Excessive and asymmetrical removal of heterozygous sites by *maxSH* biases downstream population genetic inference: Implications for hybridization between two primroses

Running title: Exploring biases from bioinformatics processes

Jie Zhang[1, 2, 3], Francisco Pina-Martins[4], Zu-Shi Jin[5], Yong-Peng Cha[1,3], Zu-Yao Liu[6], Jun-Chu Cha[1,3] Jian-Li Zhao[1, 3], Qing-Jun Li[1, 3]

1. Laboratory of Ecology and Evolutionary Biology, School of Ecology and Environmental Sciences, Yunnan University, Kunming, 650500, China

2. Institute of International Rivers and Eco-security, Yunnan University, Kunming, 650500, China

3. Yunnan Key Laboratory of Plant Reproductive Adaptation and Evolutionary Biology, Yunnan University, Kunming, 650500, China

4. Computational Biology and Population Genomics Group, Departamento de Biologia Animal, Faculdade de Ciências, Centre for Ecology, Evolution and Environmental Changes, Universidade de Lisboa, 1649-019 Lisboa, Portugal

5. School of Food Science, Tibet Agriculture and Animal Husbandry University, Nyingchi, 860000, China

6. Division of Evolutionary Ecology, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland.

Abstract

Techniques of reduced-representation sequencing (RRS) have revolutionized ecological and evolutionary genomics studies. Precise establishment of orthologs is a critical challenge for RRS, especially when a reference genome is absent. The proportion of shared heterozygous sites across samples is an alternative criterion for filtering paralogs, as divergent lineages should be less likely to share heterozygosity. In the prevailing pipeline for variant calling of RRS data - PYRAD/IPYRAD, *maxSH* is an often overlooked parameter with implications to detecting and filtering paralogs according to shared heterozygosity. Using empirical GBS data of two primroses (*Primula alpicola* Stapf and *Primula florindae* Ward) and their putative hybrids, and extra datasets of Californian golden cup oaks, we explore the impact of *maxSH* on filtering paralogs and further downstream analyses. Our study sheds light on the simultaneous validity and risk of filtering paralogs using *maxSH*, and its significant effects on downstream analyses of outlier detection, population assignment, and demographic modelling, emphasizing the importance of attention to detail during bioinformatics processes. The mutual confirmation between results of population assignment and demographic modelling in this study suggested *maxSH* = 0.10 has a potentially excessive and asymmetrical effect on the removal of truly shared heterozygous sites as paralogs. These results indicate that hybridization origin hypotheses of putative hybrids represented by results with *maxSH* = 0.25 and 0.50 are more credible. In conclusion, we revealed the critical hazard of paralogs filtration according to sharing heterozygosity at first, so that we propose to use specific protocols, rather than *maxSH*, to filter potential paralogs for closely related lineages.

Key words

Heterozygous site, bioinformatics, demographic modelling, paralog, *Primula*

1. Introduction

Techniques of reduced-representation sequencing (RRS) such as restriction site-associated DNA sequencing (RADseq; Baird et al., 2008) and genotyping by sequencing (GBS; Davey et al., 2011) are increasingly prevalent, especially in species lacking reference genomes, having revolutionized ecological and evolutionary genomics studies for their effective generating of genome-wide molecular markers with a desirable cost (Rodríguez-Ezpeleta et al., 2016; McKinney et al., 2017). Such markers allow inferring not only basic statistics of population genetics, but also phylogenetic reconstruction (e.g. Escudero et al., (2014)), population clustering (e.g. Ortego et al., (2017)), local adaptation (e.g. Pina-Martins et al., (2019)), or demographic inference (e.g. Excoffier et al., (2013)). Nevertheless, RRS has obvious shortcomings, such as precise establishment of homologous loci; a key challenge for all sequencing techniques, which is particularly difficult in RRS, since sequence similarity is often a crucial criterion for assembling demultiplexed reads into orthologous loci (Ilut et al., 2014; Harvey et al., 2015; McCartney-Melstad et al., 2019).

Several bioinformatics pipelines and programs have been developed for RRS loci assembly and data analyses, designed with various algorithms, logic and computer languages, such as STACKS (Catchen et al., 2013), IPYRAD (Eaton & Overcast, 2020),

dDOCENT (Puritz et al., 2014), RAINBOW (Chong et al., 2012) and so on. While considerable attention has been paid to reveal and reduce biases associated with wet laboratory, sequencing process and species-specific genome properties, biases from downstream bioinformatics analyses are less concentrated (but see Illut, et al., 2014; Mastretta-Yanes et al., 2015; Shafer et al., 2017; O'Leary et al., 2018).

PYRAD/IPYRAD is one of the most widely-used programs for quality filtering and variant calling of RRS data with or without a reference genome, and is superior in handling paired-end sequencing data and INDEL variation (Eaton, 2014; Eaton & Overcast, 2020). In PYRAD/IPYRAD, "clustering threshold" is a key parameter with respect to establishing homology, with a default value 85%. Several articles have soundly evaluated its influence and proposed advice on choosing its optimal value (Ilut ea al., 2014; McCartney-Melstad et al., 2019). But another important parameter, perhaps severely overlooked in most studies, is the maximum number (or proportion) of shared heterozygous sites in a locus. It can be set as both integer and decimal, abbreviated as *maxSH* or *maxsharedH* in PYRAD and *max_shared_Hs_locus* in IPYRAD (Eaton, 2014, Eaton & Overcast, 2020, in order to avoid confusion, we unified its name as *maxSH* hereafter). This parameter allows identification of paralogs according to the extent of shared heterozygosity across samples. It is worth mentioning that, in PYRAD the default value of *maxSH* is 4, which means heterozygous site shared by more than four individuals will be removed. While in IPYRAD the default value of *maxSH* is 0.50 (50%), allowing half of tested samples to share any given heterozygous site. Given bi-allelic SNPs were only generated by mutation, sharing

heterozigosity among species could only be originated by incomplete lineage sorting (ILS), gene flow and sharing gene/genome duplication (Spofford, 1969; Innan and Kondrashov, 2010; Fijarczyk and Babik, 2015). On this basis, highly divergent species (lineages) should be less likely to share heterozygous sites regarding their respective evolutionary history, and heterozygous sites simultaneously presented in many samples are more likely to represent clustering of paralogs with a fixed difference rather than a true heterozygous site (Eaton et al. 2015; Eaton & Overcast, 2020). PYRAD was primitively designed for phylogenetic reconstruction, thus extreme low default value are benefit and reasonable. However, numerous studies have used this default value or some extreme low values to generate datasets for population genomics study on close related lineages (e.g. Cavender-Bares et al., 2015; Eaton et al. 2015; Ortego et al., 2017; Tonzo et al., 2020). Closely related lineages possess relatively high proportion of heterozygous sites (Sota and Vogler 2003, Lischer et al., 2014). Besides, the high probability of introgression, hybridization, and sharing ancestral polymorphism can potentially contribute to extensively sharing heterozygosity among close lineages. Therefore, truly shared heterozygous sites will be simultaneously filtered if we set a lower threshold of *maxSH* in population genomics studies. In this regard, tuning *maxSH* is essential for unbiased inference, and its influence on downstream analyses is to be expected. However, it remains undiscovered to what extent *maxSH* affects such downstream analyses and further biological inference in population genomics frameworks.

We mainly performed our empirical test on the influence of *maxSH* values using two distylous primroses, *Primula alpicola* Stapf and *P. florindae* Ward co-occuring in the Shergyla mountains, southeast of Qinghai-Tibet region. *Primula L.* (Primulaceae) is arguably a high-profile genus for heterostyly since Darwin's seminal book (Darwin, 1877), exhibiting extreme species richness and diversity in the eastern Sino-Himalayan region (Richards, 2003). Although frequent hybridization and introgression have been considered critical factors for the complex phylogenetic relationships within *Primula* (Richards, 2003; Ren et al., 2018), only a few natural hybridization events were well documented in the complex's distribution center (see Zhu et al., 2009; Ma et al., 2014; Xie et al., 2017). Natural hybridization between *P. alpicola* and *P. florindae* was reported according to field observations by Ward, and artificial crossing is compatible in the garden (Richards, 2003). These two species are somewhat difficult to distinguish in sympatry due to morphological similarity, but still can be identified according to some critical differences on leaves and flowers. Moreover, their putative hybrids have been identified in middle elevation area of the Shergyla mountains, where *P. alpicola* and *P. florindae* share most pollinators with relative long flowering phenology overlapping (personal observation).

In this work, we sequence two primrose species and their putative hybrids using GBS to mainly disentangle 1) whether *maxSH* is effective on handling potential paralogs; 2) How, and to what extent could downstream bioinformatics analyses be influenced regarding different *maxSH* thresholds; 3) whether biological inference can tolerate potential excessive removal of heterozygous sites resulting from low *maxSH*

thresholds. In order to further verify the influence of *maxSH* on population genomics studies, we also conducted extra demographic modelling for another datasets discussing the introgression between two Californian golden cup oaks.

2. Materials and Methods

2.1 Sample collection

In 2015 and 2018, we sampled *P. alpicola* and *P. florindae* in the putative hybrid zone (elevation = 3672.71 m, latitude = 29.6704°N, longitude = 94.7157°E) from the Shergyla mountains. We collected ten individuals of *P. alpicola*, nine individuals of *P. florindae*, and fifteen putative hybrids identified by their flower color and leaf shape traits from this zone. Furthermore, three *P. sikkimensis* individuals collected from higher elevation of Shergyla mountains were sampled as outgroup species. Fresh leaf samples were quickly dried and stored with silica gel until DNA extraction.

2.2 DNA extraction and sequencing

Total DNA extraction followed a modified CTAB protocol (Doyle & Doyle 1987). The purity and amount of all extracted DNA was assessed using Nanodrop 1000 and Agarose gel. “Genotyping by sequencing” (GBS) technique was used for genotyping DNA samples and obtaining high density SNPs. In brief, DNA was double digested using *Mse*I+*Hae*II restriction enzymes, and ligated Illumina adapters and barcodes. After libraries were constructed, Qubit 2.0 was used for preliminary quantification, then DNA samples were uniformly diluted to about 1 ng/μl. At last, libraries were

pooled, and then paired-end sequenced following the standard protocol using Illumina HiSeq PE150 platform by Novogene Bioinformatics Technology Co., Ltd., Beijing, China (www.novogene.cn).

2.3 Bioinformatics and data filtering

We assembled *de novo* loci and called SNPs using IPYRAD v.0.7.30. We kept non-target parameters of IPYRAD as default value for prescriptive quality control steps, then designed parameter assemblies to test and compare their influence on downstream analyses: 1) We clustered our quality-filtered reads considering three thresholds: 85%, 90% and 95%; 2) As *maxSH* was the main-tested parameter, we set it to two frequently adopted values: 0.10 and 0.50 plus an intermediate value of 0.25. Complete information for each parameter assembly is available in supplemental files as datafile S1. In order quantify filtered paralogous loci, we used the output file "stats.txt" from IPYRAD step 7 following McCartney-Melsted et al. (2019) to plot the percentage of flagged paralogs ([filtered_by_max_indels+filtered_by_max_snps+filtered_by_max_shared_het+filtered_by_max_alleles]/total_prefiltered_loci). We further filtered processed data for keeping only biallelic SNPs, requiring a minimum allele frequency (MAF) > 0.03, and setting missing data rate (proportion of samples does not contain data at a given SNP) as 60% using VCFTOOLS v.0.1.14 (Danecek et al., 2011) since the number of retained SNPs sharply decreased with further missing data constrains. In the final step, we kept only the center SNP per locus so that we can minimize the effect of linkage disequilibrium using a python script "*vcf_parser.py*"

(https://github.com/CoBiG2/RAD_Tools/blob/master/vcf_parser.py) as of commit "0893296". In total, 9 datasets were created representing assemblies resulting from all parameter combinations. Each dataset was entitled as the combination of the initial of tested parameter (clustering threshold and *maxSH*) and their representative value, like *c85m10.*

2.4 Outlier detection

Outlier SNP detection was performed to obtain unbiased population structure and further demographic modelling. We detected outlier loci using two programs: BAYESCAN v.2.1 (Foll & Gaggiotti, 2008) based on Bayesian approach and R package *pcadapt* v.4.3.2 (Luu et al., 2017) based on principal component analysis (PCA). For both programs, individuals were preliminarily grouped according to sampling categories. BAYESCAN was run using 20 pilot runs of length 5,000, a burn-in length of 50,000, a main output iterations of 10,000, a thinning interval of 10, and a detecting threshold of 0.05. For running *pcadapt*, we firstly evaluated the number of principal components using "score plot" wrapped in *pcadapt* due to the poor resolution of Cattell's graphic rule in our case, a list of candidate SNPs under an expected FDR $\alpha = 0.05$ were identified as outliers following a standard *pcadapt* workflow. Outliers identified by either program were excluded from population structure inference analyses.

2.5 Population structure and hybrids identification

We then inferred population genetic structure and preliminarily identified hybrids for each data set using STRUCTURE v.2.3.4 (Pritchard et al., 2000) wrapped in the program *Structure_threader* v.1.3.4 (Pina-Martins et al., 2017). This program is characterized by parallelizing multiple runs of genetic clustering software and automatically assessing the best *K* as well as drawing the "meanQ" plots. We used filtered SNPs with 20 independent runs for *K* values from 1 to 6 to estimate the optimal number of clusters with a burn-in of 100 000, followed by 200 000 Markov chain Monte Carlo (MCMC) repetitions. The best *K* was estimated according to the widely used ΔK method (Evanno et al., 2005) implemented in *Structure_threader*. Principal component analysis was also performed using an R script "*snp_pca_static.R*" (https://github.com/CoBiG2/RAD_Tools/blob/master/snp_pca_static.R) as of commit "bb2fc45", in order to improve presentation, we slightly tweaked this script for our case regarding colours.

2.6 Demographic modelling

We used FASTSIMCOAL2 v.2.6.0.2 (Excoffier et al., 2013) for comparing different demographic models via coalescent simulations. Because demographic modelling depends on well-defined population structure, three populations (demes) defined based on STRUCTURE results were prepared for modelling. The folded joint SFS and unbiased estimation of allele frequency were performed using the Python script *easySFS.py* (https://github.com/isaacovercast/easySFS) as of commit "aaf80ea", which can effectively downsample populations for generating input ".obs" files for FASTSIMCOAL2. Because no invariable loci were involved in our SFS, we enabled

demographic modelling by introducing a calculated effective population size of *P. alpicola* (Papadopoulou & Knowles, 2015; Ortego et al., 2017). As $Ne = \pi/4\mu$, we inferred average mutation rate per site per generation μ from *Arabidopsis thaliana* (Nordborg et al., 2005) following Gossmann et al. (2012) as it is the genetically closest species with known μ value. π value was computed in DNASP v.6.12.03 (Rozas, et al., 2017) using ".allele.loci" file containing both polymorphic and non-polymorphic loci generated by IPYRAD. Average generation time was set as 1 yr (personal observation). Finally, we roughly bounded upper limit of divergence time between *P. alpicola* and *P. florindae* as 25 Ma (million years ago) according to estimated time when *Primula* diverged from *Soldanella* (de Vos et al., 2014).

To disentangle how putative hybrids speciated, three models were designed, respectively describing putative hybrids diverged from *P. florindae*, putative hybrids diverged from *P. alpicola*, and putative hybrids originated from hybridization between *P. alpicola* and *P. florindae*, considering post-divergence asymmetric gene flow (Fig. S1). Meanwhile, three alternative models describing similar divergence scenarios but without gene flow were also prepared as comparison (Fig. S1). Each model was run 100 independent replicates following 250,000 simulations with 60 expectation-conditional maximization (ECM) cycles, a stop criterion of 0.001, and zero SFS removed using FASTSIMCOAL2. Akaike's information criterion (AIC) was used to select the best model. The replicate with the maximum estimated likelihood of each model was selected for AIC and ΔAIC calculation.

Because Ortego et al. (2017) has conducted the comparison of genetic clustering results on different parameter assemblies of both stacks and IPYRAD, we mainly performed the comparison of coalescent analyses results among the STACKS datasets author used for his downstream analyses and two available PYRAD datasets with *maxSH* = 0.1 available as supplementary material. We named two PYRAD datasets as oak_*c85* and oak_*c90* respectively according to their clustering threshold, and the STACKS datasets as oak_*stacks*. As demographic modelling has been done for STACKS datasets (Ortego et al., 2017, Table1, Table2, Fig 2). For two PYRAD datasets, we followed Ortego's filtering steps to extract the unlinked neutral bi-allelic SNPs, then we used *easySFS.py* for downsampling and generating input files for FASTSIMCOAL2 containing folded joint SFS information, respectively. The alternative models and execution of FASTSIMCOAL2 were also in line with Ortego et al. (2017). At last, we compared the difference of best model and corresponding parameter estimates among these three datasets.

3. Results

3.1 Sequencing output and variation in data processing

The number of usable paired-end sequence reads ranged from 3,108,740 to 7,799,062 with an average of 5,291,571 per sample (SD=1,063,312). Total loci assembled by IPYRAD increased with both clustering threshold and *maxSH* values, ranging from 33,328 to 88,630 (Table 1). Total SNPs called primarily by IPYRAD varied similarly to total loci, from 226,965 to 605,829, with an average of 422,815 (Table 1). The

number of filtered SNPs also varied similarly, ranging from 5276 to 24166 (Table 1), but its differences among different *maxSH* values are further enlarged despite changing clustering thresholds.

Changing either clustering threshold or *maxSH* values effectively alters the proportion of flagged paralogs, suggesting their remarked association to detect and filter potential paralogs (Fig. 1). The percentage of flagged paralogs steeply decreased (~3% to ~5%) when changing *maxSH* from 0.10 to 0.25, and then tend to be approximate (< 0.8%) between *maxSH* = 0.25 and 0.50. Changing clustering threshold values from 0.85 to 0.95 resulted in a stepwise reduction of flagged paralogs. In dataset *c85m10*, more than 25% of assembled loci were identified as paralogs, yet *c95m50* contained ~12% flagged paralogs (Fig. 1).

BAYESCAN and *pcadapt* showed a large discrepancy in detecting loci under selection: the mean number of detected outliers of all datasets was about 21 for BAYESCAN and 1036 for *pcadapt*. In addition, the number of detected outliers increases with *maxSH* value for *pcadapt* except for datasets with the highest clustering threshold, but decreases for BAYESCAN in all datasets. It is also worth noting that BAYESCAN detected no outliers on 4 of 6 parameter assemblies with *maxSH* = 0.25 and 0.50, and outliers common to both programs were only detected when *maxSH* = 0.10 (Table 1).

After all filtering steps, the final number of neutral SNPs used for genetic clustering and demographic modelling ranged from 5002 for *c85m10* to 22126 for *c95m50* (Table 1).

3.2 Variation in population assignment

Most bayesian clustering results confirmed that samples labeled as hybrids exhibited a genetic mixture of *P. alpicola* and *P. florindae*, especially for *maxSH* values of 0.25 and 0.50 (Fig. 2D-I). Additionally, differences between results of *maxSH* = 0.25 and 0.50 were relatively small when keeping clustering threshold constant. However, hybrid ancestry proportion of *P. florindae* was relatively increased in datasets with the lowest *maxSH* (Fig 2A-C). Particularly, five putative hybrids were genetically clustered to *P. florindae* for *c95m10* (Fig. 2C). For all bayesian clusters, the optimal K value is 2 (datafile S2). Interactive version of plots for all K values were all available in supplemental files as datafile S3.

PCA results were roughly consistent with those obtained by the bayesian clustering approach, especially the similarity between plots with *maxSH* = 0.25 and 0.50 (Fig. 3D-I). Besides, when *maxSH* > 0.10, PCA results supported *P. alpicola* and *P. florindae* as genetically separated clusters, and samples of putative hybrids were located between *P. alpicola* and *P. florindae*, indicating their genetically admixed background. Unexpectedly, these datasets collectively segregated four putative hybrids from other samples marked along PC1 or PC2. By contrast, part of putative hybrids always showed their genetic similarity to *P. florindae* along at least one PC when *maxSH* = 0.10 (Fig. 3A-C), particularly, plot of *c95m10* distinctively exhibited a fusion of genetic clusters (Fig. 3C).

3.3 Variation in demographic modelling

296 Choice of *maxSH* showed dramatic impact on demographic modelling (Fig. 4).
297 Results of modelling were directed to two drastically different scenarios, therefore, all
298 results of parameter assemblies with *maxSH* = 0.25 and 0.50 inferred putative hybrids
299 were speciated from hybridization between *P. alpicola* and *P. florindae* accompanied
300 by interspecific gene flow (Model C1; Fig. 4D-I; Table s4-s9). While the best model
301 of three datasets with *maxSH* = 0.10 alternatively fit the scenario that putative hybrids
302 were diverged from *P. florindae* with post-divergence asymmetric gene flow (Model
303 A1; Fig. 4A-C; Table s1-3). Additionally, for datasets *c95m10*, the model indicating
304 putative hybrids were speciated from hybridization (model C1, Fig.S2) was
305 statistically equivalent to the best model to some extent (ΔAIC = 3.85, Burnham &
306 Anderson, 1998).

307

308 Demographic estimations also varied dramatically between datasets, including split
309 time, effective population size, proportion of migrants, and migration rates (Fig. 4).
310 Seven of nine assemblies indicated *P. alpicola* and *P. florindae* diverged more than 20
311 Ma (Fig. 4B, D-I), while *c85m10* suggested they split about 13.5 Ma (Fig. 4A).. For
312 the speciation time of putative hybrids, three models with *maxSH* = 0.50 referring to
313 hybridization origin with gene flow all directed to near 0.18 Ma (Fig. 4G-I), while the
314 rest showed that speciation time of putative hybrids varied from 0.04 to 0.19 Ma
315 regardless of ancestral lineage. Unlike split time, effective population size inferences
316 were irregularly variable, yet all results suggested putative hybrids hold the smallest
317 effective population size. Besides, eight of nine models indicated expansion of

effective population size at first, then coming to the recent constriction. *P. florindae* contributed a higher proportion of ancestry (from 0.62 to 0.75) to the hybrid lineage than *P. alpicola* according to five of six models referring to hybridization origin with gene flow (Fig. 4E-I), only *c85m25* supported both *P. florindae* and *P. alpicola* contributed the same proportion (0.5) to the hybrid lineage (Fig. 4D).. Interestingly, regardless of how putative hybrids originated, All models with gene flow got higher AIC scores compared to models without gene flow in the same parameter assembly, and eight of them shared a similar gene flow pattern: weak or moderate (only in *c95m10*) continuous gene flow between *P. alpicola* and *P. florindae* ; asymmetric gene flow between *P. alpicola* and putative hybrids, varying from moderate to strong; moderate gene flow from putative hybrids to *P. florinade*, while the reverse was subtle (Fig. 4A-E, 4G-I). Yet *c95m25* supported a different gene flow pattern: moderate gene flow from *P. alpicola* to *P. florindae* was supported. Besides, contrary to previous pattern, gene flow from putative hybrids to *P. florindae* was subtle, while the reverse was strong (Fig. 4F). It is also worth mentioning that almost all results of parameter estimation for *c95m25* were distinct from other datasets.

2.7 Verification from PYRAD datasets of Californian golden cup oaks

The best model for both PYRAD datasets with *maxSH* = 0.1 is Model B1 (Table s10, s11), indicating the southern lineage of *Quercus chrysolepis* was diverged from *Q. tomentella*. which is the second-best model for coalescent results of oak_*stacks*. Additionally, ModelC1 representing hybridization origin was even not the second-best model for both oak_*c85* and oak_*c90*. Additionally, for oak_*c90*, modelA1 indicating

the southern lineage of *Q. chrysolepis* was diverged from the northern lineage of *Q. chrysolepis* was statistically equivalent to the best model (Model B1) to some extent (ΔAIC = 2.28, Burnham & Anderson, 1998).

The advent of reduced-representation sequencing (RRS) has definitely facilitated studies on ecology and evolution in depth (Twyford & Ennos, 2012; Andrews et al., 2016). However, complex software and absence of standard analyses pipelines could mislead the analyses process, requiring conclusions to be drawn with caution (Shafer et al., 2017). Although various studies are dedicated to exploring biases from bioinformatics analyses processes, more attention should be paid to each and every detail, due to their potentially immeasurable influence (Gautier et al., 2013; Arnold et al., 2013; Shafer et al., 2017). The comparison of different values of *maxSH* in this study showed that an undesirable filtration of shared heterozygous sites can have a large impact on downstream analyses in a population genomics framework, altering the final biological inference. Since we carried out this research using empirical data, parts of our results could merely reflect some unique characteristics of the two tested datasets. However, the consistent influence of *maxSH* presented by two different datasets, especially on demographic modelling, has bolstered our confidence to draw a conclusion that a strict *maxSH* threshold is improper when conducting population genomics analyses.

In this study, we filtered missing data using moderate thresholds. Changing missing data rate could inevitably bring variations into downstream analyses, but for this first approach, we avoided using too many variables, which may result in focus reduction.

In fact, although a handful of studies have discussed its influence, dealing with missing data remains controversial (Huang and Knowles, 2016; Paris et al., 2017; Shafer et al., 2017; Yi and Latch, 2021).

4.1 Influence of *maxSH* and its interactions with clustering threshold

Identification of paralogs has unendingly been a challenge we have to cope with, because it can certainly act, as demonstrated here, on almost all downstream analyses, for example, outlier detection (Table 1), phylogenetic reconstruction (Fitz-Gibbon et al., 2017; McCartney-Melstad et al., 2019), demographic inferences (Fig. 4; Shafer et al., 2017), and to some extent, population clustering (Fig. 2 & 3; Rodríguez-Ezpeleta et al., 2016). For the increasingly prevalent PYRAD/IPYRAD, McCartney-Melstad et al. (2019) have illustrated that clustering thresholds strongly affect paralogs filtering and subsequent phylogenetic resolution in detail. In this study, we confirmed its significance on filtering paralogs and additionally estimated its influence on typical population genomic analyses. More importantly, we demonstrated that an underestimated parameter of PYRAD/IPYRAD, *maxSH*, is as influential as clustering threshold on handling potential paralogs. A low threshold of *maxSH* has remarkably increased the proportion of flagged paralogs in the tested data set. What's more, downstream analyses were all significantly impacted by *maxSH*, since variation of population assignment and demographic modelling were far more closely associated with *maxSH*, rather than clustering threshold. This could be mainly interpreted by their totally different rules for filtering paralogs. For clustering threshold, paralogs are identified via comparison of sequencing similarity, a lower threshold could lead to

underestimation of the number of loci and thus undersplitting (Rodríguez-Ezpeleta et al., 2016). However, biases from undersplitting are somewhat unpredictable, as we cannot know what such information represents. By contrast, a lower threshold of *maxSH* can directly filter heterozygous sites across many samples, which should only be originated from interspecific gene flows or sharing ancestral polymorphism if it is a true heterozygous site (Fijarczyk and Babik, 2015). This could explain why changing *maxSH* can strikingly alter the choice of the best model for coalescent simulations, while clustering thresholds mainly impacted parameter estimates. Besides, other downstream analyses should also clearly be more vulnerable once we improperly filtered these informative sites as paralogs.

Besides the different expected behaviors of clustering threshold and *maxSH*, there may also be some unexpected interactions between them. Oversplitting due to extremely high clustering thresholds has been demonstrated to cause a split between true allelic variants of orthologous loci into putatively separate loci (McCartney-Melsted et al., 2019). On the one hand, exorbitant request of sequence similarity by high clustering threshold can directly limit the proportion of heterozygosity for assembled loci. On the other hand, a low *maxSH* value can further exclude loci regarding dissimilarity components. Thus interactions between clustering threshold and *maxSH* can lead to considerable but unaccounted decrease of genetic distance among lineages, which could explain the fusion of genetic clustering intensively represented by PCA result for *c95m10*. It is also strongly supported by STRUCTURE results of golden cup oaks. in the STACKS datasets, part of individuals of *Q. tomentella*

406 in CAT population and most individuals of *Q.chrysolepis* in MOJ, LAG, BER, GAB,
407 FIG, and HAS populations exhibited some extent genetic admixture. While both
408 proportion of individuals exhibiting genetic admixture and the extent of admixture has
409 decreased in two PYRAD datasets, and higher clustering threshold has resulted in
410 heavier decrease (Ortego et al., 2017, Fig. S4). Extending this study to other datasets
411 would be helpful for confirming how prevalent the issue really is. Yet it has already
412 implied an optimal clustering threshold is urgent, as it is the first and great influential
413 filtering step in PYRAD/IPYRAD.

414 Handling heterozygous sites has frequently been neglected for RRS data
415 processing. On the one hand, most studies do not clearly exhibit information on
416 whether heterozygous sites were dropped or retained. On the other hand, phasing
417 between loci is almost impossible when a reference genome is unavailable (Garrick et
418 al., 2010; Lischer et al., 2014). As a consequence, heterozygous sites are improperly
419 filtered or totally excluded. However, multidimensional information of introgression
420 and incomplete lineage sorting from heterozygous positions is undoubtedly precious
421 for population genomic studies tackling closely related lineages. Our study revealed
422 the fathomless influence from processing heterozygous sites at first, and illustrated
423 filtering paralogs according to shared heterozygosity is risky and unreliable.

424 **4.2 Interpretation of divergent biological inference**

425 Clustering threshold and *maxSH* biological inference were determined by tuning said
426 parameters, with special emphasis on *maxSH*. Individuals labeled as hybrids were

genetically admixed in most cases according to STRUCRURE and PCA. These results are in agreement with their intermediate morphology, indicating potential hybridization origin of putative hybrids. Demographic models with *maxSH* = 0.25 and 0.50 supported a hybridization origin of putative hybrids scenario. However, a low *maxSH* value shifts this conclusion to a putative divergent origin, especially when combined with high clustering threshold values. Given that patterns of gene flow across eight modelling results collectively indicated gene flow from putative hybrids to *P. florindae* is always several times higher than to *P. alpicola*, putative hybrids are less likely to share heterozygous sites with *P. alpicola*. When we set a small *maxSH* value, those limited shared heterozygous sites have to be preferentially filtered, while on the other hand, shared heterozygous sites between hybrids and *P. florindae* can be more likely retained. These results were also partially verified by the coalescent results for golden cup oaks. The best model of STACKS datasets without filtering sharing heterozygous sites indicated hybrid origin of the southern lineage of *Q.chrysolepis*, while two PYRAD datasets collectively tend to the model describing *Q.chrysolepis* was diverged from *Q.tomentella* with post-divergence gene flow. As such, the mutual confirmation between population assignment and demographic modelling illustrated that extreme *maxSH* has brought excessive and asymmetrical removal of truly sharing heterozygous sites as paralogs in this study. We thus tend to infer putative hybrids to be originated by hybridization between *P. alpicola* and *P. florindae*, with an asymmetrical pattern of gene flow.

448 Hybridization could be widespread and play vital role on the diversification of
449 *Primula* (Schimidt-Lebuhn et al., 2012; Boucher et al., 2016; Keller et al., 2021).
450 Besides, increasing evidences have proofed the existence of multiple gene/genome
451 duplication and their significance on the origin of heterostyly, the most famous feature
452 of *Primula* (Li et al., 2016; Huu et al., 2020; Potente et al., 2022). On this basis,
453 sharing heterozygous sites among close related *Primula* species could contain plenty
454 of both paralogs and truly heterozygous sites. It could be the reason why a harsh
455 *maxSH* threshold can sharply reduce the number of retained SNPs. While it also
456 suggested that we should depend on other specific ways to filter paralogs when
457 performing population genomics studies on *Primula* and other similar taxa.

458 ise e

459 **4.3 Differentiated behavior of similar pipelines and approaches**

460 The flourishing of sequencing technology has prompted software development around
461 all aspects of downstream analyses. Yet differences between underlying algorithms
462 and logic of different software can lead the same analysis to divergent inference. For
463 example, Chen et al. (2021) demonstrated incredibly different behaviors of two

464 mainstream lines—McDonald-Kreitman (MK) test and PAML test—for positive
465 selection detection. Considering outlier removal is not only prior for inferring
466 unbiased population structure and estimating demographic history, but also crucial for
467 tackling adaptive divergence, numerous approaches have been developed for

detecting outliers, mostly lending *F*st-related statistics as criteria, such as BAYESCAN, OUTFLANK (Whitlock & Lotterhos 2015), SELESTIM (Vitalis et al., 2014) and so on. Unlike them, *pcadapt* identifies outliers regarding their relationship with population structure ascertained with principal component analysis. In this study, *F*st based BAYESCAN and PCA based *pcadapt* showed distinct efficiency on outlier identification. *pcadapt* could always flag a large number of loci under selection, while BAYESCAN conservatively detected quite a few number of outliers, furthermore, for only part of parameter assemblies. Likewise, large discrepancy and limited intersection of outliers detected by BAYESCAN and *pcadapt* have been elaborated by several studies (e.g. Kotsakiozi et al., 2017, Bekkevold et al., 2019). Nevertheless, quite few studies took interpretation of their discrepancy into consideration. Luu et al. (2017) pointed out the power of BAYESCAN decreased sharply when admixed individuals are included, which has been verified by our results and partially explains the distinct behavior of BAYESCAN and *pcadapt* in this study. Except for the influence of admixed individuals, *maxSH* has resulted in extra difference for these two approaches. Turning up *maxSH* can strikingly increase the number of total SNPs, while it can also decrease the genetic distance (*F*st value) among populations. Thereby those *F*st methods would be more vulnerable to tuning *maxSH*. This could mainly explain the decrease of detected outliers when increasing *maxSH* and the absence of outliers in some parameter assemblies when *maxSH* = 0.25 and 0.50, while non-sensitive PCADAPT will detect more outliers along with the increasing total SNPs.

STRUCTURE and PCA are some of the most popular approaches for tackling population assignment in population genetics/genomics. In this study, results of these two approaches are mostly concordant. Yet, we also find a number of differences on how they respond to variations of parameter assemblies. Tuning *maxSH* has brought greater impact on PCA, because when *maxSH* =0.10, part of putative hybrids constantly can not separate from or *P. florindae* or *P. alpicola* in any PCs. Yet when clustering threshold = 0.85 and 0.90, putative hybrids still kept their genetic admixture in STRUCTURE. According to the identity of putative hybrids. Genetic clustering by STRUCTURE could have offered robuster results in this study. In *c95m10*, both STRUCTURE and PCA have showed the fusion of genetic clusters,This could be a result of oversplitting effect on reducing genetic distances among populations due to extreme clustering threshold values (Harvey et al., 2015; Rodríguez-Ezpeleta et al., 2016). At last, four putative hybrids were repetitively drifted to others. According to their higher missing data rate than other putative hybrids, we inferred their irregular drift should stem from the impute limitations for missing data in PCA (Yi and Latch, 2021).

Demographic modelling was most vulnerable to *maxSH* variation among all population genomics analyses in this study. As we performed demographic modelling using FASTSIMCOAL2, the filtration of sharing heterozygous sites can intensively alter site frequency spectrum, the most important impute information for modelling. Thus, estimation of the best model would be close related to tuning *maxSH*. Taking consideration of increasing popularity and significance of coalescent simulation in

population genomics, only a handful of studies shed light on how bioinformatics processes affect demographic modelling or inference (Harvey et al., 2015; Shafer et al., 2017). Based on our results, we would like to highlight again the extreme importance of precise establishment of orthologs before performing demographic modelling for reliable biological inference.

Conclusions

Overall, this study highlights the feasibility but risk of tuning *maxSH* values on filtering paralogs. Our results illustrate its remarkable effect on almost all downstream analyses within a population genomics framework. According to the mutual confirmation between population assignment and demographic modelling, we inferred that *maxSH* = 0.10 has brought excessive and asymmetrical removal of truly sharing heterozygous sites as paralogs into this study. On this basis, we tend to approve the hybrid origin of putative hybrids between *P. alpicola* and *P. florindae* with an asymmetrical gene flow pattern deserving further investigation.

Here we give some suggestions on how to minimize the influence of *maxSH* from excessive and asymmetrical removal of heterozygous sites. Foremost, no single value could be expected for *maxSH* to be universal for all studies. Setting optimal clustering threshold following McCartney-Melstad et al (2019) would be beneficial as we have demonstrated the amplified biases from interactions between clustering threshold and *maxSH*. Then no matter what kind of analyses are arranged for closely related lineages, especially those with potential hybridization or introgression, one should not rely on *maxSH* for filtering paralogs, when we use PYRAD to generate datasets for

population genomics study, do not forget to turn up this parameter as the number of half samples (same as the default value of IPYRAD). And if we use IPYRAD, just keep it as default value.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (U1202261, U1602263), We especially thanks the South-East Tibetan Plateau Station for Integrated Observation and Research of Alpine Environment for fieldwork assistance. We also thanks Joaquín Ortego, Evan McCartney-Melsted, Julio Rozas for their help with data analysis. The authors declare that there are no conflict of interests.

Data Accessibility

Sequencing Data for this study will be available as soon as acceptance at NCBI Sequennce Read Archives as PRJNA669915.

References

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17: 81–92.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, 22(11): 3179-3190.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson, EA. 2008. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *Plos One*, 3(10): e3376.

Bekkevold D, Höjesjö J, Nielsen EE, Aldvén D, Als TD, Sodeland M, Kent MP, Lien S, Hansen MM. 2020. Northern European *Salmo trutta* (L.) populations are genetically divergent across geographical regions and environmental gradients. *Evolutionary Applications*, 13(2): 400-416.

560 Burnham KP, Anderson DR. 1998. *Model selection and inference: a practical information-theoretic*
561 *approach*. New York: Springer.

562 Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: An analysis tool set for
563 population genomics. *Molecular Ecology*, 22(11): 3124-3140.

564 Cavender-Bares J, Gonzalez-Rodriguez A, Eaton DAR, Hipp AAL, Beulke A, Manos PS. 2015.
565 Phylogeny and biogeography of the American live oaks (*Quercus* subsection *Virentes*): a
566 genomic and population genetics approach. *Molecular Ecology*, 24(14): 3688-3687.

567 Chen, QP, Yang H, Feng X, Chen QJ, Shi SH, Wu CI, He ZW. 2021. Two decades of suspect evidence
568 for adaptive molecular evolution-Negative selection confounding positive selection signals.
569 *National Science Review*. doi: 10.1093/nsr/nwab217

570 Chong Z, Ruan J, Wu CI. 2012. Rainbow: an integrated tool for efficient clustering and assembling
571 RAD-seq reads. *Bioinformatics*, 28(21): 2732-2737.

572 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, 1000 Genomes
573 Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15):
574 2156–2158.

575 Darwin C. 1877. *The Different Forms of Flowers on Plants of the Same Species*. London: John Murray.

576 Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. 2011. Genome-wide genetic
577 marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics*,
578 12: 499-510.

579

580 de Vos, JM, Wüest RO, Conti E. 2014. Small and ugly? Phylogenetic analyses of the ″selfing
581 syndrome″ reveal complex evolutionary fates of monomorphic primrose flowers. *Evolution*,
582 68(4): 1042-1057.

583 Doyle JJ, Doyle JL. 1987. A Rapid DNA Isolation Procedure from Small Quantities of Fresh Leaf
584 Tissues. *Phytochem Bull*, 19: 11-15.

585 Eaton DAR. 2014. PYRAD: assembly of de novo RADseq loci for phylogenetic analyses.
586 *Bioinformatics*, 30(13): 1844-1849.

587 Eaton DAR, Hipp AL, González-Rodríguez A, Cavender-Bares J. 2015. Historical introgression among
588 the American live oaks and the comparative nature of tests for introgression. Evolution,
589 69(10): 2587-2601.

590 Eaton DAR, Overcast I. 2020. IPYRAD: Interactive assembly and analysis of RADseq datasets.
591 *Bioinformatics*, 36(8): 2592-2594.

592 Escudero M, Eaton DAR, Hahn M, Hipp AL. 2014. Genotyping-by-sequencing as a tool to infer
593 phylogeny and ancestral hybridization: A case study in *Carex* (Cyperaceae). *Molecular*
594 *Phylogenetics and Evolution*, 79: 359–367.

595 Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the
596 software STRUCTURE: A simulation study. *Molecular Ecology*, 14(8): 2611-2620.

597 Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust Demographic Inference
598 from Genomic and SNP Data. *Plos Genetics*, 9(10): e1003905

599 Fijarczyk A, Babik W. 2015. Detecting balancing selection in genomes: limits and prospects. *Molecular*
600 *ecology*, 24(14): 3529-3545.

601 Fitz-Gibbon S, Hipp AL, Pham KK, Manos PS, Sork VL. 2017. Phylogenomic inferences from
602 reference-mapped and de novo assembled short-read sequence data using RADseq sequencing
603 of California white oaks (*Quercus* section *Quercus*). *Genome*, 60(9): 743-755.

604 Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both
605 dominant and codominant markers: a Bayesian perspective. *Genetics*, 180(2): 977-993.

606

607 Garrick RC, Sunnucks P, Dyer RJ. 2010. Nuclear gene phylogeography using PHASE: dealing with
608 unresolved genotypes, lost alleles, and systematic bias in parameter estimation. *BMC*
609 *evolutionary biology*, 10(1): 1-17.

610 Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J, Estoup A. 2013. The
611 effect of RAD allele dropout on the estimation of genetic variation within and between
612 populations. *Molecular Ecology*, 22(11): 3165-3178.

613 Gossmann TI, Keightley PD, Eyre-Walker A. 2012. The effect of variation in the effective population
614 size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biology and*
615 *Evolution*, 4(5): 658-667.

616 Harvey MG, Judy CD, Seeholzer GF, Maley JM, Graves GR, Brumfield RT. 2015. Similarity
617 thresholds used in DNA sequence assembly from short reads can reduce the comparability of
618 population histories across species. *PeerJ*, 3: e895.

619 Huang H, Knowles LL. 2016. Unforeseen consequences of excluding missing data from next-
620 generation sequences: simulation study of RAD sequences. *Systematic Biology*, 65(3): 357–
621 365.

622 Ilut DC, Nydam ML, Hare MP. 2014. Defining loci in restriction-based reduced representation genomic
623 data from nonmodel species: sources of bias and diagnostics for optimal clustering. *BioMed*
624 *Research International*, 2014: 675158.

625 Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing
626 between models. *Nature Reviews Genetics*, 11(2): 97-108.

627 Kotsakiozi P, Richardson JB, Pichler V, Favia G, Martins AJ, Urbanelli S, Armbruster PA, Caccone A.
628 2017. Population genomics of the Asian tiger mosquito, *Aedes albopictus*: insights into the
629 recent worldwide invasion. *Ecology and evolution*, 7(23): 10143-10157.

630 Lischer HEL, Excoffier L, Heckel G. 2014. Ignoring heterozygous sites biases phylogenomic estimates
631 of divergence times: implications for the evolutionary history of *Microtus* voles. *Molecular*
632 *Biology and Evolution*, 31(4): 817-831.

633 Luu K, Bazin E, Blum MG. 2017. *pcadapt*: an R package to perform genome scans for selection based
634 on principal component analysis. *Molecular Ecology Resources*, 17(1): 67-77.

635 Ma YP, Xie WJ, Tian XL, Sun WB, Wu ZK, Richard M. 2014. Unidirectional hybridization and
636 reproductive barriers between two heterostylous primrose species in north-west Yunnan,
637 China. *Annals of Botany*, 113(5): 763-775.

638 Mao XG. Zhang JP. Zhang SY. Rossiter SJ. 2010. Historical male-mediated introgression in horseshoe
639 bats revealed by multilocus DNA sequence data. *Molecular Ecology*, 19(7): 1352-1366.

640 Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson, BC. 2015. Restriction site-
641 associated DNA sequencing, genotyping error estimation and de novo assembly optimization
642 for population genetic inference. *Molecular Ecology Resources*, 15(1): 28-41.

643 McCartney-Melstad E, Gidis M, Shaffer HB. 2019. An empirical pipeline for choosing the optimal
644 clustering threshold in RADseq studies. *Molecular Ecology Resources*, 19(5): 1195-1204.

645 McKinney GJ, Waples RK, Seeb LW, Seeb JE. 2017. Paralogs are revealed by proportion of
646 heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural
647 populations. *Molecular Ecology Resources*, 17(4): 656-669.

648 Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng HG, Bakker E, Calabrese P, Gladstone J,
649 Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA,
650 Shah C, Wall JD, Wang J, Zhao KY, Kalbfleisch T, Schulz V, Kreitman M, Bergelson, J.
651 2005. The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS biology*, 3(7): e196.

652 Ortego J, Gugger PF, Sork VL. 2018. Genomic data reveal cryptic lineage diversification and
653 introgression in Californian golden cup oaks (section *Protobalanus*). *New Phytologist*, 218(2):
654 804-818.

655 O'Leary SJ., Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. 2018. These aren′t the loci you′e
656 looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*,
657 27(16): 3193-3206.

658

659 Papadopoulou A, Knowles LL. 2015. Species-specific responses to island connectivity cycles: Refined
660 models for testing phylogeographic concordance across a Mediterranean Pleistocene
661 Aggregate Island Complex. *Molecular Ecology*, 24(16): 4252-4268.

662 Paris JR, Stevens JR, Catchen JM. 2017. Lost in parameter space: A road map for stacks. *Methods in*
663 *Ecology and Evolution*, 8(10): 1360–1373.

Pina-Martins F, Silva DN, Fino J, Paulo OS. 2017. *Structure_threader*: An improved method for automation and parallelization of programs STRUCTURE, FASTSTRUCTURE and *MavericK* on multicore CPU systems. *Molecular Ecology Resources*, 17(6): e268–e274.

Pina-Martins F, Baptista J, Pappas Jr. G, Paulo OS. 2019. New insights into adaptation and population structure of cork oak using genotyping by sequencing. *Global Change Biology*, 25(1): 337-350.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2): 945–959.

Puritz JB, Hollenbeck CM, Gold JR. 2014. *dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, 2: e431.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Richards J. 2003. *Primula*. Oregon: Timber Press, Inc.

Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, Álvarez P, Cotano U, Irigoien X. 2016, Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Molecular Ecology Resources*, 16(4): 991-1001.

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA Sequence Polymorphism analysis of Large Datasets. *Molecular Biology and Evolution*, 34(12): 3299-3302.

Ren T, Yang Y, Zhou T, Liu ZL. 2018. Comparative plastid genomes of *Primula* species: Sequence divergence and phylogenetic relationships. *International journal of molecular sciences*, 19(4): 1050.

Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JBW. 2017. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 8(8): 907-917.

Sota T, Vogler AP. 2003. Reconstructing species phylogeny of the carabid beetles *Ohomopterus* using multiple nuclear DNA sequences: heterogeneous information content and the performance of simultaneous analyses. *Molecular Phylogenetics and Evolution*, 26(1): 139-154.

Spofford JB. 1969. Heterosis and the evolution of duplications. *The American Naturalist*, 103(932): 407-432.

Tonzo V, Papadopoulou A, Ortego, J. 2020. Genomic footprints of an old affair: Single nucleotide polymorphism data reveal historical hybridization and the subsequent evolution of reproductive barriers in two recently diverged grasshoppers with partly overlapping distributions. *Molecular Ecology*, 29(12): 2254-2268.

699 Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. *Heredity*, 108(3): 179-
700 189.

701 Vitalis R, Gautier M, Dawson KJ, Beaumont MA. 2014. Detecting and measuring selection from gene
702 frequency data. *Genetics*, 196 (3): 799–817.

703 Whitlock MC, Lotterhos KE. 2015. Reliable Detection of Loci Responsible for Local Adaptation:
704 Inference of a Null Model through Trimming the Distribution of F_{ST}. *The American Naturalist*,
705 186(S1): S24-S36.

706 Xie YP, Zhu XF, Ma YP, Zhao JL, Li L, Li QJ. 2017. Natural hybridization and reproductive isolation
707 between two *Primula* species. *Journal of Integrative Plant Biology*, 59(8): 526-530.

708 Yi X, Latch EK. 2022. Nonrandom missing data can bias Principal Component Analysis inference of
709 population genetic structure. *Molecular Ecology Resources*, 22(2): 602-611.

710 Zhu XF, Li Y, Wu GL, Fang ZD, Li QJ, Liu JQ. 2009. Molecular and morphological evidence for
711 natural hybridization between *Primula secundiflora* franchet and *P. poissonii* franchet
712 (Primulaceae). *Acta Biologica Cracoviensia*, 51(2): 29-36.

Tables

Table 1 Summary of IPYRAD output and outlier detection of different parameter assemblies

Parameter assemblies	Total loci	Total SNPs	Filtered SNPs	BAYESCAN outliers	*pcadapt* outliers	Outliers by both	Neutral SNPs
c85m10	33328	226965	5276	25	267	18	5002
c85m25	45671	351147	14374	0	816	0	13558
c85m50	47652	371317	16346	0	836	0	15510
c90m10	44783	307539	7076	33	342	21	6722
c90m25	58940	453269	17362	0	876	0	16486
c90m50	61046	475112	19456	0	1130	0	18326
c95m10	70321	427733	10870	73	1948	60	8909
c95m25	86620	586421	22187	36	1090	0	21061
c95m50	88630	605829	24166	25	2015	0	22126

Notes: Total loci and Total SNPs were generated by IPYRAD with at least 20% individuals containing data at a given locus; Filtered SNPs were generated by total SNPs further filtered by missing data, minimum allele frequency and keeping only the center one SNP per locus; BAYESCAN outliers, outliers detected by the software BAYESCAN; *pcadapt* outliers, outliers detected by the R package *pcadapt*; Neutral SNPs, filtered SNPs with detected outlier removed.

Figure Legends

Fig. 1. Proportion of loci flagged as paralogs and filtered by IPYRAD. Results are grouped according to clustering threshold in order to highlight differences of flagged paralogs resulted from *maxSH*.

Fig. 2. Comparison of genetic clustering by the bayesian clustering approach for *P. alpicola*, *P.florindae* and their putative hybrids implemented in the program STRUCTURE, Only K = 2 is shown here for being the Best K value. Each column shared the same clustering threshold and each row shared the same *maxSH* value. In every plot, each individual is represented by a vertical bar for every independent plot and color composition of each bar is referred to the individual's ancestry.

Fig. 3. Comparison of genetic clustering by PCA approach for *P. alpicola*, *P. florindae* and their putative hybrids. Each column shared the same clustering threshold and each row shared the same *maxSH* value. In every plot, each individual is represented by a dot and dots are colored according to sampling classification.

Fig. 4. The optimal demographic model for nine tested parameter assembly indicated by AIC and ΔAIC computation. Value of estimated parameters for the best model are showed in each plot, including divergence time (T_{DIV}), admixture time (T_{ADMIX}) for admixture model, effective population size (θ), rates of gene flow (m), and proportion of lineages transfer (α).

Support Information

Fig. S1. Alternative demographic models for exploring the origin of putative hybrids. The only difference between upper three models and lower three is the existence of interspecific gene flow. Parameter estimation include divergence time (T_{DIV}), admixture time (T_{ADMIX}) for admixture model, effective population size (θ), rates of gene flow (*m*), and proportion of lineages transfer (α).

Fig. S2. Results of genetic clustering by the BAYESCAN clustering approach for p95_60_10 after removing six samples totally similar to *P. florindae* in genetics. Only K = 2 is shown here for being the Best K value. Each individual is represented by a vertical bar for every independent plot and color composition of each bar is referred to the individual's ancestry.

Fig. S3. Demographic model statistically equivalent to the best model for p95_60_10. Value of estimated parameters includes divergence time (T_{DIV}), admixture time (T_{ADMIX}) for admixture model, effective population size (θ), rates of gene flow (*m*), and proportion of lineages transfer (α).

Table S1. Comparison of demographic models for *c85m10*.

Table S2. Comparison of demographic models for *c90m10*.

Table S3. Comparison of demographic models for *c95m10*.

Table S4. Comparison of demographic models for *c85m25*.

Table S5. Comparison of demographic models for *c90m25*.

Table S6. Comparison of demographic models for *c95m25*.

Table S7. Comparison of demographic models for *c85m50*.

Table S8. Comparison of demographic models for *c90m50*.

Table S9. Comparison of demographic models for *c95m50*.

datafile S1. Complete information of each parameter assembly for running IPYRAD.

767 **datafile S2.** The best *K* estimation for each datasets according to Δ*K* method.

768 **datafile S3.** Interactive version of plots of STRUCTURE results for all K values.