

1 Regarding the *F*-word: the effects of data *Filtering* on inferred
2 genotype-environment associations

3 Running title: Filtering impacts on GEAs

4

5 Collin W Ahrens¹, Rebecca Jordan², Jason Bragg³, Peter A Harrison⁴, Tara Hopley⁵, Helen

6 Bothwell⁶, Kevin Murray⁶, Dorothy A Steane^{2,4}, John W Whale¹, Margaret Byrne⁵, Rose

7 Andrew⁷, Paul D. Rymer¹

8

9 ¹ Hawkesbury Institute for the Environment, Western Sydney University, Richmond NSW

10 Australia

11 ² CSIRO Land & Water, 15 College Rd, Sandy Bay 7005, Tasmania, Australia

12 ³ Research Centre for Ecosystem Resilience, Australian Institute of Botanical Science, The

13 Royal Botanic Garden Sydney, NSW 2000, Australia

14 ⁴ School of Natural Sciences and Australian Research Council Training Centre for Forest Value,

15 University of Tasmania, Hobart, Tasmania, Australia

16 ⁵ Biodiversity and Conservation Science, Department of Biodiversity, Conservation and

17 Attractions, Locked Bag 104, Bentley Delivery Centre, WA 6983, Australia

18 ⁶ Australian National University, Acton, ACT, 2601, Australia

19 ⁷ School of Environmental and Rural Science, University of New England, Armidale, Australia

20

21 Author Correspondence:

22 Collin Ahrens

23 Email: c.ahrens@westernsydney.edu.au

24 Phone: +61 2 4570 1862

25

26 Abstract

27 Genotype-environment association (GEA) methods have become part of the standard
28 landscape genomics toolkit, yet, we know little about how to filter genotype-by-sequencing data
29 to provide robust inferences for environmental adaptation. In many cases, default filtering
30 thresholds for minor allele frequency and missing data are applied regardless of sample size,
31 having unknown impacts on the results. These effects could be amplified in downstream
32 predictions, including management strategies. Here, we investigate the effects of filtering on
33 GEA results and the potential implications for adaptation to environment. Using empirical and
34 simulated datasets derived from two widespread tree species to assess the effects of filtering on
35 GEA outputs. Critically, we find that the level of filtering of missing data and minor allele
36 frequency affect the identification of true positives. Even slight adjustments to these thresholds
37 can change the rate of true positive detection. Using conservative thresholds for missing data
38 and minor allele frequency substantially reduces the size of the dataset, lessening the power to
39 detect adaptive variants (i.e. simulated true positives) with strong and weak strength of
40 selections. Regardless, strength of selection was a good predictor for GEA detection, but even
41 SNPs under strong selection went undetected. We further show that filtering can significantly
42 impact the predictions of adaptive capacity of species in downstream analyses. We make
43 several recommendations regarding filtering for GEA methods. Ultimately, there is no filtering
44 panacea, but some choices are better than others, depending largely on the study system,
45 availability of genomic resources, and desired objectives of the study.

46

47 Keywords: *Eucalyptus*; climate adaptation; genome sequencing; genomic simulation; GEA;
48 reduced representation; SNP analysis

49

50

51 Introduction

52 Identifying genomic patterns associated with adaptation in wild populations can provide
53 information to support management strategies as well as facilitate fundamental discoveries
54 (Garner et al., 2016; Sgrò, Lowe, & Hoffmann, 2011). We can improve our understanding of the
55 response of species to changing climates and their evolutionary potential by leveraging
56 knowledge about adaptive genetic variation in natural populations (Browne, Wright, Fitz-Gibbon,
57 Gugger, & Sork, 2019; Razgour et al., 2019; Sork, 2017). Genotype–environment association
58 (GEA) methods are used to identify potentially adaptive loci in non-model systems based on
59 correlations between allele frequencies and environmental data. In recent years, there has been
60 a proliferation of genomic studies on landscape adaptation using GEA analyses (Ahrens et al.,
61 2018), which is becoming a standard part of the analytical pipelines for landscape genomics.

62 The utility of GEA analyses is limited by several problems, including the presence of false
63 positives (type I errors) (Storz, 2005). While, false negatives (type II errors) are likely common
64 due to controlling for population structure (Sork et al., 2013), they are unlikely to limit or
65 confound the GEA results. False positives are present in GEA outputs regardless of filtering,
66 significance thresholds or false discovery corrections (Forester et al., 2018). From a biological
67 perspective, false positives are genomic variants significantly associated with the environment
68 through random, neutral processes. For example, demographic processes can generate clines
69 in allele frequencies that covary with environmental gradients, leading to neutral SNPs
70 potentially being falsely identified as adaptive (François, Martins, Caye, & Schoville, 2016;
71 Hoban et al., 2016; Lotterhos & Whitlock, 2015). However, these impacts will vary depending on
72 the unique demographic history (e.g. bottlenecks, population growth, or rapid expansion) of a
73 species. Many GEA methods control for patterns of population structure, to reduce false positive
74 call rates, but by doing so, true positives are also at risk of becoming false negatives (Nadeau,

75 Meirmans, Aitken, Ritland, & Isabel, 2016; Orsini, Mergeay, Vanoverbeke, & Meester, 2013).
76 One way to control for false positive call rates is to combine the results of multiple approaches
77 in the hope of identifying loci with well-supported associations with environmental variables
78 (Meirmans, 2015). However, the outcomes of these approaches are variable (Nadeau et al.,
79 2016) and, this is not surprising given the numerous statistical models and methods used to
80 mitigate the confounding effects of genetic structure. Also, the consequences of false positives
81 could vary, depending on the conservation or management applications associated with the
82 analysis. For example, the presence and overrepresentation of false positives could have
83 implications for conservation actions, through the identification of patterns of putative adaptation
84 that are supported more by false positives than true positives (i.e. the noise is stronger than the
85 signal).

86 The occurrence of false positives is partially attributable to incomplete genome sampling (Lowry
87 et al., 2017). The proportion of the genome sampled can be influenced at many stages of the
88 workflow, including choice of genotyping method, library preparation method (e.g. enzyme
89 choice), bioinformatic processing, and data quality filtering (O'Leary, Puritz, Willis, Hollenbeck, &
90 Portnoy, 2018). Most GEA studies of non-model organisms employ reduced representation
91 approaches, as they are cost-effective, do not require extensive genomic resources (e.g.
92 reference genomes) (Manel et al., 2016) and often yield thousands of loci scattered across a
93 species' genome. Yet, even small genomes are poorly sampled through reduced representation
94 library preparation. For example, a dataset of 20 k SNPs only represents ~0.7% of a 550 Mbp
95 genome with a linkage disequilibrium decay of 200 bp (2.75 million linkage blocks). Thus, for
96 many reduced representation approaches, the likelihood of detecting positive associations is
97 limited by querying a very small proportion of the genome. Previous studies have amply
98 reviewed how choices made during library preparation and bioinformatic processing impact the
99 level of genome sampling that can be achieved for any given reduced representation dataset
100 (Mastretta-Yanes et al., 2015; O'Leary et al., 2018). In addition, total sample size is also known
101 to have an impact on the power of GEA analyses and identification of false positives (Lotterhos

102 & Whitlock, 2015). While the importance of sample size alone has been discussed previously as
103 an important factor for sample design for GEA analyses (Lotterhos & Whitlock, 2015; de Mita et
104 al., 2013), it is unknown how sample size interacts with filtering choices. Therefore, in this study
105 we explore the explicit impact of data quality filtering on downstream GEA results.

106 Filtering remains incredibly challenging, and a highly important aspect of population genomics
107 data analysis (Andrews & Luikart, 2014). Optimal, default filtering settings suitable for all GEA
108 studies are unlikely, given the range of organisms and research questions explored. Even so,
109 documenting the effects of data filtering on analyses has proved highly useful for other
110 population genetic applications, assisting researchers to set filters that are appropriate for their
111 experimental design and individual study goals (Narum, Buerkle, Davey, Miller, & Hohenlohe,
112 2013). For example, it has been shown previously that SNP calling and filtering settings can
113 affect estimates of heterozygosity and F_{ST} (Díaz-Arce & Rodríguez-Ezpeleta, 2019; Shafer et
114 al., 2017), routinely used in conservation decision making (Gautier et al., 2012; Pool, Hellmann,
115 Jensen, & Nielsen, 2010). Minor allele frequency (MAF) filtering settings can change F_{ST}
116 estimates (Hendricks et al., 2018; Linck & Battey, 2019), due to the inclusion of locally isolated
117 alleles increasing the perceived dissimilarity of populations. Liberal thresholds of missing data
118 have been shown to reduce estimates of expected heterozygosity and increased inference of
119 inbreeding; however, the results vary across species (Fu, 2014). Stringent filtering increases
120 completeness of the dataset at the expense of the number of SNPs retained and the proportion
121 of the genome sampled. While it is general practice to filter missing data to low levels, no
122 studies to date, as far as we are aware, have investigated the impact of missing data on
123 downstream GEA results. In addition, filtering of reduced representation datasets from
124 organisms without genomic resources is even more critical, because *de novo* alignment can
125 introduce errors (O'Leary et al., 2018). While the importance of filtering has been
126 acknowledged, the impacts of filtering thresholds on GEA analyses have yet to be fully
127 investigated.

128 In many cases, GEA analyses and outputs are cited as being useful for downstream
129 applications, including the improvement of management, conservation, and breeding programs.
130 While commendable, we do not know how filtering choices might impact final recommendations.
131 As the dataset changes due to filtering, so too will the identified set of putatively adaptive SNPs,
132 and these differences could be compounded when extrapolating across environmental space.
133 Often these extrapolated maps, of adaptive genomic variation across species' ranges, are the
134 currency of interpretation for stakeholders and decision-makers. The connections between
135 geospatial predictions of adaptation and genomic variation to support management /
136 conservation outcomes is evident in studies on birds (Bay et al., 2018) and grasses (Ahrens et
137 al., 2020), where researchers quantify the heterogeneity of genomic vulnerability to climate
138 change. However, these predictive outputs could be affected by filtering choices.

139 Filtering requires subjective decisions about how best to compile the best available dataset to
140 investigate genomic adaptation across landscapes, while limiting the proportions of false
141 positives and false negatives identified by GEA analyses. No definitive filtering guidelines for
142 GEA currently exist. Instead researchers are left to iteratively change filtering thresholds and
143 subjectively choose a perceived optimal dataset for the question at hand (as demonstrated by
144 the range of filtering settings identified in a GEA meta-analysis; Ahrens et al., 2018). This
145 subjective process may result in ambiguous interpretation and the potential for bias in the
146 reporting of results. As the incorporation of GEAs into analytical pipelines increases, it is
147 important to establish objective guidelines to assist researchers in determining the impact that
148 filtering can be expected to have on downstream GEA results. Therefore, we ask two questions:
149 1) how does filtering affect the identification of putatively adaptive loci? and, 2) how does our
150 ability or inability to identify associations affect downstream applications? To answer these
151 questions, we test four common assumptions:

152 (1) More stringent filtering reduces identification of false positives.

153 (2) Loci with strong selection strengths will be identified as significant, regardless of filtering
154 choices.

155 (3) Combining GEA analyses reduces false positive call rates.

156 (4) Extrapolation of adaptive variants across the landscape reveals consistent areas of
157 climate adaptation.

158 We test these assumptions using both empirical and simulated data sets, the latter matched to
159 the empirical demographic scenarios with the addition of known true positives. We explore how
160 early filtering decisions affect conservation and management decisions and provide guidelines
161 for data filtering to optimise the effectiveness of GEA methods.

162 Methods

163 SNP and climate data

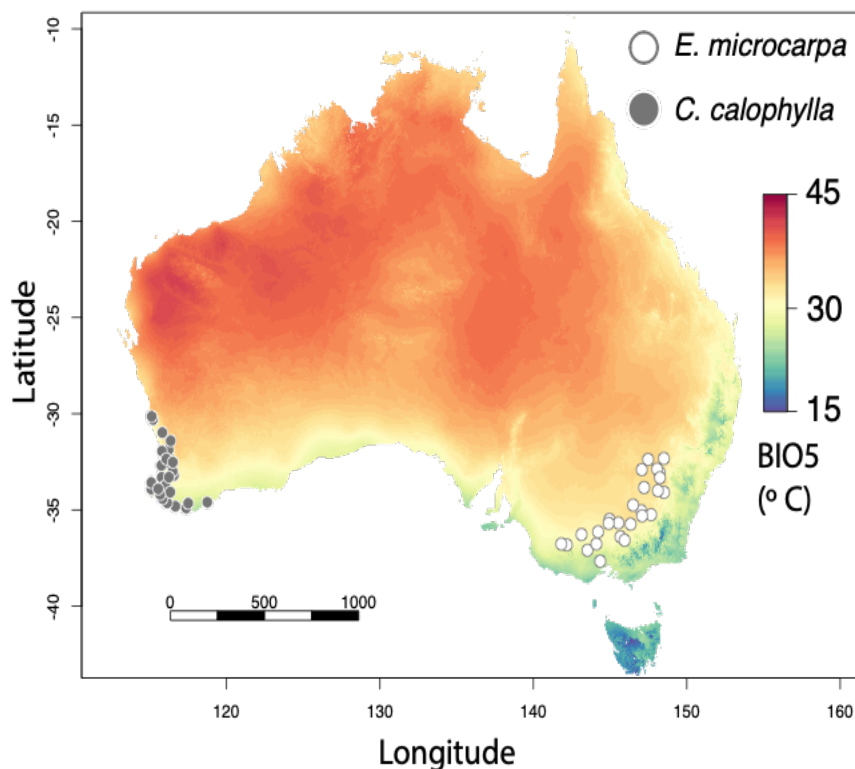
164 We chose two reduced representation SNP datasets from different genera within the eucalypt
165 group: *Eucalyptus microcarpa* (Maiden) Maiden (Jordan, Hoffmann, Dillon, & Prober, 2017) and
166 *Corymbia calophylla* (Lindl.) K.D.Hill & L.A.S.Johnson (Ahrens, Byrne, & Rymer, 2019). Both
167 species are native to south-eastern and south-western Australia respectively (Figure 1). By
168 comparing phylogenetically close species, we minimised potential confounding effects arising
169 from using species with very different genomes, thereby allowing us to focus on how filtering
170 affects GEA results.

171

172 The datasets were based on sampling across the range of each species. The *E. microcarpa*
173 dataset consisted of a total of 577 samples from 26 populations and the *C. calophylla* dataset
174 comprised 263 samples from 27 populations. Genomic data for both species were generated
175 using DArTseq (Diversity Arrays Technology P/L, Canberra, Australia), with the same library
176 preparation, multiplexing, and sequencing protocols. The raw, unfiltered genotype data were
177 used as the input datasets, with different filtering applied as described below. Genotypes were
178 quality filtered prior to analysis, retaining those with an individual minimum read-depth of 10x,
179 minimum genotype quality Phred-score of 30 and a maximum mean read-depth of 100x,
180 retaining only biallelic SNPs.

181

182 Climate data were extracted from WorldClim (Fick & Hijmans, 2017) for each sampling location
183 using the R package *raster* (R core team 2019). We chose the mean maximum temperature of
184 the warmest month (BIO5) to test the effect of filtering on genotype-environment association
185 (GEA) analyses. Temperature was selected as it is commonly used in GEA analyses and a key
186 selective force given projected increases into the future; BIO5 represents the high temperature
187 extremes, presumably a greater selective pressure than mean annual temperatures in Australia
188 (Prober et al., 2016; Costa e Silva, Potts, Harrison, & Bailey, 2019). To assess the potential
189 effect of multiple variables confounding GEA results, we also tested mean precipitation of the
190 driest month (BIO14), representing a second key selective force of precipitation. Assessments
191 of spatial autocorrelation (Moran's I) and effective population size, given the environment ($n_{\text{eff-}}$
192 $_{\text{env}}$) was performed, provide critical metrics for determining which climate variables have greater
193 power to detect SNPs under selection (details in Supplementary information).



194

195 **Figure 1.** Map of the sampled locations for the two study species with maximum temperature of
196 the warmest month (BIO5) shown across Australia.

197
198
199
200
201
202

Table 1. Attributes of empirical and simulated datasets. Pearson’s correlation coefficient (r); spatial autocorrelation (Moran’s I); effective sample size due to environment ($n_{\text{eff-env}}$); BIO5 - maximum temperature of the warmest month; BIO14 - precipitation of the driest month; number of SNPs remaining after filtering for largest and smallest analysis datasets (#SNP).

	Empirical				Simulation			Structure	r	Moran’s I	$n_{\text{eff-env}}$
species	# samples	# pops	#SNP largest	#SNP smallest	# samples	#SNP largest	#SNP smallest	F_{ST}	BIO5 ~ BIO14	BIO5	BIO5
<i>E. microcarpa</i>	577	26	25,826	2,931	650	20,685	3,494	0.01	0.014	0.267	15.1
<i>C. calophylla</i>	263	27	25,811	5,595	270	21,255	5,031	0.05	-0.75	0.327	13.6

203
204

205 Simulated data set creation

206 Simulated SNP datasets were generated to be comparable to the empirical datasets, with two
207 main motivations. First, the effect of missing data can be studied by generating complete
208 simulated SNP datasets, and then implementing different levels of ‘missingness’. Second,
209 simulated datasets enable evaluation of the performance of GEA (rates of detection of false
210 positives and negatives) in relation to different filtering treatments. This can be accomplished by
211 including known true positives (TP) with different magnitudes of selection pressure.

212

213 Simulated datasets were generated using the *simulate.baypass* R function in BayPass (Gautier,
214 2015). This function creates simulated datasets under a BayPass model (see Coop, Witonsky,
215 Rienzo, & Pritchard, 2010; Günther & Coop, 2013) using an empirical matrix of allelic
216 covariances (the Ω matrix). It generates SNPs whose allele frequencies vary across populations
217 according to the covariance matrix previously estimated from the empirical datasets, with an
218 additional associations of prescribed strength to a bioclimatic variable. Two simulated datasets
219 were generated based on the species’ empirical data, hereafter referred to as ‘*Sim microcarpa*’
220 and ‘*Sim calophylla*’ to distinguish from empirical datasets of *E. microcarpa* and *C. calophylla*,
221 respectively.

222

223 We simulated population-level allele counts for ~25 000 ‘neutral’ SNPs plus 200 ‘adaptive’ (i.e.
224 simulated SNPs that are correlated with a specific climate variable) SNPs whose coefficients of
225 association with each of the two bioclimatic variables were drawn from a uniform distribution
226 between -0.3 and 0.3 (beta.coef). We chose these selection coefficients knowing that, at their
227 extremes, they are likely greater than the values we would find in wild populations. We did this
228 intentionally to verify that loci with very strong selection coefficients were highly likely to be
229 identified in the GEA analyses. Other *simulate.baypass* parameters were chosen so that the
230 simulated data resembled our empirical datasets. For example, the simulation function uses a
231 beta distribution to describe the frequencies of ancestral alleles among loci. We chose the
232 parameters for this distribution by fitting the beta distribution to the minor allele frequencies
233 observed in the empirical datasets. *Corymbia calophylla* returned shape1 = 0.54 and shape2 =
234 0.53, whereas *E. microcarpa* returned shape1 = 0.43 and shape2 = 0.43. Fixed loci were
235 removed from the simulated datasets, resulting in a loss of 1000-1600 SNPs per dataset. We
236 also wanted to approximate, in the simulations, the way missing data were distributed across
237 samples and across loci in the empirical data sets. We therefore began by fitting statistical
238 distributions to frequencies of missing genotypes across loci and samples in the empirical data.
239 We used the estimated distributions to impose missing alleles on the loci and samples across
240 the simulated datasets (Figure S1). If we sampled from a distribution and obtained a negative
241 number of missing genotypes for a locus, we set the value of missingness for that locus to 0.

242

243

244 Subsetting datasets

245 To understand how filtering choices affect the ability of GEAs to identify true positives ,we
246 filtered each data set by minor allele frequency (MAF), missing data (MD), and the number of
247 samples per population (all 150 data sets represented in Table 2). We chose three MAF to
248 explore (0.01, 0.05, and 0.1; Table 2) based on the most commonly applied thresholds (Ahrens
249 et al., 2018). We applied five MD thresholds (10%, 20%, 30%, 40%, and 50%; Table 2). The

250 most commonly applied MD thresholds are between 10 and 30%; we included thresholds up to
 251 50% to test how less-stringent MD thresholds would behave with GEA methods.

252

253 The importance of biological sampling design on GEA analyses has been demonstrated
 254 previously (see Forester et al., 2018; Lotterhos & Whitlock, 2015; de Mita et al., 2013 for more
 255 thorough treatments of sampling design), and do not try to replicate these studies but rather. Six
 256 individuals per population is often regarded as the minimum sample size for population genetics
 257 analyses when thousands of SNPs are available (Nazareno, Bemmels, Dick, & Lohmann, 2017;
 258 Willing, Dreyer, & Oosterhout, 2012) and GEA studies (Lotterhos & Whitlock, 2015). We
 259 therefore tested the effect of using 6 or 10 individuals per population for both species, as well as
 260 25 individuals per population for *E. microcarpa*, reflecting the empirical *C. calophylla* and *E.*
 261 *microcarpus* datasets, respectively.

262

263 **Table 2.** Matrix detailing the 150 data filtering combinations explored in the present study.
 264 Numbers within the table represent the total number of individuals per dataset: 6, 10, or 25
 265 individuals per population. The number of populations remained constant throughout the study
 266 (*C. calophylla* - 27 populations; *E. microcarpa* - 26 populations). MAF - minor allele frequency.

MAF Dataset	Proportion of Missing Data				
	50%	40%	30%	20%	10%
<i>C. calophylla</i>	6, 10	6, 10	6, 10	6, 10	6, 10
<i>Sim calophylla</i>	6, 10	6, 10	6, 10	6, 10	6, 10
0.01 <i>E. microcarpa</i>	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25
<i>Sim microcarpa</i>	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25
<i>C. calophylla</i>	6, 10	6, 10	6, 10	6, 10	6, 10
<i>Sim calophylla</i>	6, 10	6, 10	6, 10	6, 10	6, 10
0.05 <i>E. microcarpa</i>	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25
<i>Sim microcarpa</i>	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25
<i>C. calophylla</i>	6, 10	6, 10	6, 10	6, 10	6, 10
<i>Sim calophylla</i>	6, 10	6, 10	6, 10	6, 10	6, 10
0.1 <i>E. microcarpa</i>	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25
<i>Sim microcarpa</i>	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25	6, 10, 25

267

268

269 GEA Analyses

270 We focused on three commonly used GEA methods with different underlying computational
271 models to identify SNP-climate associations. We compared two univariate methods, LFMM2
272 (Caye, Jumentier, Lepeule, & François, 2019) and BayPass (Gautier, 2015) which associates
273 each SNP individually with a given climate variable, and one multivariate method, redundancy
274 analysis, RDA following the usage in Forester et al., (2018).

275

276 LFMM2 uses discrete ancestral clusters computed via principal component analysis (PCA) to
277 control for population structure, and a least-squares approach for confounder estimation of
278 genomic data (Caye et al., 2019). As LFMM2 requires a full data set, we imputed data using the
279 mean method with the *impute* function as the default and may be considered the 'worst case
280 scenario' imputation method (note: the mean method is a naive imputation method, and we
281 suggest using other imputation methods). We ran PCAs for each data set to assess the change
282 in population structure as a result of filtering choices (data not shown). As expected, population
283 structure varied across datasets, likely due to the low, but present population structure ($F_{ST} =$
284 0.05 & 0.01; Table 1). We observed only very slight changes from a $K = 3$ to a $K = 4, 5,$ or $6,$
285 with $K = 3$ being the most consistent solution for both species. Therefore, we used $K = 3$ for all
286 LFMM2 analyses to allow direct comparisons across data sets. Significant associations were
287 called at $\alpha = 0.001$ after applying a false discovery rate as suggested by Caye et al., (2019). We
288 explored lower significance thresholds but found they were too permissive, returning high
289 numbers of false positives; 0.001 seemed to be similar to the BayPass significance factor, a
290 Bayes Factor (BF), of 20.

291

292 BayPass uses an Ω matrix to account for population structure based on allelic covariance
293 between populations. BayPass analyses were run following the methods described in the
294 BayPass manual. We ran the standard model twice to obtain the Ω matrix, and averaged the Ω

295 matrix across runs. The mean Ω matrix was used as the covariance matrix within the auxiliary
296 model, which calculates a BF to assist with identification of SNP-climate associations. The
297 auxiliary model was run twice, and results averaged across runs. The parameters used for both
298 models (standard and auxiliary) were 20 pilot runs for 1000 iterations, 2500 burn-in, and 1000
299 MCMC samples. Significant associations were called at a BF > 20, considered ‘decisive’
300 evidence (Jeffreys, 1961). As above, for LFMM2, we explored other significance levels with
301 results returning high numbers of false positives.

302

303 Complementing the univariate GEA analyses, we also performed a redundancy analysis (RDA).
304 This multivariate method has been shown to be robust across a wide range of selection
305 strengths, demographic histories, sampling designs, and in the presence of many levels of
306 population structure (Forester et al., 2018). To address the RDA requirement of a complete data
307 set, we calculated and used population-level allele frequencies, instead of imputation. For RDA,
308 an $\alpha = 0.05$ was used to extract significant SNPs along the two climate axes, across the three
309 main RDA axes. Variance inflation factors (VIF) were used to check multicollinearity between
310 the two climate variables, *C. calophylla* returned 2.35 VIF for both climatic variables and *E.*
311 *microcarpa* returned 1.00 VIF for both, indicating that these are sufficiently independent to
312 identify associations via RDA because they are below 10 (Zuur et al., 2010).

313

314 For each dataset and analysis, we recorded the SNPs that were identified as having significant
315 associations with environment. For simulated datasets, we recorded which SNPs were true
316 positives (TP) and which were false positives (FP). We also recorded ‘pseudo positives’ (PP),
317 defined as SNPs that were found to be significantly associated with one climate variable but
318 were in fact TP for the other climate variable i.e. were identified as significantly associated with
319 BIO5 but were actually adapted to BIO14.

320

321 In order to test whether there is a strength of selection threshold for which GEA methods
322 achieve a 100% TP call rate, we plotted strength of selection (beta coefficient applied during

323 simulations) against the significance of association for BayPass (BF) and LFMM2 (calibrated P -
324 value) for *Sim calophylla*. We also calculated the difference between the significance values for
325 each MAF threshold and the standard deviation. This estimate allowed us to quantify the mean
326 differences and variance between data sets differentiated only by MAF.

327

328 Impacts of filtering on extrapolation and interpretation of adaptive variation

329 To determine how filtering thresholds may affect the downstream extrapolation of putatively
330 adaptive genomic variation across geographic space, we estimated the genomic-informed
331 ‘climate selection surface’ for both species. Here, a climate selection surface refers to the
332 prediction of adaptation through geographic space. This extrapolation followed the logic of
333 Steane et al. (2014), but using RDA instead of canonical analysis of principal coordinates
334 (details provided in supplementary information). The effect of each filtering parameter was
335 explored separately in the simulated datasets, holding other filtering parameters constant (e.g.,
336 when assessing the effect of MAF, the MD and sample size thresholds were held constant). We
337 also compared the impact of different filtering methods on the empirical datasets for the most
338 liberal (MD = 50%; MAF = 0.01) and conservative (MD = 10%; MAF = 0.1) datasets. Significant
339 differences between climate selection surfaces were determined using a pixel pairwise z-score
340 test. Here, the liberal dataset was compared to the conservative dataset, such that a positive
341 difference between the two resulting climate selection surfaces corresponds to the liberal
342 dataset predicting more adaptive variation, and a negative difference corresponds to the
343 conservative dataset predicting more adaptive variation.

344 Results

345 Effects of filtering on GEA outputs - simulated data

346 Using simulated data that reflected natural population structure and climate gradients across *C.*
347 *calophylla* and *E. microcarpa* (*‘Sim calophylla’* and *‘Sim microcarpa’* respectively), we found
348 that data filtering influenced the identification of ‘adaptive’ SNPs. Filtering regimes differentially

349 impacted the data sets and GEA programs in various ways. Both filtering thresholds (missing
350 data (MD), minor allele frequency (MAF)) and biological sample size influenced the number of
351 significant SNP-climate associations. Furthermore, filtering thresholds also impacted the
352 number of true positives (TP), false positives (FP) and pseudo-positives (PP).

353

354 With the exception of RDA for *Sim calophylla*, the GEA methods identified SNP associations
355 with BIO5, including TPs (Figures 2 & 3). The multivariate RDA approach performed
356 exceedingly poorly for *Sim calophylla* and only moderately well for *Sim microcarpa* compared to
357 the other two GEA methods in all aspects, particularly in identifying TPs. For *Sim calophylla*, this
358 finding was surprising and might be due to the fact that the climate variable is closely associated
359 with the population structure (see Ahrens et al., 2019 for details), identifying all TPs as false
360 negatives; alternatively, the demographic history *C. calophylla* may make RDA less sensitive to
361 true associations, as no associations were found in the empirical dataset either. Because of this
362 complication, we focus the results on BayPass and LFMM2.

363

364 The numbers of TPs and FPs increased with higher proportions of missing data (Figure 2). This
365 pattern reflects, in part, the total number of SNPs retained in each filtered dataset, with fewer
366 SNP-climate associations and TPs retained when more stringent filtering was applied (Figure
367 S2). There were significant relationships between the number of TPs found and the total
368 number of SNPs kept in the analysis for both species (*Sim microcarpa* - $r^2 = 0.93$, $p = <0.0001$;
369 *Sim calophylla* - $r^2 = 0.89$, $p = <0.0001$) (Figure S2). On the other hand, the amount of missing
370 data had little influence on the proportion of TPs in 'All Associations' (AA) and, thus, the ratio of
371 TPs to AAs remained constant within method and species (TP:AA; Figure 3). Although the
372 TP:AA ratio was markedly different between species and between methods within species
373 (Figure 3).

374

375 In general, a smaller MAF identified more TPs and more FPs than a large MAF (Figure 2). The
376 increase in FPs was especially apparent in the LFMM2 analysis for the *Sim calophylla* data,

377 where a MAF of 0.01 yielded nearly twice as many FPs as TPs (Figure 2b). For the *Sim*
378 *microcarpa* dataset, a MAF of 0.1 identified substantially fewer TPs than lower MAFs, although
379 the decrease in FPs was not as clear because of the already low FP call rate. The proportion of
380 TPs in AAs varied with MAF (Figure 3). For the *Sim calophylla* data, a larger MAF generally
381 resulted in a higher proportion of TPs (higher ratio of TP:AA). For the *Sim microcarpa* data, a
382 MAF of 0.01 generally had the lowest proportion of TPs (lowest ratio of TP:AA), with the highest
383 proportion of TPs varying between MAF 0.05 and 0.1 depending on the program used and
384 amount of missing data.

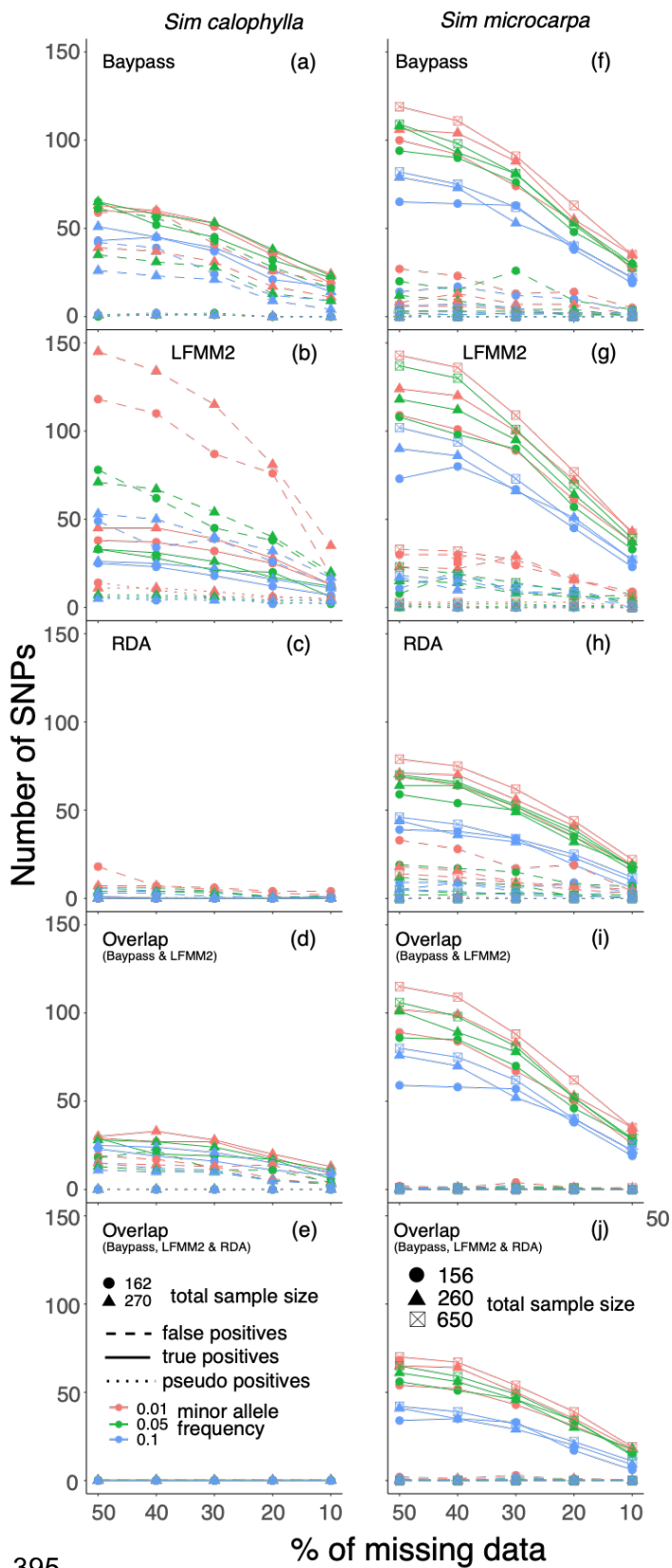
385

386 Sample size and pseudo positives differed between species and method. Larger biological
387 sample sizes consistently identified more TPs for *Sim microcarpa*, whereas sample size had
388 less influence on TP identification (Figure 2; more detailed results about sample size are in the
389 supplementary information). Pseudo positives (PP) were at or near zero for *Sim microcarpa* for
390 both BayPass and LFMM2, but PPs were detected for *Sim calophylla* in LFMM2, but few in
391 BayPass (Figures 2a).

392

393

394



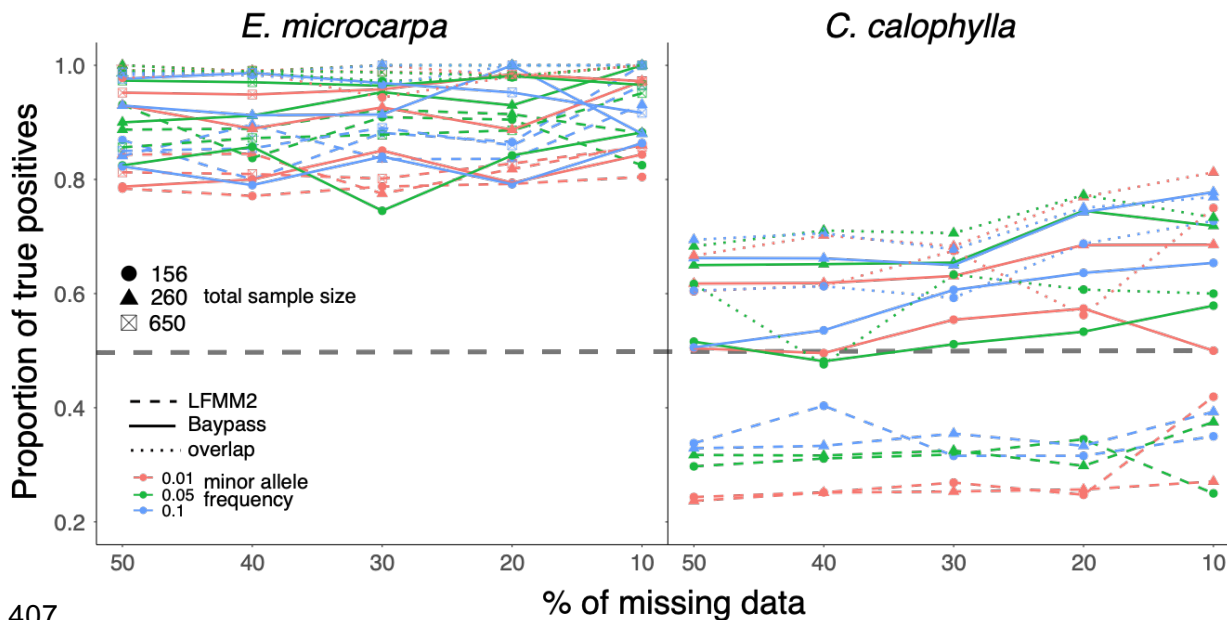
395

396

397 **Figure 2.** The number of significant SNP-climate associations with BIO5 (maximum temperature
 398 of the warmest month), for the simulated datasets (a-e) *Sim calophylla* and (f-j) *Sim microcarpa*

399 using three GEA analytical approaches: (a,e) BayPass, (b,g) LFMM2, and (c,h) RDA; including
 400 'overlap' – common identified associations – between (d,i) BayPass and LFMM2, and (e,j)
 401 BayPass, LFMM2 and RDA. Associations called false positives (FP) – significant 'non-adaptive'
 402 SNPs; true positives (TP) – significant 'adaptive' SNPs; and pseudo positives (PP) – SNPs
 403 'adaptive' for BIO14 (precipitation of the warmest month) but found to be significantly associated
 404 with BIO5.
 405

406



407

408 **Figure 3.** The proportion of *True Positives* (TP) among all identified associations (AA) called in
 409 BayPass, LFMM2, and the SNPs shared between them. The dashed horizontal line indicates
 410 50% TPs in AA; equal to a 1:1 ratio of TPs vs false positives (FP). For values above this line
 411 TPs > FPs, while below the line TPs < FPs.
 412

413 Overlapping results

414 A common approach for determining putatively 'adaptive' SNPs is to select those SNPs
 415 identified in multiple, independent analyses (Lotterhos et al., 2017), the rationale being that
 416 these SNPs are more likely to be TPs. Our results show a slight increase in the proportion of
 417 TPs identified (increased TP:AA) when results from independent analyses were combined
 418 (Figure 3). This was due to a small reduction in the number of FPs compared to the most
 419 conservative method (i.e. BayPass). However, this reduction in FPs came at the cost of fewer
 420 TPs being retained. In general, the number of TPs retained was reduced to the level of the more
 421 conservative dataset. For *Sim microcarpa*, the number of TPs was reduced to BayPass
 422 numbers for the BayPass-LFMM2 overlap (Figure 2i) and reduced to RDA numbers for the

423 BayPass-LFMM2-RDA overlap in *Sim microcarpa* (Figure 2j). *Sim calophylla* had a substantially
424 greater decrease in TPs when comparing the overlap between BayPass and LFMM2, dropping
425 to less than either Baypass or LFMM2 (Figure 2d). There were no identified TPs common to all
426 three analyses for *Sim calophylla*, reflecting the lack of TPs from RDA (Figure 2e). Using
427 multiple methods decreased the number of FPs, to the point of there being very few or zero FPs
428 for *Sim microcarpa* (Figure 2). This decrease in FPs compared to TPs when using multiple
429 methods slightly increased the proportion of TPs in the set of SNPs common to multiple GEA
430 methods (Figure 3).

431

432 The influence of selection strength on identifying associations

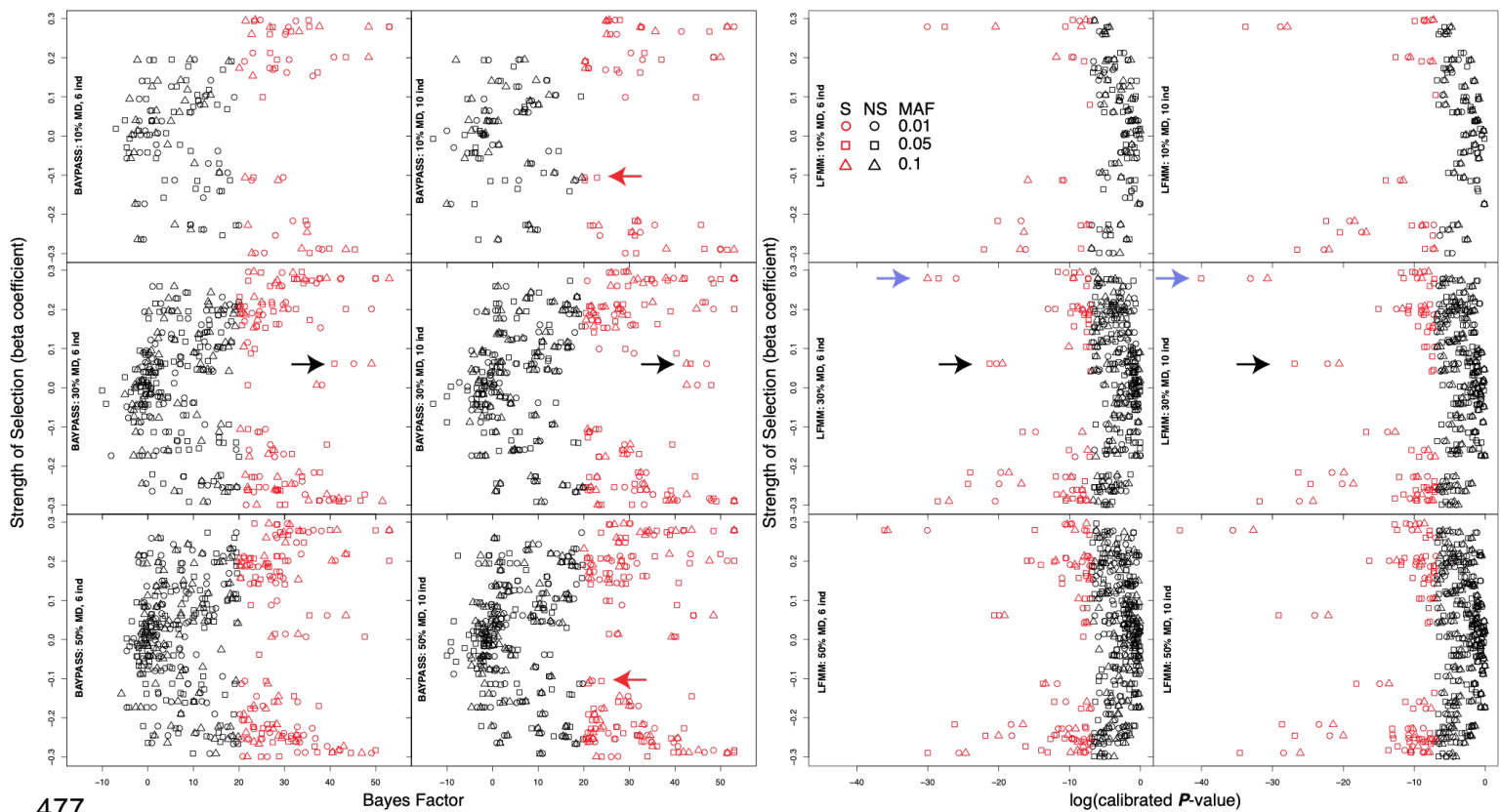
433 We hypothesised that the strength of selection prescribed in the simulations would influence the
434 magnitude of the association inferred, and ultimately, the likelihood of detecting TPs. In
435 particular, we wanted to know if it was possible to identify a threshold above which the call rate
436 for TP was 100%. The strength of selection for individual TPs did impact the identification of
437 significant associations. While never 100% accurate at any strength of selection, linear models
438 revealed significant relationships between the strength of selection and levels of significance for
439 BayPass and LFMM2 ($r^2 = 0.47$ and 0.22 respectively), showing the strength of selection does
440 have some effect on results (Figure S3). However, a threshold for high TP call rates was only
441 observed at low levels of missing data (10%) for BayPass (strength of selection +/- 0.28; Figure
442 4). This threshold disappeared when we included more SNPs through filtering and no threshold
443 was identified for LFMM2.

444

445 Increasing the strength of selection increased the rate of TP detection. However, false negatives
446 (SNPs under selection not detected as significant) occurred across all selection strengths
447 (Figure 4). The proportion of missing data appeared to have more of an effect on identifying TPs
448 than the number of samples, but this is likely due to differences in the total number of SNPs in
449 the dataset and not due to missing data *per se* (Figure S2). There was little change in the
450 number of TPs identified whether 6 or 10 individuals per population were sampled. Furthermore,

451 as more data were retained through less stringent filtering of missing data, we could identify TPs
452 under weaker selection for both BayPass and LFMM2. However, even with adjustments to
453 sample size and the amount of missing data (number of SNPs retained), a large proportion of
454 TPs were not identified irrespective of filtering parameters. For example, in *Sim calophylla* only
455 20% (\pm 3% SD) of the simulated adaptive SNPs were identified by LFMM2, 30% (\pm 3% SD) by
456 BayPass, and 0% (\pm 0% SD) by RDA. In analyses of *Sim microcarpa* a higher proportion of the
457 adaptive variants was identified, with 75% (\pm 5% SD) of the SNPs under selection being
458 identified by LFMM2, 62% (\pm 3% SD) by BayPass, and 38% (\pm 2% SD) by RDA.

459
460 Minor allele frequency, in combination with biological sample size, impacted the significance of
461 individual SNPs. There were multiple examples where a SNP was considered significant for one
462 MAF but not another (Figure 4 highlights three SNPs indicted by red, black, and blue arrows).
463 One SNP (red arrow, Figure 4) was identified as significant when MAF = 0.01 and 0.05 but not
464 when MAF = 0.1, while holding the number of individuals to 10 and MD at 10%. However, these
465 three SNPs were significant at all MAFs when allowing 50% missing data. Furthermore, the
466 significance of the same SNP with different MAFs can change depending on the method or
467 sample size. For example, one SNP (black arrows, Figure 4) in the Baypass analysis using six
468 individuals per population (162 total), was most significant when MAF = 0.1. When there were
469 10 individuals per population (270 total) the significance of this SNP was greatest when MAF =
470 0.01, and lowest when MAF = 0.1. We investigated whether these differences might be due to
471 variation in the covariance (Ω) matrices but found that the covariation among covariance
472 matrices were highly correlated (correlation coefficients ranged between 0.87 and 0.93; all p -
473 values $<$ 0.001) and had little effect on the observed differences. One SNP detected in the
474 LFMM2 analyses (blue arrows, Figure 4) showed a significance pattern with MAF 0.1 $>$ 0.05 $>$
475 0.01 when there were six individuals per population, but the significance rank changed to MAF
476 0.05 $>$ 0.01 $>$ 0.1 when there were 10 individuals per population.



477

478 **Figure 4.** The strength of selection for each SNP and the resulting power of association for
 479 BayPass (Bayes Factor) and LFMM2 (calibrated P -value) for *Sim calophylla*. S = significant
 480 (red); NS = not significant (black). See text for explanations of red, blue and black arrows.
 481

482 While significance levels were significantly ($p < 0.001$) consistent across datasets, LFMM2 had
 483 higher consistency with all values >0.98 correlation values while BayPass were between 0.8
 484 and 0.87 for both species (Table S2), slight changes of filtering thresholds did affect outcomes
 485 in some circumstances. The influence of MAF on individual SNP significance was observed
 486 when comparing significance levels of individual SNPs identified for *Sim calophylla* (Table S3).
 487 For BayPass, MAF had a greater effect on the significance level of individual SNPs when using
 488 smaller sample sizes (162 vs 270 individuals); more SNPs became non-significant when the
 489 biological sample size was smaller. Although the difference in significance level varied with
 490 biological sample size, the variation (SD) was similar (Table S3). The opposite was observed
 491 with LFMM2 where MAF had less impact (i.e. smaller differences and less variation) on the
 492 significance levels of individual SNPs in analyses that used smaller biological sample sizes

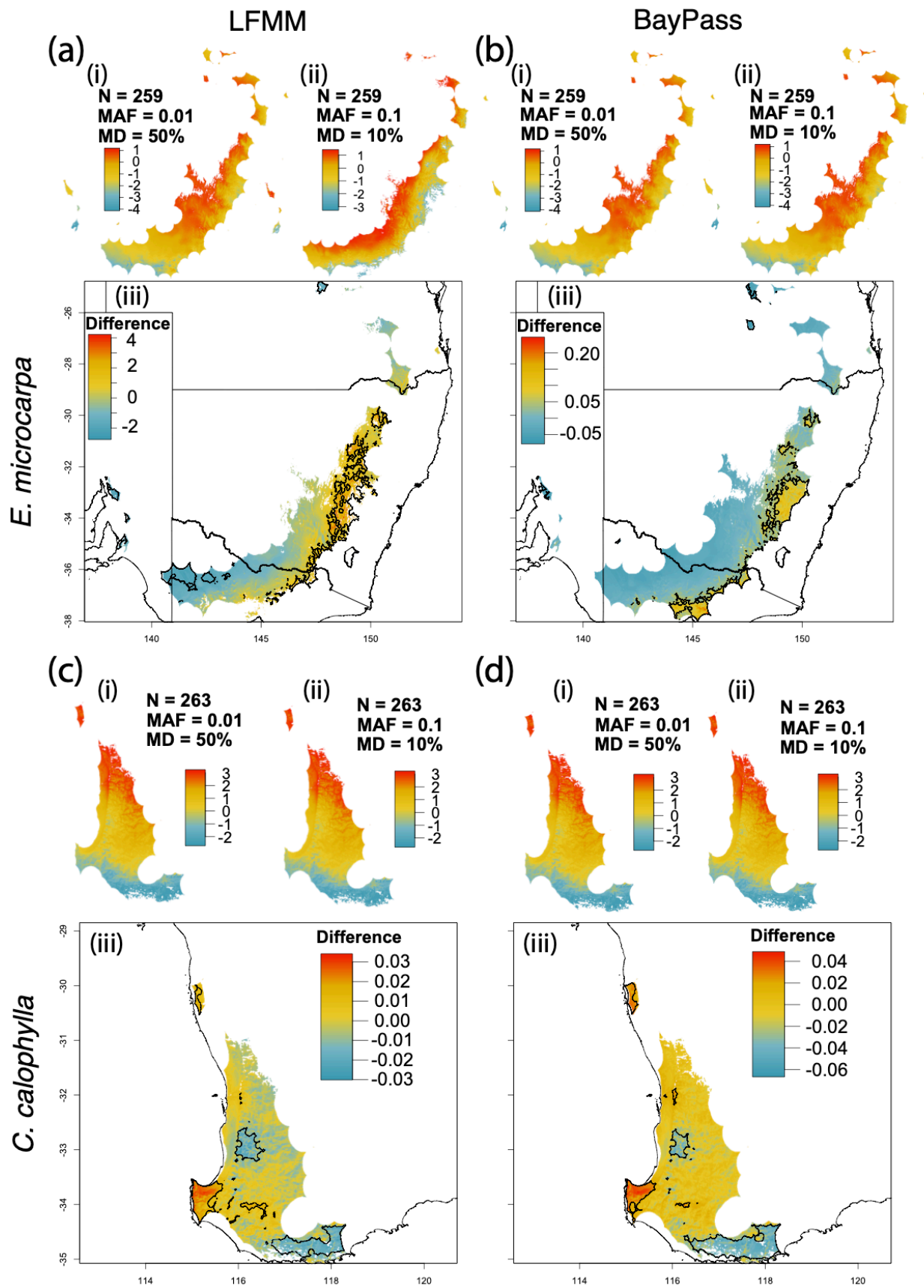
493 (Table S3), yet, compared to BayPass, more SNPs became non-significant when changing from
494 MAF = 0.01 to MAF = 0.05.

495

496 Impacts of filtering on extrapolation and interpretation

497 The impact of filtering the SNP datasets on downstream extrapolation of ‘adaptive’ genomic
498 variation across geographic space varied depending on the thresholds applied and the GEA
499 method. The greatest difference in the climate selection surface (i.e. adaptive predictions
500 through geographic space) between the two approaches and two species was observed for the
501 empirical dataset for *E. microcarpa* using LFMM2 (Figure 5). Applying the conservative
502 thresholds for MAF and MD (while keeping sample size constant) resulted in a significantly
503 different pattern of adaptive genomic variation across the landscape (climate selection surfaces)
504 with different geographic areas predicted to be locally adapted (e.g. red surfaces in Figure 5).
505 This is evident in the comparison between the surfaces produced using the conservative and
506 liberal thresholds using LFMM2 on *E. microcarpa*, where more liberal SNP filtering tended to
507 have a north-south pattern compared to an east-west pattern for the conservative filtering.
508 These contrasting spatial patterns resulted in large differences between predictions (an adaptive
509 index change of > 4). The liberally filtered dataset for LFMM2 was more consistent with both of
510 the predictions for BayPass. Conversely, the effect of filtering was not as apparent when using
511 BayPass on *E. microcarpa*, nor on any of the *C. calophylla* predictions, where, though
512 significant, only subtle differences between filtering thresholds and GEA methods were
513 observed (Figure 5). Nevertheless, the incongruences for *E. microcarpa* occurred along the
514 margins of the species distribution with the conservative filtering method slightly underpredicting
515 putative adaption compared to the liberally filtered dataset. Likewise, incongruences for *C.*
516 *calophylla* adaptive predictions showed statistically significant differences along the species
517 margins but also within the interior region for both GEA methods, although LFMM2 had slightly
518 larger incongruences compared to BayPass. Consistent with the empirical datasets, the general
519 spatial patterns of adaptive variation predicted using the simulated datasets remained
520 qualitatively the same despite filtering for MAF, MD, and sample size, indicating that the signal

521 to noise remained quite similar despite the higher number of FPs in the more liberal datasets
522 (Figure S4). However, we detected regions where there were statistically significant differences
523 among datasets, particularly along the margins of the species' ranges. Those differences were
524 driven by different filtering parameters and GEA methods. For instance, the biggest changes for
525 *Sim microcarpa* in BayPass are driven by MAF, but missing data and number of samples had
526 the biggest impact in LFMM2 (Figure S4). The increase in the number of individuals from 260
527 total individuals to 650 individuals had very little impact on landscape-wide patterns of genomic
528 variation (i.e. adaptive index).



529

530

531 **Figure 5.** Analysis of the effect of filtering on spatial extrapolation of adaptive variation within
532 the empirical datasets. The maps within boxes (iii) show the differences between the 'liberal' (i)
533 and 'conservative' (ii) maps (smaller maps directly above). Combinations of species are *E.*
534 *microcarpa* and LFMM2 (a), *E. microcarpa* and Baypass (b), *C. calophylla* and LFMM2 (c), and
535 *C. calophylla* and BayPass (d). Red surface colours in the smaller maps represent regions of
536 each species gene pool putatively adapted to hotter and drier climates while the blue surface
537 represents the regions putatively adapted to increasingly cooler and wetter climates. The red
538 surface in the main differential maps represent regions where the liberal dataset predicted
539 stronger adaptation, whereas the blue surface corresponds to regions where the conservative
540 dataset predicted stronger adaptation. Note: the differential scales are different across
541 comparisons, this was done to highlight the differences within each comparison. Areas of
542 significant differences in predicted magnitude of adaptation are outlined with a black polygon.
543 Liberal dataset = missing data (MD) = 50%; minor allele frequency (MAF) = 0.01. Conservative
544 dataset = MD = 10%; MAF = 0.1; MAF = minor allele frequency; MD = missing data; N = sample
545 size.

546 Discussion

547 Most studies filter data prior to GEA analysis with the aim of improving the quality of the input
548 data to obtain better inferences of environmental adaptation. While several studies have
549 explored the influence of demographic history, population structure, sampling strategy,
550 landscape configuration, and strength of selection on the capacity of various approaches to
551 detect loci under selection (Forester et al., 2018; Lotterhos & Whitlock, 2014, 2015; Luu, Bazin,
552 & Blum, 2016; de Mita et al., 2013; Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015;
553 Schlamp et al., 2015; de Villemereuil et al., 2014), the impact of filtering thresholds on GEA
554 outputs has not been thoroughly evaluated previously. Given the wide range of filtering choices
555 used and the lack of broad scale patterns of adaptation (Ahrens et al., 2018), we explored the
556 impact of filtering on the capacity of various approaches to identify putatively adaptive SNPs,
557 and demonstrated that filtering thresholds do impact the outcomes of GEA analyses. We reveal
558 that filtering for minor allele frequency and missing data affects GEA outputs in various ways
559 depending upon species, sample size, and GEA analytical method. To summarise how our
560 study challenges the four common assumptions addressed in the introduction:

561 (1) More stringent filtering reduces the identification of FPs but the rate of identifying FPs
562 remains constant across most filtering thresholds.

563 (2) Loci with strong selection strengths are more likely to be identified as TPs but a strong
564 selection strength does not guarantee a significant identification.

- 565 (3) Combining GEA analyses slightly reduces FPs but at the expense of TPs.
- 566 (4) Predictions across the landscape, for the most part, were biologically robust but
- 567 statistically different across all filtering thresholds, although some circumstances led to
- 568 biologically and statistically different adaptive patterns.

569 Ultimately, we found that filtering choices can have multiplicative effects for downstream

570 interpretation, meaning that small filtering changes could change estimates of genomic

571 predicted adaptation to the environment. While we focused on widespread tree species, the

572 concepts drawn from our results are applicable across other organisms and we suggest that

573 some common practices employed in GEA studies should be reconsidered.

574

575 Effects of filtering on GEA outputs

576 Missing data are usually minimised in order to improve the reliability of the dataset. However,

577 our results suggest that filtering data with strict missing data thresholds does not necessarily

578 improve GEA outcomes. In fact, filtering missing data seemed to have little effect on the ratio of

579 TP to AA. This is in line with other population genetic studies that found that missing data (within

580 reason) do not affect calculations of F_{ST} or H_e (Binks, Gibson, Ottewell, Macdonald, & Byrne,

581 2019; Díaz-Arce & Rodríguez-Ezpeleta, 2019; Shafer et al., 2017). Indeed, we found that

582 BayPass, LFMM2, and RDA (specific to *Sim microcarpa*) were robust to missing data with

583 respect to TP:AA but the actual number of TPs and FPs identified varied. For BayPass and

584 RDA, this could be partially due to the use of population-level allele counts or allele frequencies

585 as the input data, a strategy that effectively ignores missing data. Because LFMM2 uses

586 individual genotypes, and we naively imputed the gaps using loci means (default parameter), we

587 expected that missing data would result in more FPs and thereby provide a possible source of

588 differentiation among methods (de Villemereuil et al., 2014). While this was apparent for *Sim*

589 *calophylla*, LFMM2 performed well for *Sim microcarpa*. The 'missingness' was similar among

590 species, meaning that the different responses between species suggest that the relatively high

591 FP call rate for *Sim calophylla* is likely due to a combination of missingness and other

592 underlying differences between the species.

593

594 The lack of improvement to GEA outputs with decreasing proportions of missing data suggests
595 that the number of SNPs in a dataset is more important than dataset completeness, within
596 reason, bearing in mind that we only tested up to 50% missing data. More SNPs allow
597 sufficiently large numbers to statistically define 'neutral demographic structure', an important
598 aspect to all GEA analyses, and thus increase the number of putatively adaptive SNPs identified
599 (see further discussion below). The relative importance placed on filtering missing data should
600 depend on the downstream application of putatively adaptive loci. This is borne out by the maps
601 in Figure S4 (particularly between the missing data thresholds), where the presence of more
602 FPs do not affect the adaptive signal to non-adaptive noise, at least when the signal from TPs is
603 sufficiently large. However, this interpretation must be qualified, because the discovery of TPs in
604 empirical datasets is unknown, and it is the strength and number of TPs that will override a
605 contrasting FP signal.

606

607 Minor allele frequency is an important threshold, because nonsynonymous SNPs are likely to
608 have a MAF less than 0.05 (Cargill et al., 1999) and, in human studies, inclusion of SNPs with
609 low MAF increases the rate of identification of causal variants (Gorlov, Gorlova, Sunyaev, Spitz,
610 & Amos, 2008). Our data suggest that a low minor allele frequency has a type I error (FP) rate
611 close to nominal levels (i.e. FP rate is similar among datasets), which has been found in other
612 studies (Moskvina, Craddock, Holmans, Owen, & O'Donovan, 2006; Tabangin, Woo, & Martin,
613 2009). These findings suggest that low MAF should not be excluded from GEA datasets if
614 sampling design is sufficiently large. However, in our study, MAF influenced FP call rates with
615 varying impacts between programs and species. It is important to note that MAF filtering is also
616 a function of sample size and missing data. The larger the sample size, the smaller the MAF
617 threshold can be. This is most apparent when considering MAF as minor allele counts (MAC;
618 see O'Leary et al., 2018 for discussion), where a low MAF could still result in a high MAC for
619 larger sample sizes, allowing for sequencing error issues to be resolved by maintaining SNPs

620 that are called confidently (higher MAC). Ultimately, like missing data, MAF affects the total
621 number of SNPs in the dataset, but it can also influence a SNP's significance.

622

623 While more stringent filtering may, theoretically, improve the quality of the dataset, the reduction
624 in the overall size of the data set and the potential loss of informative loci may influence the null
625 models underlying GEA analyses and thus the identification of SNP-environment associations.

626 This is evident in the impact of both missing data and MAF on the detection of adaptive SNPs
627 under different strengths of selection. Both Baypass and LFMM2 missed TPs at all selection

628 strengths, even for SNPs under strong selection pressure. However, datasets that included

629 more missing data yielded TPs that were under weak selection (~ 0.05). This is likely because

630 less stringent filtering of missing data results in larger datasets, thereby increasing the overall

631 number of TPs. In addition, despite the missing data, larger datasets (relative to reduced

632 representation datasets with 2-20k SNPs) may enable a more statistically significant 'null model'

633 for the GEA and therefore greater power to detect loci under selection (Morin, Martien, & Taylor,

634 2009); and the power of the number of SNPs in genome-wide association studies has been

635 discussed previously (Hong & Park, 2012; Klein, 2007; Spencer, Su, Donnelly, & Marchini,

636 2009), and the same logic applies for GEAs. Filtering of MAF may also influence the null model,

637 changing the significance of TPs and, thus, their potential to be identified as TPs. While

638 stringently filtering genomic data may create a more reliable dataset in theory, having fewer

639 data points appears to reduce the overall power and effectiveness of GEAs.

640

641 Combining results

642 Using loci identified across multiple analyses reduced both the number of FPs and TPs. This is
643 common practice and in one sense, our results support this commonly-used approach (Forester

644 et al., 2018; Lotterhos & Whitlock, 2015) in that we observed a slight increase in TP:AA.

645 However, the TPs retained reflected the more conservative analysis, and most of the TPs

646 identified by the other methods were lost. Each method uses unique approaches to identify

647 SNPs (e.g. controlling for population structure and statistical model) and different methods are

648 likely to identify different suites of putatively adaptive SNPs. This output agrees with findings
649 from Forester et al., (2018); that combining results will bias the results to strong selective
650 sweeps and limit findings to the least powerful (or most conservative) method. The trade-off
651 between reduced FPs and the loss of informative TPs therefore needs careful consideration,
652 particularly given that downstream extrapolation of results tends to be largely unaffected by the
653 presence of FPs. If one uses an overlapping approach we suggest using the Lotterhos et al.
654 (2017) composite measure to improve the identification of adaptive signals by using the outputs
655 across many GEA methods.

656

657 Influence on downstream applications

658 For the most part, the patterns of geospatial predictions were biologically similar but statistically
659 different within species and methods, but across filtering thresholds. However, this was not the
660 case for *E. microcarpa* and LFMM2. The difference between the liberal and conservative
661 datasets revealed different biological and statistical geospatial adaptive patterns. The more
662 liberal dataset was more similar to both BayPass outputs, suggesting that the LFMM2
663 conservative prediction was spurious. While it is possible that both patterns are correct due to
664 hierarchically complex relationships between adaptation and climate, this pattern is likely due to
665 the fact that the FPs had a larger impact on the adaptive signal because there were fewer TPs
666 overall (i.e. the noise was greater than the signal), as only five putatively adaptive SNPs
667 associated with BIO5 (eight for BIO14) were identified in the conservative dataset compared to
668 101 putatively adaptive SNPs associated with BIO5 (36 for BIO14) in the liberal dataset. This
669 outcome suggests that FPs can affect predictions when fewer TPs are found for LFMM2, but
670 this effect was lost when more TPs are kept through larger datasets and liberal filters.

671

672 Pseudo positives (PP) were found to be a confounding factor, particularly for the *Sim calophylla*
673 dataset. Indeed, the correlation coefficients of the two environmental variables suggested that
674 PPs would have a much greater impact on the *Sim calophylla* dataset than on the *Sim*
675 *microcarpa* dataset. While we did find PPs in *Sim calophylla*, they numbered only about 20% of

676 the number of TPs; this was less than expected considering the strong correlation of the two
677 climatic variables across the distribution of *C. calophylla*. This suggests that it is preferable to
678 include environmental variables that are not correlated in GEA analyses (see Hoban et al.,
679 2016); however, the inclusion of variables with correlation coefficients around 0.7 seems to be
680 adequate (agreeing with the findings in Dormann et al., (2013)), particularly if they were chosen
681 a priori with hypothesis-driven questions.

682

683 In our simulations, we chose higher than expected strength-of-selection coefficients to try to
684 identify selection coefficients that would enable identification of adaptive SNPs above a given
685 threshold. We were unable to identify a consistent threshold and therefore conclude that strong
686 selection pressure is not sufficient to identify adaptive SNPs, and that the SNPs must be
687 distributed throughout the populations in specific ways. However, we did find a strong
688 relationship between the significance of SNPs and strength-of-selection, indicating that, not
689 surprisingly, there is a much higher probability of identifying SNPs of large effect using either of
690 the univariate methods than with RDA.

691

692 Differences among species

693 The datasets that we examined showed different responses to the effects of filtering despite
694 being (i) derived from related species that span similar climate gradients, and (ii) produced
695 using the same reduced representation approach. One reason for these differences could be
696 the different genome sizes of these species. The genome of *C. calophylla* is estimated to be
697 400 Mb while that of *E. microcarpa* is around 700 Mb. Although genome size is likely not
698 evolutionarily significant (Vu et al., 2015), it could influence the search for adaptive SNPs, as a
699 smaller genome size would provide better representation of coding regions. A second reason for
700 the differences between datasets could be that, even though the two species inhabit similar
701 temperature gradients, the broader climate of each species is fundamentally different: *C.*
702 *calophylla* occurs in a Mediterranean-type climate and *E. microcarpa* occurs in a temperate
703 climate. A third reason for the differences between species could be that similar levels of global

704 population structure does not dictate how genetic variance is distributed within species. For
 705 instance, it is possible that when more SNPs are kept due to filtering thresholds, the estimated
 706 population structure may change in different ways for each species. Finally, species' geographic
 707 range size, as well as demographic and evolutionary history, may explain differences in results.
 708 *Eucalyptus microcarpa* has a larger geographic range than *C. calophylla*, indicating that
 709 underlying demographic history could be fundamentally different (e.g. expansion/contraction).

710

711 **Table 4.** Outcomes and suggestions of different filtering approaches for different project aims
 712 employing GEA analyses. TP = True Positive; FP = False Positive; MD = missing data; MAF =
 713 minor allele frequency; MAC = minor allele count.

Aim / Concern	Example research question or application	Filtering approach	Analysis outcome
Conservation	Understanding general landscape patterns of genomic diversity for conservation or management. Reference genome may not be available.	More relaxed filtering to create larger overall SNP dataset. Smaller permissible MAF (given sample size and thus MAC). Larger amount of MD. Pool unique candidate SNPs across multiple methods.	Large overall pool of 'adaptive' SNPs, including mix of TPs and FPs; providing overview of the adaptive landscape.
Maximise TP	Patterns of genomic adaptation across major environmental gradients. Reference genome available.	More relaxed filtering. Larger amount of MD. Smaller permissible MAF. Pool candidate SNPs from multiple methods (don't just select overlapping results). Refine SNP sets with location and/or functional annotation.	Larger overall dataset of 'adaptive' SNPs; maximising number of TPs and improving dataset for downstream applications.
Minimise FP	Looking for candidate large-effect loci under selection for further investigation, especially where no genome is available.	More stringent filtering. Fewer MD. Consider MAF as a function of sample size and missing data (MAC) as well as impacts on significance. Focus on SNPs occurring in multiple programs using the Lotterhos et al., (2017) composite method.	Decreased absolute number of FPs at the expense of the number of TPs. Reduced identification of loci under weaker selection.
Identify loci under weak selection	Quantification of genome-wide levels of adaptation driven by environmental selection. Reference genome may or may not be available.	More relaxed filtering. Larger amount of MD. Lower permissible MAF with larger biological sample sizes. Refine SNP sets with location and/or functional annotation.	Increased power of GEA analyses. Greater number of loci providing more informative null models for GEA analyses. Improved ability to detect loci under weaker selection.

714

715

716 Conclusions

717 While we provide a filtering roadmap that enables users to understand how filtering might affect
718 GEA outputs, all organisms and datasets we study are unique, and the questions developed for
719 each will be different. Therefore, there is no universally 'best' way to perform filtering for GEA
720 analyses. Datasets should be developed in ways that best fit the objectives of the study (some
721 possible examples and recommendations are given in Table 4). Another important component
722 that we have not addressed, and is outside the scope of this study, is the use of genomic
723 resources for the betterment of GEA outputs. Additional genomic resources, such as an
724 annotated reference genome, provide further chances to refine the SNP sets used for
725 downstream analyses or applications. For example, it might be useful to examine whether SNPs
726 that putatively mediate local adaptation are located near genes whose function is relevant to the
727 environmental variable (Manel et al., 2016), or whose expression is induced by relevant
728 environmental challenges. Collectively, if a large proportion of putatively adaptive SNPs are
729 located near genes with relevant functions, it might promote confidence in the associations, and
730 their application to management actions.

731

732 Identifying true adaptive variants is difficult, particularly for non-model organisms, and this is
733 true even when strengths-of-selection are large. When we try to create and use the most
734 complete datasets through stringent filtering, we filter out many of those strongly adaptive SNPs
735 that are likely to be identified as TPs. When we have fewer putatively adaptive SNPs, then the
736 noise of FPs might lead to spurious adaptive signals through predictions, as we show. On the
737 other hand, if we filter our datasets more liberally, the adaptive signal seems to overpower
738 spurious signals. Together, as we identify clearer signals of adaptation, we are likely to better
739 understand how non-model species have adapted to the environment, moving the field of
740 landscape genomics toward a more complete understanding of our natural systems.

741 Data accessibility

742 All data will be uploaded to dryad upon acceptance and R code will be made available through
743 github or dryad.

744 Author contributions

745 CA developed the original idea. All authors contributed to further development of the idea at a
746 workshop hosted at Western Sydney University. CA, TH, KM, PH, RA, and JB developed the
747 code and analytics pipeline. CA and RJ wrote the first draft. All authors edited various versions
748 of the manuscript.

749 Acknowledgements

750 We acknowledge Prof James Seeb's use of "F-word" at the CONGEN 2013 meeting which is
751 discussed in Andrews & Luikart (2014) manuscript. This paper gained traction at the Eucalyptus
752 Australia meeting in 2019 in Hobart, where there was interest to compare adaptive patterns
753 across eucalypt species, and therefore thank the organisers and presenters at the meeting for
754 the opportunity to meet and discuss the genomics of eucalypts.

755

756 References

- 757 Ahrens, C. W., Byrne, M., & Rymer, P. D. (2019). Standing genomic variation within coding and
758 regulatory regions contributes to the adaptive capacity to climate in a foundation tree
759 species. *Molecular Ecology*, 28(10), 2502–2516.
- 760 Ahrens, C. W., Rymer, P. D., Stow, A., Bragg, J., Dillon, S., Umbers, K. D. L., & Dudaniec, R. Y.
761 (2018). The search for loci under selection: trends, biases and progress. *Molecular Ecology*,
762 27(6), 1342–1356.
- 763 Ahrens, C.W., James, E.A., Miller, A.D., Ferguson, S., Aitken, N.C., Jones, A.W., Lu-Irving, P.,
764 Borevitz, J.O., Cantrill, D.J. and Rymer, P.D. (2020). Spatial, climate, and ploidy factors drive
765 genomic diversity and resilience in the widespread grass *Themeda triandra*. *Molecular*
766 *Ecology*, doi:10.1111/mec.15614

- 767 [dataset] Ahrens, C.W., Jordan, R., Bragg, J., Harrison, P.A., Hopley, T., Bothwell, H.,... (2020).
768 Regarding the F-word: the effects of data Filtering on inferred genotype-environment
769 associations. DOI: (to be provided upon acceptance via dryad – data and R code)
- 770 Andrews, K. R., & Luikart, G. (2014). Recent novel approaches for population genomics data
771 analysis. *Molecular Ecology*, 23(7), 1661–1667.
- 772 Bay, R. A., Harrigan, R. J., Le Underwood, V., Gibbs, H. L., Smith, T. B., & Ruegg, K. (2018).
773 Genomic signals of selection predict climate-driven population declines in a migratory bird.
774 *Science*, 359(6371), 83-86.
- 775 Binks, R. M., Gibson, N., Ottewell, K. M., Macdonald, B., & Byrne, M. (2019). Predicting
776 contemporary range-wide genomic variation using climatic, phylogeographic and
777 morphological knowledge in an ancient, unglaciated landscape. *Journal of Biogeography*,
778 46(3), 503–514.
- 779 Browne, L., Wright, J. W., Fitz-Gibbon, S., Gugger, P. F., & Sork, V. L. (2019). Adaptational lag
780 to temperature in valley oak (*Quercus lobata*) can be mitigated by genome-informed assisted
781 gene flow. *Proceedings of the National Academy of Sciences*, 116(50), 25179–25185.
- 782 Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., ... Lander, E. S. (1999).
783 Characterization of single-nucleotide polymorphisms in coding regions of human genes.
784 *Nature Genetics*, 22(3), 231–238.
- 785 Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM 2: Fast and Accurate
786 Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology
787 and Evolution*, 36(4), 852–860.
- 788 Coop, G., Witonsky, D., Rienzo, A. D., & Pritchard, J. K. (2010). Using environmental
789 correlations to identify loci underlying local adaptation. *Genetics*, 185(4), 1411–1423.
- 790 Costa e Silva, J., Potts, B., Harrison, P. A., & Bailey, T. (2019). Temperature and rainfall are
791 separate agents of selection shaping population differentiation in a forest tree. *Forests*,
792 10(12), 1145.
- 793 Díaz-Arce, N., & Rodríguez-Ezpeleta, N. (2019). Selecting RAD-Seq data analysis parameters
794 for population genetics: the more the better? *Frontiers in Genetics*, 10, 533.
- 795 Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B,
796 Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B,
797 Skidmore AK, Zurell D, Lautenbach S (2013) Collinearity: a review of methods to deal with it
798 and a simulation study evaluating their performance. *Ecography*, 36, 27-46.
- 799 Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces
800 for global land areas. *International Journal of Climatology*, 37(12), 4302–4315.
- 801 Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for
802 detecting multilocus adaptation with multivariate genotype–environment associations.
803 *Molecular Ecology*, 27(9), 2215–2233.
- 804 François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in
805 genome scans for selection. *Molecular Ecology*, 25(2), 454–469.

- 806 Fu, Y.-B. (2014). Genetic diversity analysis of highly incomplete SNP genotype data with
807 imputations: an empirical assessment. *G3: Genes|Genomes|Genetics*, 4(5), 891–900.
- 808 Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., ... Luikart,
809 G. (2016). Genomics in conservation: case studies and bridging the gap between data and
810 application. *Trends in Ecology & Evolution*, 31(2), 81–83.
- 811 Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-
812 specific covariates. *Genetics*, 201(4), 1555–1579.
- 813 Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., ... Estoup, A. (2012).
814 The effect of RAD allele dropout on the estimation of genetic variation within and between
815 populations. *Molecular Ecology*, 22(11), 3165–3178.
- 816 Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R., & Amos, C. I. (2008). Shifting
817 paradigm of association studies: value of rare single-nucleotide polymorphisms. *The
818 American Journal of Human Genetics*, 82(1), 100–112.
- 819 Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies.
820 *Genetics*, 195(1), 205–220.
- 821 Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., ... Luikart,
822 G. (2018). Recent advances in conservation and population genomics data analysis.
823 *Evolutionary Applications*, 11(8), 1197–1211.
- 824 Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., ... Whitlock,
825 M. C. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and
826 future directions. *The American Naturalist*, 188(4), 379–397.
- 827 Hong, E. P., & Park, J. W. (2012). Sample size and statistical power calculation in genetic
828 association studies. *Genomics & Informatics*, 10(2), 117–122.
- 829 Jeffreys, H. (1961). *Theory of probability*, 3rd Edn Oxford: Oxford University Press. Oxford, UK.
- 830 Jordan, R., Hoffmann, A. A., Dillon, S. K., & Prober, S. M. (2017). Evidence of genomic
831 adaptation to climate in *Eucalyptus microcarpa*: implications for adaptive potential to
832 projected climate change. *Molecular Ecology*, 26(21), 6002–6020.
- 833 Klein, R. J. (2007). Power analysis for genome-wide association studies. *BMC Genetics*, 8(1),
834 58.
- 835 Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population
836 structure inference with genomic data sets. *Molecular Ecology Resources*, 19(3), 639–647.
- 837 Lotterhos, K. E., Card, D. C., Schaal, S. M., Wang, L., Collins, C., & Verity, B. (2017).
838 Composite measures of selection can improve the signal-to-noise ratio in genome scans.
839 *Methods in Ecology and Evolution*, 8(6), 717–727.
- 840 Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral
841 parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, 23(9), 2178–
842 2192. doi: 10.1111/mec.12725

- 843 Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local
844 adaptation depends on sampling design and statistical method. *Molecular Ecology*, 24(5),
845 1031–1046.
- 846 Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A.
847 (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA
848 sequencing for genome scans of adaptation. *Molecular Ecology Resources*, 17(2), 142–152.
- 849 Luu, K., Bazin, E., & Blum, M. G. B. (2016). pcadapt : an R package to perform genome scans
850 for selection based on principal component analysis. *Molecular Ecology Resources*, 17(1),
851 67–77.
- 852 Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., & Aurelle, D.
853 (2016). Genomic resources and their influence on the detection of the signal of positive
854 selection in genome scans. *Molecular Ecology*, 25(1), 170–184.
- 855 Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C.
856 (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de
857 novo assembly optimization for population genetic inference. *Molecular Ecology Resources*,
858 15(1), 28–41.
- 859 Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid
860 them. *Molecular Ecology*, 24(13), 3223–3231.
- 861 de Mita, S., Thuillet, A., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., & Vigouroux, Y. (2013).
862 Detecting selection along environmental gradients: analysis of eight methods and their
863 effectiveness for outbreeding and selfing populations. *Molecular Ecology*, 22(5), 1383–1399.
- 864 Morin, P. A., Martien, K. K., & Taylor, B. L. (2009). Assessing statistical power of SNPs for
865 population structure and conservation studies. *Molecular Ecology Resources*, 9(1), 66–73.
- 866 Moskvina, V., Craddock, N., Holmans, P., Owen, M. J., & O'Donovan, M. C. (2006). Effects of
867 differential genotyping error rate on the type I error probability of case-control studies.
868 *Human Heredity*, 61(1), 55–64.
- 869 Nadeau, S., Meirmans, P. G., Aitken, S. N., Ritland, K., & Isabel, N. (2016). The challenge of
870 separating signatures of local adaptation from those of isolation by distance and colonization
871 history: the case of two white pines. *Ecology and Evolution*, 6(24), 8649–8664.
- 872 Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013).
873 Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*,
874 22(11), 2841–2847.
- 875 Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes
876 for population genomics: an empirical study from an Amazonian plant species. *Molecular
877 Ecology Resources*, 17(6), 1136–1147.
- 878 O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These
879 aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists.
880 *Molecular Ecology*, 27(16), 3193–3206.

- 881 Orsini, L., Mergeay, J., Vanoverbeke, J., & Meester, L. (2013). The role of selection in driving
882 landscape genomic structure of the waterflea *Daphnia magna*. *Molecular Ecology*, 22(3),
883 583–601.
- 884 Pool, J. E., Hellmann, I., Jensen, J. D., & Nielsen, R. (2010). Population genetic inference from
885 genomic sequence variation. *Genome Research*, 20(3), 291–300.
- 886 Prober, S. M., Potts, B. M., Bailey, T., Byrne, M., Dillon, S., Harrison, P. A., ... Vaillancourt, R.
887 E. (2016). Climate adaptation and ecological restoration in eucalypts. *Proceedings of the*
888 *Royal Society of Victoria*, 128(1), 40.
- 889 Razgour, O., Forester, B., Taggart, J. B., Bekaert, M., Juste, J., Ibáñez, C., ... Manel, S. (2019).
890 Considering adaptive genetic variation in climate change vulnerability assessment reduces
891 species range loss projections. *Proceedings of the National Academy of Sciences*, 116(21),
892 201820663.
- 893 Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical
894 guide to environmental association analysis in landscape genomics. *Molecular Ecology*,
895 24(17), 4348–4370.
- 896 Schlamp, F., Made, J. van der, Stambler, R., Chesebrough, L., Boyko, A. R., & Messer, P. W.
897 (2015). Evaluating the performance of selection scans to detect selective sweeps in
898 domestic dogs. *Molecular Ecology*, 25(1), 342–356.
- 899 Sgrò, C. M., Lowe, A. J., & Hoffmann, A. A. (2011). Building evolutionary resilience for
900 conserving biodiversity under climate change. *Evolutionary Applications*, 4(2), 326–337.
- 901 Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B.
902 W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream
903 population genetic inference. *Methods in Ecology and Evolution*, 8(8), 907–917.
- 904 Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., & Neale, D. B. (2013). Putting
905 the landscape into the genomics of trees: approaches for understanding local adaptation and
906 population responses to changing climate. *Tree Genetics & Genomes*, 9(4), 901–911.
- 907 Sork, Victoria L. (2017). Genomic studies of local adaptation in natural plant populations.
908 *Journal of Heredity*, 109(1), 3–15.
- 909 Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing genome-wide
910 association studies: sample size, power, imputation, and the choice of genotyping chip.
911 *PLoS Genetics*, 5(5), e1000477.
- 912 Storz, J. F. (2005). Using genome scans of DNA polymorphism to infer adaptive population
913 divergence. *Molecular Ecology*, 14(3), 671–688.
- 914 Tabangin, M. E., Woo, J. G., & Martin, L. J. (2009). The effect of minor allele frequency on the
915 likelihood of obtaining false positives. *BMC Proceedings*, 3(Suppl 7), S41.
- 916 Thornhill, A. H., Crisp, M. D., Külheim, C., Lam, K. E., Nelson, L. A., Yeates, D. K., & Miller, J. T.
917 (2019). A dated molecular perspective of eucalypt taxonomy, evolution and diversification.
918 *Australian Systematic Botany*, 32(1), 29–48.

919 Vu, G. T. H., Schmutzer, T., Bull, F., Cao, H. X., Fuchs, J., Tran, T. D., ... Schubert, I. (2015).
920 Comparative genome analysis reveals divergent genome size evolution in a carnivorous
921 plant genus. *The Plant Genome*, 8(3), 1–14.

922 Willing, E.-M., Dreyer, C., & Oosterhout, C. van. (2012). Estimates of genetic differentiation
923 measured by F_{ST} do not necessarily require large sample sizes when using many SNP
924 markers. *PLoS ONE*, 7(8), e42649.

925 Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common
926 statistical problems. *Methods in ecology and evolution*, 1(1), 3-14.

927

928

929