomics analyses. We recommend prefacing any missing-data treatment with an
497 analysis of the prevalence of missing sites and the level of linkage disequilibrium to determine
498 whether imputation is the appropriate method for each dataset.

499

500 **4.2 Very small sample sizes skew the estimated DFE**

501 A review on DFE estimated in 139 plant and animal species (Chen et al., 2017), each with
502 between 2–50 chromosomes sampled, shows very different DFE distributions. We evaluated
503 the effects of the number of sampled individuals on the estimated DFE when the number of
504 sites was not a limiting factor. We found that DFE estimated from few individuals (<8) were
505 strongly skewed compared to larger sample sizes. In simulated datasets with no missing data,
506 the accuracy of the estimated DFE was highest in the largest sample (100 individuals) and
507 lowest in the smallest samples (4 and 8 individuals), and the samples with >8 individuals
508 markedly improved DFE estimates. Similarly, DFE estimates based on 4 individuals produced
509 the least accurate results using both downsampling and imputation for missing-data treatment.

510 In the empirical trials, DFE estimates between random sets of 4 individuals were rather
511 unstable in the Austrian population. In the Norwegian dataset subsampled at 10% that kept only
512 2 diploid individuals, the proportion of slightly deleterious mutations was greatly overestimated
513 compared to that of the full population size. Results stabilized with a sample size of 8 or more,
514 which is consistent with the findings from the simulated datasets. This suggests that a relatively
515 small number of individuals is needed for reliable DFE estimates when there are many sites
516 available, but that very limited sample sizes increases the risk of producing non-representative
517 results. We thus deem the potential effects of low sample size to be alarming due to their
518 unpredictable and stochastic nature, and caution against using sample sizes below 4 diploid
519 individuals (8 haploids).

520

521 **4.3 Limited sites cause high variability in DFE results**

522 Reducing the number of sites resulted in highly variable and unpredictable DFE estimates even
523 with larger sample sizes. Overall, the negative correlation was observed between the number
524 of SNPs and EMD values in the simulated datasets indicates that the accuracy of SFS-based
525 DFE estimation is limited by the number of SNPs available. This trend was also observed in
526 the empirical data, where estimates based on 1M, 10M and 55M sites in 29 individuals all
527 looked similar, but using 1K ~ 10K sites (59 ~ 571 SNPs) produced highly dissimilar results,
528 demonstrating the importance of having a sufficient number of sites and SNPs for reliable SFS-
529 based analyses. The DFE is estimated from SFS, i.e. the distribution of SNPs of different
530 frequencies in the population. Thus, the number and specific subset of SNPs directly affect the
531 resolution to which we can estimate the shape of the DFE. This would explain why the 95%
532 CIs increased in size as the number of sites decreased. At 1K ~ 10K sites, the confidence
533 intervals spanned the entire range of possible values for several of the mutational categories
534 (Fig. 3b). For these datasets, we are therefore left with no confidence that our predicted DFE is
535 close to the true DFE. If the CIs are ignored, the very different DFE estimates from subsets of
536 the same dataset could lead to different interpretations of the selection pressures acting on the
537 population. This result illustrates a clear type 1 error; the estimated DFE from our samples of
538 1K, 10K and 100K sites are not representative of the full sites and produce incorrect inferences
539 that imply differences in the underlying DFE, despite being random subsets of the same dataset.

540 Based on both the empirical and simulated trials, we conclude that DFE estimates of
541 DFE become stochastic and unpredictable with very small number of sites/SNPs, and accuracy
542 is expected to increase significantly with the number of SNPs included; at least 5K SNPs are
543 required to obtain reliable DFE estimates using DFE-alpha.

544

545 **4.4 Population structure may skew DFE estimates**

546 By combining samples from the Austrian and Norwegian populations into merged populations,
547 we were able to see how the composition of populations affects DFE estimates. One trend was
548 immediately clear: the estimated proportion of strongly deleterious mutations was higher in the
549 merged populations than in the contributing single population subsets. A high F_{ST} may skew
550 the DFE towards higher estimated proportions of strongly deleterious mutations and lower
551 proportions of slightly and moderately deleterious mutations. This correlation may not be
552 conclusive, but it indicates that population structure can indeed affect DFE and should be taken
553 into consideration when performing these analyses at a species level. Studies on DFE often
554 include multiple or combined populations to gain a global estimate that characterizes the

555 organism or species (Chen et al., 2017; Hämälä & Tiffin, 2020; Slotte et al., 2010; Zhao et al.,
556 2020). We cannot presently state that pooled samples will always skew the DFE distribution,
557 but it is advisable to estimate the DFE separately in individual populations, as well as from
558 pooled samples to evaluate any deviations caused by pooling that might inform conclusions
559 drawn from the results. A recent study developed a joint DFE approach that enables the analysis
560 of pairs of populations (Huang et al., 2021), which could be practical in examining variance of
561 DFE among populations.

562

563 **5 CONCLUSION**

564 Accurate estimation of DFE from genomic data hinges on several factors, including the number
565 of sampled individuals, the availability of sites and SNPs, and the approach employed to address
566 missing data. Our study, which utilized both empirical data and forward simulations, explored
567 all these aspects and offers guidance for experimental design of DFE estimation studies. We
568 found that downsampling is a dependable method of handling missing data, though it may still
569 impact the DFE to some extent. Imputation, while generally accurate, may be less suitable for
570 small samples (≤ 100 individuals, $< 10K$ SNPs) or when genome-wide linkage disequilibrium is
571 very low (as is often the case with highly outbreeding species). We demonstrated that DFE
572 estimates derived from datasets with less than four diploid individuals or less than 5K SNPs
573 may be unreliable due to the risk of sampling error and the limited information in the SFS.
574 Furthermore, strong population structure within samples can potentially skew DFE estimates.

575 More advanced methods of DFE estimation employ an unfolded SFS, where each SNP
576 is categorized as ancestral or derived based on an outgroup reference genome. While model
577 species can benefit from these sophisticated techniques, most studies must still rely on methods
578 utilizing the folded SFS, and frequently deal with limited sample sizes. Given the extensive
579 body of previously published work employing folded SFS, it is imperative to be able to
580 understand the expected accuracy of DFE estimates in comparative analyses. This study
581 highlights the factors that should be considered when interpreting DFE estimates, thereby
582 enhancing the reliability and relevance of future research.

583

584 **ACKNOWLEDGEMENTS**

585 Genomic data processing and analyses were performed using resources provided by the
586 Swedish National Infrastructure for Computing (SNIC), through the High Performance
587 Computing Centre North (HPC2N). This study was supported by grants from the Swedish
588 Research Council (VR) and T4F program to XRW. The authors declare no conflict of interest.

589

590 AUTHOR CONTRIBUTION

591 WZ and XRW designed the empirical study. All authors contributed to designing the simulation
592 study. BH provided support for simulations in SLiM 4.0. BA and WZ performed empirical data
593 analyses. ÅB provided statistical advice. BA, WZ and XRW wrote the manuscript draft. All
594 authors contributed to the revision of the manuscript.

595

596 DATA AVAILABILITY

597 All sequencing data are retrieved from the NCBI SRA database with accession numbers listed
598 in Supplemental Table 1. Procedures associated with the SLiM simulations are provided to
599 GitHub repository at <https://github.com/beaangelica/DFE-filtering>.

600

601 REFERENCES

- 602 Bataillon, T., & Bailey, S. F. (2014). Effects of new mutations on fitness: insights from models
603 and data. *Annals of the New York Academy of Sciences*, *1320*, 76-92.
604 doi:10.1111/nyas.12460
- 605 Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D.,
606 Lohmueller, K. E., . . . Bustamante, C. D. (2008). Assessing the evolutionary impact of
607 amino acid mutations in the human genome. *Plos Genetics*, *4*, e1000083.
608 doi:10.1371/journal.pgen.1000083
- 609 Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from
610 next-generation reference panels. *The American Journal of Human Genetics*, *103*, 338-
611 348. doi:<https://doi.org/10.1016/j.ajhg.2018.07.015>
- 612 Castellano, D., Macia, M. C., Tataru, P., Bataillon, T., & Munch, K. (2019). Comparison of the
613 full distribution of fitness effects of new amino acid mutations across great apes.
614 *Genetics*, *213*, 953-966. doi:10.1534/genetics.119.302494
- 615 Chen, J., Glemin, S., & Lascoux, M. (2017). Genetic diversity and the efficacy of purifying
616 selection across plant and animal species. *Molecular Biology and Evolution*, *34*, 1417-
617 1428. doi:10.1093/molbev/msx088
- 618 Chen, J., Glemin, S., & Lascoux, M. (2020). From drift to draft: how much do beneficial
619 mutations actually contribute to predictions of Ohta's slightly deleterious model of
620 molecular evolution? *Genetics*, *214*, 1005-1018. doi:10.1534/genetics.119.302869
- 621 Chen, S. F., Zhou, Y. Q., Chen, Y. R., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ
622 preprocessor. *Bioinformatics*, *34*, 884-890. doi:10.1093/bioinformatics/bty560

623 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . 1000
624 Genomes Project Analysis Group. (2011). The variant call format and VCFtools.
625 *Bioinformatics*, 27, 2156-2158. doi:10.1093/bioinformatics/btr330

626 Eyre-Walker, A., & Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution
627 in the presence of slightly deleterious mutations and population size change. *Molecular*
628 *Biology and Evolution*, 26, 2097-2108. doi:10.1093/molbev/msp119

629 Gossmann, T. I., Song, B. H., Windsor, A. J., Mitchell-Olds, T., Dixon, C. J., Kapralov, M. V.,
630 . . . Eyre-Walker, A. (2010). Genome wide analyses reveal little evidence for adaptive
631 evolution in many plant species. *Molecular Biology and Evolution*, 27, 1822-1832.
632 doi:10.1093/molbev/msq079

633 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring
634 the joint demographic history of multiple populations from multidimensional SNP
635 frequency data. *Plos Genetics*, 5, e1000695. doi:10.1371/journal.pgen.1000695

636 Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2019). Tree-sequence
637 recording in SLiM opens new horizons for forward-time simulation of whole genomes.
638 *Molecular Ecology Resources*, 19, 552-566. doi:10.1111/1755-0998.12968

639 Haller, B. C., & Messer, P. W. (2023). SLiM 4: Multispecies eco-evolutionary modeling.
640 *American Naturalist*. doi:10.1086/723601

641 Halligan, D. L., & Keightley, P. D. (2009). Spontaneous mutation accumulation studies in
642 evolutionary genetics. *Annual Review of Ecology Evolution and Systematics*, 40, 151-
643 172. doi:10.1146/annurev.ecolsys.39.110707.173437

644 Huang, X., Fortier, A. L., Coffman, A. J., Struck, T. J., Irby, M. N., James, J. E., . . . Gutenkunst,
645 R. N. (2021). Inferring genome-wide correlations of mutation fitness effects between
646 populations. *Molecular Biology and Evolution*, 38, 4588-4602.
647 doi:10.1093/molbev/msab162

648 Hämälä, T., & Tiffin, P. (2020). Biased gene conversion constrains adaptation in *Arabidopsis*
649 *thaliana*. *Genetics*, 215, 831-846. doi:10.1534/genetics.120.303335

650 Johri, P., Aquadro, C. F., Beaumont, M., Charlesworth, B., Excoffier, L., Eyre-Walker, A., . . .
651 Jensen, J. D. (2021). Recommendations for improving statistical inference in population
652 genomics. *Plos Biology*, 20, e3001669.

653 Keightley, P. D., & Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects
654 of deleterious mutations and population demography based on nucleotide
655 polymorphism frequencies. *Genetics*, 177, 2251-2261.
656 doi:10.1534/genetics.107.080663

657 Kim, B. Y., Huber, C. D., & Lohmueller, K. E. (2017). Inference of the distribution of selection
658 coefficients for new nonsynonymous mutations using large samples. *Genetics*, *206*,
659 345-361. doi:10.1534/genetics.116.197145

660 Kimura, M. (1968). Evolutionary rate at molecular level. *Nature*, *217*, 624-626.
661 doi:10.1038/217624a0

662 Kutschera, V. E., Poelstra, J. W., Botero-Castro, F., Dussex, N., Gennnell, N. J., Hunt, G. R.,
663 . . . Wolf, J. B. W. (2020). Purifying selection in corvids is less efficient on islands.
664 *Molecular Biology and Evolution*, *37*, 469-474. doi:10.1093/molbev/msz233

665 Larson, W. A., Isermann, D. A., & Feiner, Z. S. (2021). Incomplete bioinformatic filtering and
666 inadequate age and growth analysis lead to an incorrect inference of harvested-induced
667 changes. *Evolutionary Applications*, *14*, 278-289.
668 doi:<https://doi.org/10.1111/eva.13122>

669 Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping
670 and population genetical parameter estimation from sequencing data. *Bioinformatics*,
671 *27*, 2987-2993. doi:10.1093/bioinformatics/btr509

672 Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-
673 MEM. *arXiv:1303.3997v2 [q-bio.GN]*.

674 Liang, Y. Y., Shi, Y., Yuan, S., Zhou, B. F., Chen, X. Y., An, Q. Q., . . . Wang, B. S. (2022).
675 Linked selection shapes the landscape of genomic variation in three oak species. *New*
676 *Phytologist*, *233*, 555-568. doi:10.1111/nph.17793

677 Marburger, S., Monnahan, P., Seear, P. J., Martin, S. H., Koch, J., Paajanen, P., . . . Yant, L.
678 (2019). Interspecific introgression mediates adaptation to whole genome duplication.
679 *Nature Communications*, *10*, 5218. doi:10.1038/s41467-019-13159-5

680 Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, *246*, 96-98.
681 doi:10.1038/246096a0

682 Ohta, T. (1992). The nearly neutral theory of molecular evolution. *Annual Review of Ecology*
683 *and Systematics*, *23*, 263-286. doi:10.1146/annurev.es.23.110192.001403

684 Papadopoulou, A., & Knowles, L. L. (2015). Genomic tests of the species-pump hypothesis:
685 Recent island connectivity cycles drive population divergence but not speciation in
686 Caribbean crickets across the Virgin Islands. *Evolution*, *69*, 1501-1517.
687 doi:10.1111/evo.12667

688 Perez, M. F., Franco, F. F., Bombonato, J. R., Bonatelli, I. A. S., Khan, G., Romeiro-Brito, M.,
689 . . . Moraes, E. M. (2018). Assessing population structure in the face of isolation by

690 distance: Are we neglecting the problem? *Diversity and Distributions*, 24, 1883-1889.
691 doi:10.1111/ddi.12816

692 Pook, T., Mayer, M., Geibel, J., Weigend, S., Caverro, D., Schoen, C. C., & Simianer, H. (2020).
693 Improving imputation quality in BEAGLE for crop and livestock data. *G3-Genes*
694 *Genomes Genetics*, 10, 177-188. doi:10.1534/g3.119.400798

695 Schneider, A., Charlesworth, B., Eyre-Walker, A., & Keightley, P. D. (2011). A method for
696 inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics*,
697 189, 1427-1437. doi:10.1534/genetics.111.131730

698 Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F., Bernhard
699 Schmitzer, & Schrieber, J. (2019). transport: Computation of optimal transport plans
700 and Wasserstein distances. *R package version 0.13-0*. doi:[https://cran.r-](https://cran.r-project.org/package=transport)
701 [project.org/package=transport](https://cran.r-project.org/package=transport)

702 Slotte, T., Foxe, J. P., Hazzouri, K. M., & Wright, S. I. (2010). Genome-wide evidence for
703 efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a
704 large effective population size. *Molecular Biology and Evolution*, 27, 1813-1821.
705 doi:10.1093/molbev/msq062

706 Takou, M., Hamala, T., Koch, E. M., Steige, K. A., Dittberner, H., Yant, L., . . . de Meaux, J.
707 (2021). Maintenance of adaptive dynamics and no detectable load in a range-edge
708 outcrossing plant population. *Molecular Biology and Evolution*, 38, 1820-1836.
709 doi:10.1093/molbev/msaa322

710 Tataru, P., & Bataillon, T. (2019). polyDFEv2.0: testing for invariance of the distribution of
711 fitness effects within and across species. *Bioinformatics*, 35, 2868-2869.
712 doi:10.1093/bioinformatics/bty1060

713 Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine,
714 A., . . . DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the
715 genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*,
716 43, 11.10.11-11.10.33. doi:10.1002/0471250953.bi1110s43

717 Zhao, W., Sun, Y. Q., Pan, J., Sullivan, A. R., Arnold, M. L., Mao, J. F., & Wang, X. R. (2020).
718 Effects of landscapes and range expansion on population structure and local adaptation.
719 *New Phytologist*, 228, 330-343. doi:10.1111/nph.16619

720

721 **Figure legends**

722 **Figure 1: Experimental design**

723 We performed three sets of tests to understand their potential influence on the estimated DFE:
724 a) three procedures of missing-data treatment, b) the number of individuals and sites used, and
725 c) population structure. Each box represents a derived dataset, with the number of individuals
726 shown on top and nucleotide sites below. The study involved two populations of *Arabidopsis*
727 *lyrata* from Austria and Norway. We created merged populations with subsets of individuals
728 from Austria and Norway as specified on the left of each merged boxes (c, greyed out). The
729 estimated DFE of the merged population are compared to that of the contributing populations.

730

731 **Figure 2: Methods of missing-data treatments for SFS based analyses**

732 Illustration of the different steps involved in the three missing-data filtering methods examined
733 in this study. Each box corresponds to an individual's genotype at a site, and missing boxes
734 represent missing data for a genotype. In downsampling, step 1 excludes sites at which data is
735 missing in more than a prescribed threshold of individuals (e.g. 25%), while step 2 samples
736 genotypes without replacement from the remaining data at each site. In imputation, as in
737 downsampling, step 1 excludes sites with missing rate more than a prescribed fraction, while
738 step 2 imputes (fills in) missing data. In subsampling, step 1 excludes individuals with missing
739 data in more than a prescribed fraction of sites, while step 2 excludes all sites with missing data.

740

741 **Figure 3: Effects of number of individuals and sites on DFE**

742 DFE estimated from *Arabidopsis lyrata*, a) random samples of 4, 8, 12, 16, 20 and 24 of the 29
743 individuals of the Austrian population with 55M sites; b) all 29 individuals, c) a random sample
744 of 4 individuals with 1K, 10K, 100K, 1M, 10M and 55M sites. The complete DFE is represented
745 as percentage contribution of each of four categories of mutations: *neutral* (blue), *slightly*
746 *deleterious* (yellow), *moderately deleterious* (orange) and *strongly deleterious* (red). The DFE
747 for each sample size is represented in two ways: on the left as stacked estimated percentages of
748 the four categories of mutations, and on the right as the estimated percentage of each category
749 of mutations (black bars and light areas) together with the 95% CIs (darker coloured areas).

750

751 **Figure 4: The accuracy of DFE estimations by manipulating SLiM simulated dataset**

752 DFE estimates and 95% CIs for 4, 8, 12, 16, 20, 24 50, and a maximum of either 85 (in
753 downsampling) or 100 (in the other cases) individuals, with either a) no missing data, or 15%
754 missing data per individual and filtered with either b) downsampling, c) imputation or d)

755 subsampling. e) DFE estimation error, as represented by Earth Mover's Distance (EMD), in
756 different sample sizes without missing data (black, 10 replicates (n) per sample size), or with
757 15% missing data and filtered with either downsampling (green), imputation (red) or
758 subsampling (yellow), in four replicates each. f) DFE estimation error in samples plotted against
759 SNP number, in datasets without missing data (black) as well as with missing-data filtered by
760 downsampling (green), imputation (red) or subsampling (yellow). Linear regression lines for
761 the no missing data (black) and for all of the filtered datasets combined (brown) are displayed
762 to show the trend of EMD over SNP number in the two groups. Datasets without missing data
763 include 10 replicates of 4 ~ 100 individuals, while four replicates of 4 ~ 50 individuals are
764 included for the missing-data filtered datasets.

765

766 **Figure 5: Effect of population structure on DFE**

767 a) The estimated DFE of the Austrian (dark dots) and Norwegian (light dots) samples of
768 *Arabidopsis lyrata*, compared to merged samples (solid lines) containing both groups in
769 different combinations. The relative sample size from each population is listed along the
770 horizontal axis (bottom), as well the name of each of three replicates (top). b) Linear regression
771 of the estimated proportion of each of the four mutational categories of the DFE over the F_{ST}
772 between the merged samples, with 95% confidence intervals shown in shaded areas.

773

774

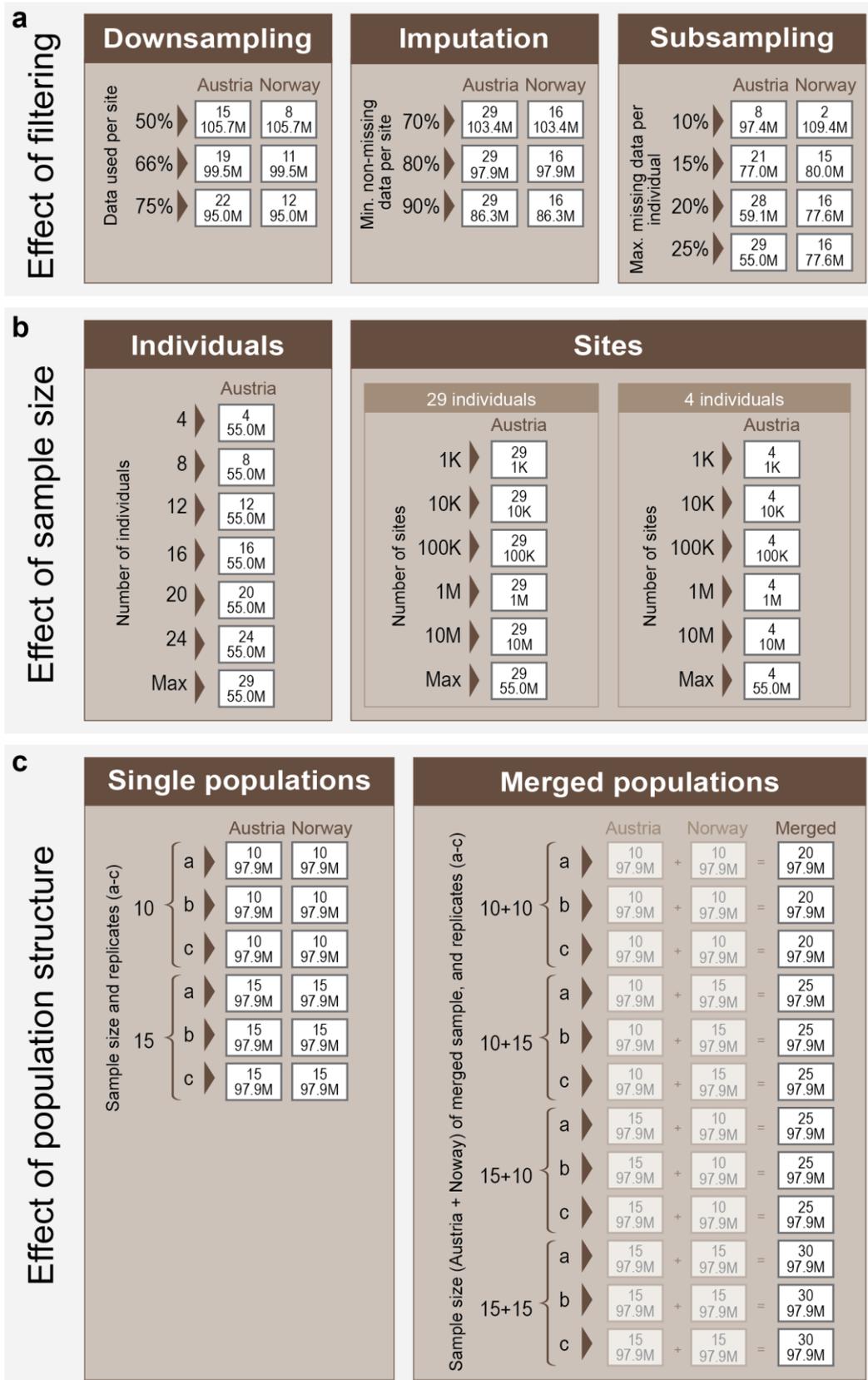
775 **SUPPORTING INFORMATION**

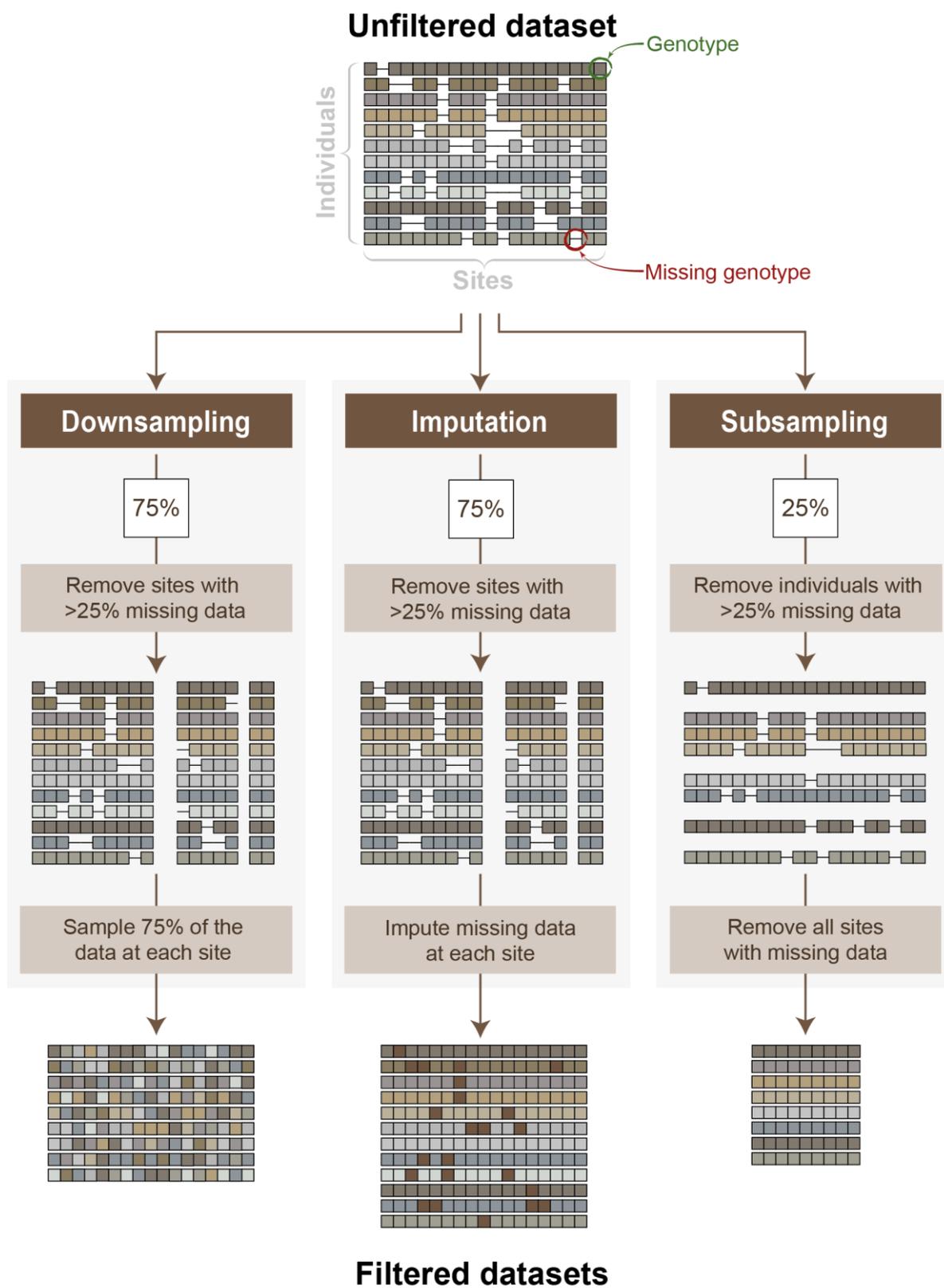
776 **Table S1.** Sequence information of the *Arabidopsis lyrata* samples included in this study.

777 **Table S2.** Accuracy of DFE estimations for different missing-data treatments and sample sizes
778 from the SLiM simulated dataset.

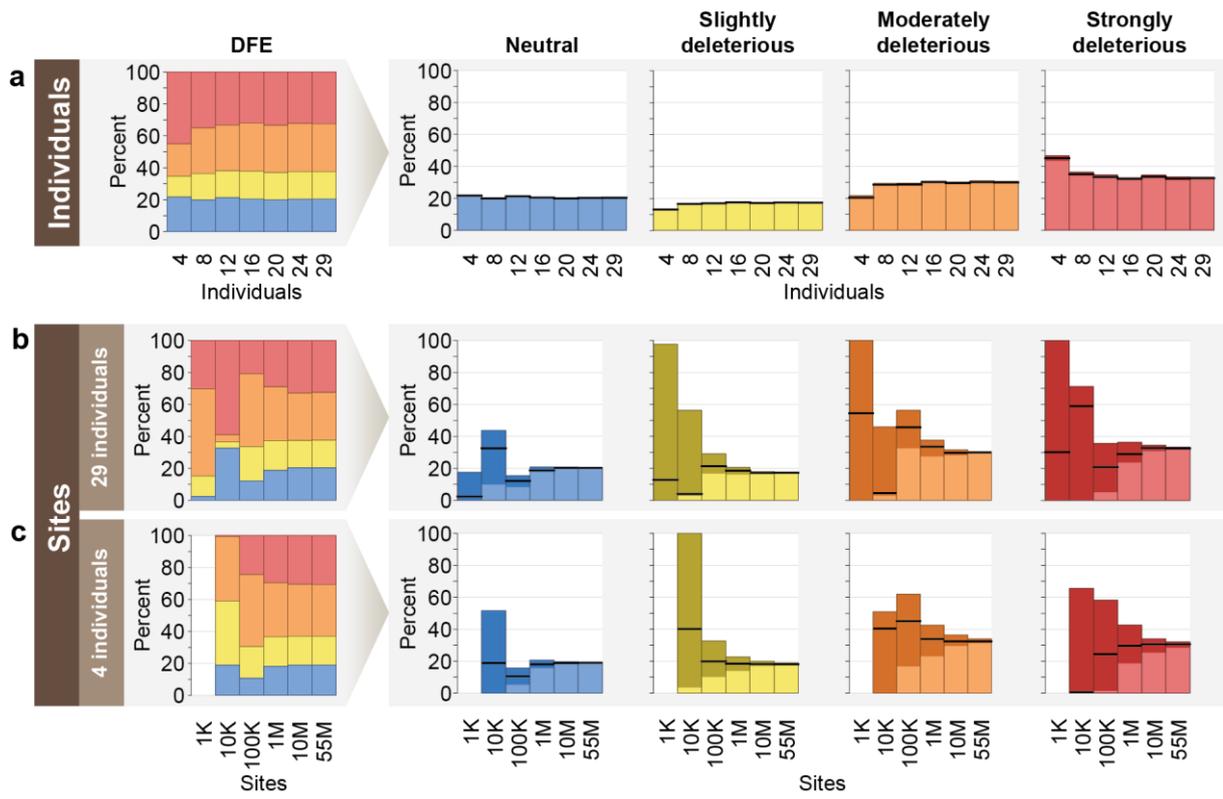
779 **Figure S1.** Principal component analysis in the 45 individuals of *Arabidopsis lyrata* sampled
780 from Austria and Norway, based on 3,921,575 SNPs.

781



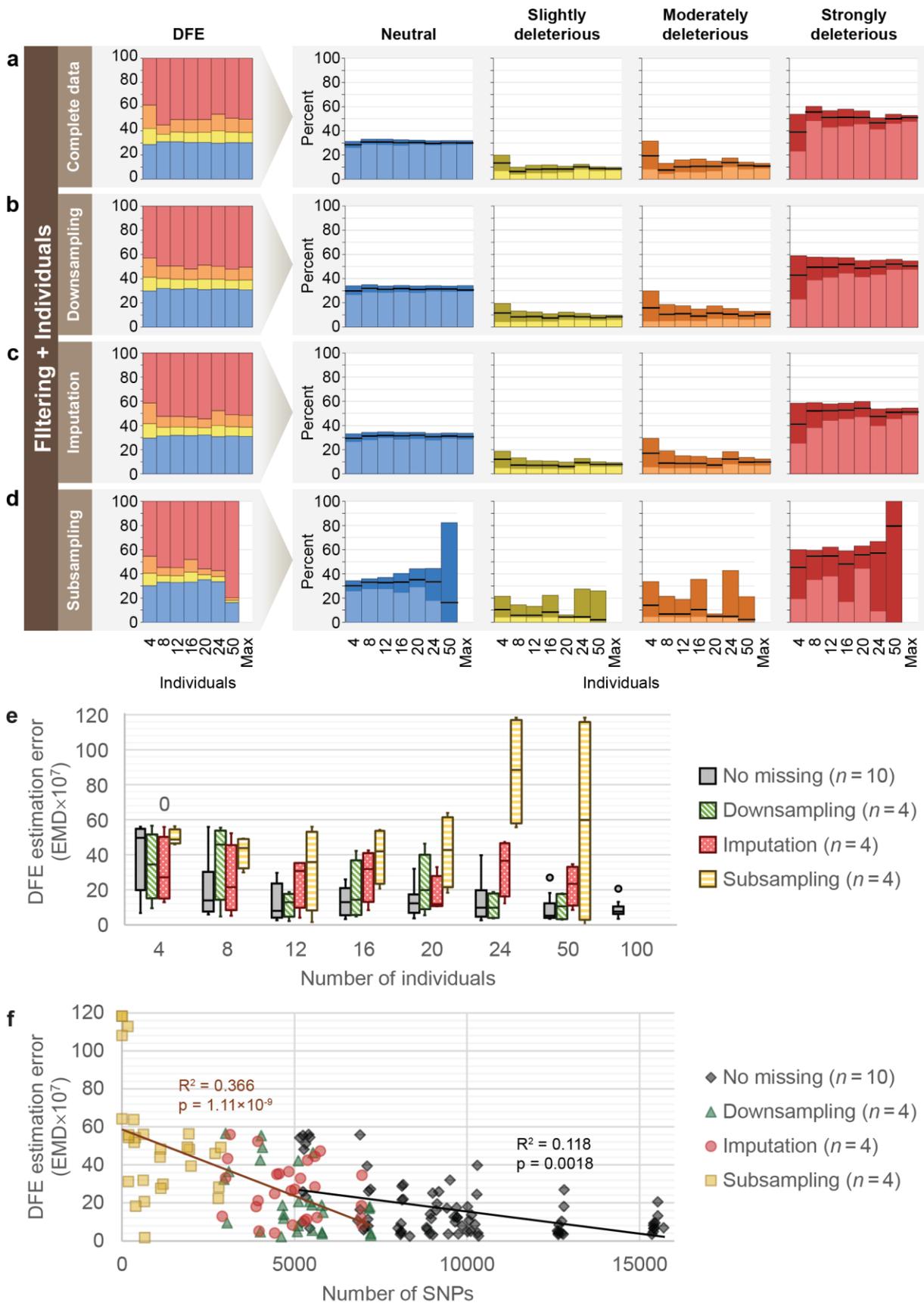


786 **Figure 3.**



787

788 **Figure 4.**



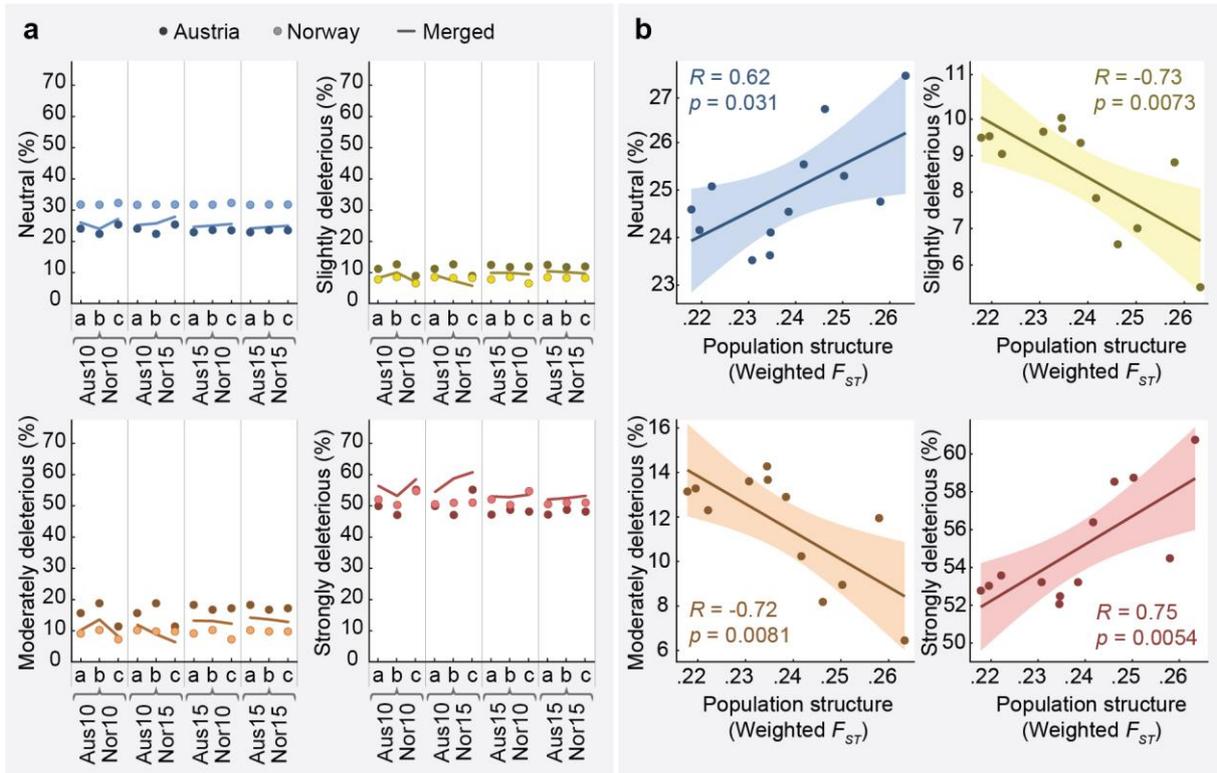
789

790

791

792

793 **Figure 5.**



794

795

Table 1. The estimated DFE using downsampling, imputation and subsampling procedures in Austrian and Norwegian populations of *A. lyrata*. For downsampling, thresholds show the percentage of data retained at each locus. Imputation thresholds signify the data quality (inverse of the max missing rate) of the individuals included in the dataset prior to imputation. Subsampling thresholds signify the max missing rates per individual.

	Filtering method	Individuals	Total sites	0-fold sites	4-fold sites	SNPs in SFS	DFE [95% CI]: % of mutations with $-N_e s$ values of:									
							[0, 1]	(1, 10]	(10, 100]	> 100						
Austria	Downsampling	50%	15	105,746,755	20,081,506	4,532,827	1,389,235	24.63	[24.53–24.73]	10.59	[10.47–10.72]	15.11	[14.87–15.34]	49.68	[49.37–49.96]	
		66%	19	99,518,927	19,796,632	4,465,916	1,455,146	23.26	[23.15–23.35]	11.73	[11.61–11.80]	17.53	[17.30–17.70]	47.48	[47.26–47.80]	
		75%	22	95,023,967	19,576,647	4,410,244	1,472,536	21.76	[21.69–22.95]	12.91	[11.45–13.00]	20.30	[17.07–20.48]	45.03	[44.75–48.57]	
	Imputation	70%	29	103,436,893	19,988,452	4,509,075	1,686,431	23.08	[22.97–24.04]	12.39	[11.26–12.51]	18.87	[16.47–19.12]	45.66	[45.35–48.25]	
		80%	29	97,909,623	19,724,874	4,444,971	1,625,688	22.67	[22.57–22.77]	12.12	[11.99–12.22]	18.44	[18.17–18.65]	46.77	[46.50–47.12]	
		90%	29	86,272,541	19,088,682	4,267,939	1,437,770	20.21	[20.09–20.33]	13.56	[13.46–13.68]	22.24	[21.99–22.52]	43.99	[43.69–44.27]	
	Subsampling	10%	8	97,383,774	19,593,039	4,364,309	843,938	22.96	[22.80–23.28]	11.21	[10.60–11.43]	16.61	[15.39–17.05]	49.22	[48.61–50.79]	
		15%	21	76,996,896	18,026,562	3,900,662	874,047	19.78	[19.60–19.90]	14.68	[14.50–14.87]	24.81	[24.40–25.28]	40.74	[40.16–41.24]	
		20%	28	59,099,749	14,532,777	3,066,767	662,641	20.11	[19.41–20.29]	16.87	[16.62–17.62]	29.21	[28.62–31.15]	33.81	[31.81–34.50]	
		25%	29	54,974,337	13,445,210	2,824,524	609,256	20.31	[20.10–20.56]	17.29	[16.96–17.56]	29.93	[29.16–30.56]	32.47	[31.72–33.38]	
	Norway	Downsampling	50%	8	105,746,755	20,081,506	4,532,827	374,403	33.13	[32.78–33.45]	9.40	[7.96–10.17]	12.06	[9.86–13.30]	45.41	[43.60–48.80]
			66%	11	99,518,927	19,796,632	4,465,916	366,254	32.05	[31.58–32.32]	7.91	[7.42–9.95]	9.86	[9.13–13.07]	50.17	[45.39–51.36]
75%			12	95,023,967	19,576,647	4,410,244	341,308	30.91	[30.70–31.15]	8.11	[7.72–8.67]	10.24	[9.64–11.10]	50.74	[49.38–51.59]	
Imputation		70%	16	103,436,893	19,988,452	4,509,075	399,078	32.83	[32.58–33.04]	6.51	[6.20–6.96]	7.80	[7.37–8.44]	52.86	[51.89–53.62]	
		80%	16	97,909,623	19,724,874	4,444,971	365,494	31.62	[31.33–31.89]	6.64	[6.26–7.20]	8.04	[7.49–8.85]	53.71	[52.54–54.49]	
		90%	16	86,272,541	19,088,682	4,267,939	268,958	29.30	[29.04–29.68]	8.06	[7.09–8.41]	10.27	[8.79–10.83]	52.38	[51.49–54.50]	
Subsampling		10%	2	109,442,991	20,192,673	4,555,968	248,706	7.26	[7.00–7.48]	86.62	[86.37–86.99]	6.12	[5.89–6.30]	0.00	[0.00–0.00]	
		15%	15	79,983,985	18,525,720	4,136,084	179,342	28.20	[27.89–28.57]	7.97	[7.46–8.23]	10.22	[9.42–10.64]	53.60	[53.01–54.72]	
		20%	16	77,586,760	18,343,099	4,091,607	171,711	28.36	[27.78–28.69]	7.55	[7.02–8.83]	9.56	[8.76–11.61]	54.53	[51.62–55.79]	
		25%	16	77,586,760	18,343,099	4,091,607	171,711	28.36	[27.78–28.69]	7.55	[7.02–8.83]	9.56	[8.76–11.61]	54.53	[51.62–55.79]	

Table S1. Sequence information of the *Arabidopsis lyrata* samples included in this study. Source codes are NCBI accession IDs.

Population	Source	Coverage (Mb, $\geq 5x$)	Mean depth ($\geq 5x$)	Median depth ($\geq 5x$)	No. reads
Norway	ERR3397904	147.26	25.71	16	63201441
Norway	ERR3397905	144.61	21.83	14	52859026
Norway	ERR3397906	143.18	22.77	13	54633798
Norway	ERR3397907	143.9	22.27	14	53839619
Norway	ERR3397908	146.44	23.74	15	59507152
Norway	ERR3397909	146.45	23.69	15	59465532
Norway	ERR3397910	145.45	24.72	14	62069317
Norway	ERR3397911	148	26.05	15	66516823
Norway	ERR3397912	145.1	23.77	14	59254019
Norway	ERR3397913	143.16	21.8	13	53951341
Norway	SRR5124977	151.64	70	38	127927590
Norway	SRR5124983	144.92	33.26	22	59142085
Norway	SRR5124985	135.63	22.97	13	34814708
Norway	SRR5124997	153.66	55.35	35	102114289
Norway	SRR5124998	149.93	49.79	30	88534336
Norway	SRR5124999	139.45	27.54	16	42830997
Austria	ERR3514864	130.18	17.02	11	20727733
Austria	ERR3514865	141.4	23.75	15	31042548
Austria	ERR3514866	140.31	21.29	15	27625035
Austria	ERR3514869	156.87	49.12	37	72515100
Austria	ERR3514870	145.82	25.47	18	34493579
Austria	ERR3514871	147.85	26.79	19	36994924
Austria	ERR3514872	130.12	17.95	11	22086024
Austria	ERR3514873	136.53	19.89	13	25584056
Austria	ERR3514874	141.42	26.26	16	34646984
Austria	ERR3514875	128.41	16.94	10	20457660
Austria	ERR3514876	148.58	29.17	22	39961518
Austria	ERR3514877	130.4	16.15	11	19708718
Austria	ERR3514878	155.98	42.5	33	62109925
Austria	ERR3514879	141.14	21.85	14	29584688
Austria	ERR3514880	153.3	38.58	27	54727972
Austria	ERR3514883	130	19.1	11	23003063
Austria	ERR3514884	130.2	17.75	11	21314626
Austria	ERR3514885	141.2	27.63	16	37910237
Austria	ERR3514886	124.05	15.02	10	17565928
Austria	ERR3514887	144.2	26.05	17	35825064
Austria	ERR3514888	142.67	26.2	17	34966847
Austria	ERR3514889	141.14	24.65	16	32554512
Austria	ERR3514892	138.26	19.26	14	24924196
Austria	ERR3514893	142.71	23.43	16	30871081
Austria	ERR3514895	148.06	30.95	20	42713949
Austria	ERR3514896	149.81	31.97	23	44511357
Austria	ERR3514897	135.69	22.12	13	27521061
Austria	ERR3514898	145.12	32.64	21	44216326
Austria	ERR3514899	144.58	29.89	19	39745319

Table S2: Number of sites and SNPs in different simulated datasets, together with their respective estimated mean (E_s) and shape (β) parameters of the DFE. Earth Mover’s Distance (EMD) signifies the accuracy of each estimated gamma distribution of DFE to the known DFE used in the simulation (shape β : 0.1, mean E_s : -100). Lower EMD values indicate a closer fit to the known DFE.

Filtering method	Individuals	Sites in VCF after filtering	Sites in SFS (min – max)	SNPs in SFS (min – max)	β		E_s		EMD (10^{-7})		
					(Median	[min – max])	(Median	[min – max])	(Median	Mean	[min – max])
No missing data (in 10 replicates)	4	50,000,000	12,162,648 – 12,162,869	5,156 – 5,448	0.0902	[0.0500 – 0.1675]	-682	[-6.36 $\times 10^6$ – -4]	49.7	38.5	[6.7 – 56.5]
	8	50,000,000		6,840 – 7,205	0.1064	[0.0500 – 0.1318]	-127	[-5.19 $\times 10^6$ – -21]	13.9	20.8	[5.8 – 56.5]
	12	50,000,000		8,002 – 8,342	0.1077	[0.0865 – 0.1326]	-103	[-1,044 – -18]	7.9	13.3	[2.5 – 30.5]
	16	50,000,000		8,832 – 9,182	0.1043	[0.0900 – 0.1308]	-148	[-714 – -22]	12.9	13.3	[3.2 – 26.5]
	20	50,000,000		9,420 – 9,859	0.1044	[0.0843 – 0.1220]	-134	[-1,293 – -37]	12.3	13.0	[3.7 – 32.5]
	24	50,000,000		10,019 – 10,336	0.1051	[0.0792 – 0.1291]	-126	[-2,773 – -24]	9.8	13.2	[2.6 – 40.5]
	50	50,000,000		12,617 – 12,818	0.1096	[0.0874 – 0.1235]	-94	[-821 – -36]	5.3	8.9	[3.4 – 27.5]
	100	50,000,000		15,357 – 15,702	0.1138	[0.0972 – 0.1280]	-79	[-276 – -33]	7.5	8.9	[3.5 – 21.5]
Downsampling (in 4 replicates)	4	28,420,524	6,911,406	2,959 – 3,099	0.1253	[0.0500 – 0.1411]	-39	[-1.23 $\times 10^7$ – -11]	34.5	33.7	[9.4 – 57.5]
	8	28,420,524		4,010 – 4,075	0.0724	[0.0500 – 0.1001]	-8,392	[-3.15 $\times 10^6$ – -147]	45.9	38.0	[4.8 – 55.5]
	12	28,420,524		4,621 – 4,694	0.0999	[0.0919 – 0.1162]	-198	[-420 – -52]	13.0	11.7	[2.2 – 19.5]
	16	28,420,524		5,085 – 5,122	0.0945	[0.0777 – 0.1082]	-336	[-3,711 – -80]	14.4	19.0	[4.9 – 42.5]
	20	28,420,524		5,446 – 5,539	0.0907	[0.0727 – 0.1096]	-456	[-6,829 – -78]	19.9	22.9	[5.5 – 46.5]
	24	28,420,524		5,795 – 5,827	0.0981	[0.0913 – 0.1030]	-229	[-417 – -136]	9.8	10.5	[3.7 – 19.5]
	50	28,420,524		7,165 – 7,216	0.0992	[0.0929 – 0.1088]	-248	[-391 – -88]	10.6	10.5	[3.0 – 18.5]
	85 (n=1)	28,420,524		8,323	0.1045		-139		4.0	4.0	
	Imputation (in 4 replicates)	4		28,420,524	6,911,406	2,907 – 3,136	0.1140	[0.0500 – 0.1498]	-137	[-5.64 $\times 10^6$ – -7]	38.3
8		28,420,524	3,892 – 3,976	0.0998		[0.0642 – 0.1171]	-293	[-54,893 – -32]	19.6	24.2	[5.2 – 52.5]
12		28,420,524	4,426 – 4,554	0.0848		[0.0815 – 0.1079]	-1,201	[-1,762 – -84]	30.0	24.9	[4.1 – 35.5]
16		28,420,524	4,840 – 5,017	0.0866		[0.0794 – 0.1111]	-811	[-1,953 – -63]	26.8	24.6	[8.3 – 36.5]
20		28,420,524	5,179 – 5,384	0.0894		[0.0760 – 0.0981]	-819	[-3,850 – -226]	22.2	24.4	[10.5 – 43.5]
24		28,420,524	5,592 – 5,750	0.0793		[0.0729 – 0.1155]	-2,976	[-8,094 – -50]	36.5	33.2	[12.3 – 47.5]
50		28,420,524	6,931 – 6,958	0.0948		[0.0814 – 0.1132]	-334	[-1,642 – -65]	15.4	18.5	[8.5 – 35.5]
100 (n=1)		28,420,524	8,396	0.0986			-247		11.6	11.6	
Subsampling (in 4 replicates)		4	26,099,661	6,347,733		2,707 – 2,870	0.1279	[0.0733 – 0.1518]	-23	[-11,804 – -6]	37.2
	8	13,629,024	3,313,984	1,897 – 2,003	0.0702	[0.0500 – 0.0737]	-11,337	[-8.96 $\times 10^6$ – -2,657]	48.7	48.3	[39.4 – 56.5]
	12	7,109,386	1,730,591	1,107 – 1,194	0.1040	[0.0696 – 0.1366]	-2,386	[-10,280 – -18]	37.0	37.5	[27.7 – 48.5]
	16	3,713,156	902,526	624 – 661	0.1109	[0.0782 – 0.1618]	-59	[-1,223 – -3]	26.3	27.6	[1.7 – 56.5]
	20	1,938,476	470,900	348 – 402	0.0907	[0.0500 – 0.1938]	-35,852	[-795,504 – -2]	53.1	47.1	[18.3 – 64.5]
	24	1,011,970	246,323	174 – 199	0.0540	[0.0500 – 0.5804]	-362,712	[-3.23 $\times 10^6$ – 0]	55.1	63.6	[31.2 – 113.5]
	50	14,863	3,737	3 – 5	0.5172	[0.0500 – 0.5227]	0	[-2.16 $\times 10^{12}$ – 0]	113.2	102.3	[64.2 – 118.5]

Figure S1. Principal component analysis in the 45 individuals of *Arabidopsis lyrata* sampled from Austria and Norway, based on 3,921,575 SNPs.

