nature reviews genetics

https://doi.org/10.1038/s41576-024-00738-6

Supplementary information

Next-generation data filtering in the genomics era

In the format provided by the authors and unedited

Supporting Information for: Next-generation data filtering in the genomics era

William Hemstrom*, Jared A. Grummer*, Gordon Luikart, Mark R. Christie*

Corresponding author information:
* William Hemstrom; email: whemstro@purdue.edu
* Jared Grummer; email: jared.grummer@flbs.umt.edu
* Mark Christie; email: christ99@purdue.edu

This PDF includes:

Supplementary Text
Supplementary Tables 1-4
Supplementary Figures 1-7

Supplementary Methods:

Empirical data preparation

We acquired empirical data for ten taxa from published studies with taxonomic coverage including mammals (killer whales [*Orcinus orca*][1], deer mice [*Peromyscus maniculatus*][2], humans [*Homo sapiens*][3], mountain goats [*Oreamnos americanus*][4], and white-tailed deer [*Odocoileus virginianus*][5]), arthropods (water fleas [*Daphnia pulex*][6], stoneflies [*Sweltsa coloradensis*][7], and monarch butterflies [*Danaus plexippus*][8]), fish (yellow perch [*Perca flavescens*][9]), and plants (*Arabidopsis thaliana*)[10] (Table S2). The data types included whole-genome sequencing, low-coverage whole-genome sequencing, exome capture, and restriction site-associated DNA (RAD) sequencing. When more than two populations were sampled, we randomly selected two with 30 individuals per population for filtering. Populations with fewer than 30 individuals were not sub-sampled. For datasets where it was possible, we also applied GATK's suggested hard-filters[11] (QD > 2, FS < 60, SOR > 3, MQ > 40, MQRankSum > -12.5, ReadPosRankSum < -8) and a genotype quality (GQ) cutoff of 13 prior to filtering. The *D. pulix* data was also filtered to remove loci with very high heterozygosities (>60%) to remove probably paralogous loci.

Simulated data preparation

We used the *scrm* coalescent simulator[12] via the *coala* R package[13] to simulate three genomic datasets under three different demographic histories: a neutral (static) scenario, a recent population bottleneck, and a recent population expansion (Table S3). For each model, we simulated three populations, all of which descended from a common ancestral population which split 1,000 generations before present to form populations A and (B + C). Populations B and C then split from each other 500 generations later. Population C then remained static for 450

generations, after which it either continued without change (neutral/static model), declined exponentially over five generations to 1/20th its original size (bottleneck model), or exponentially expanded ten-fold over the same time-frame (expansion model). Prior to demographic changes, all populations were held at a constant effective population size of 10,000. Gene flow between populations B and C was allowed following the population split at a rate of 0.1 migrants per generation. For each model, we sampled 30 individuals from populations B and C at the end of the simulation for 10 chromosomes, each with a length of 10mb and a recombination rate averaging at one per chromosome per cross per generation. Population C was used for all further analyses except F_{ST}, for which both B and C were used. An R markdown document with the code used to perform these simulations is available in the Supplemental Materials.

To simulate selection for a range of recombination rates (Table S3), we used the *msms* simulator[14], also via *coala*[13]. We used the same parameters as the neutral/static model, but each chromosome had a different recombination rate with selection on a single new mutation beginning 50 generations in the past. We varied recombination rates between 0.1 and 10 (results are reported for recombination rates of 0.1 and 1 in Box 1 of the main text). In all cases, we used a selection coefficient of 0.2 against the ancestral allele during selection. An R script with the code used to perform these simulations is available in the Supplemental Materials (Supplementary Notebook 3).

Filtering

We filtered each empirical and simulated dataset with a range of different filters and thresholds using the *filter_snps* function in the *snpR* R package[15]. Specifically, we used the following filters and thresholds (function arguments listed in italics in parentheses):

- MAF (*maf*): 0.02–0.01 in 0.01 increments using within-group filtering (*maf_facets = "pop"*) such that any locus with a MAF less than the threshold in *all* groups was removed. We did not filter at 0.01 because only 30 diploid individuals were included from each population, resulting in a minimum observable MAF of ~0.017 for polymorphic loci.

- HWP (*hwe*): $1x10^{-6}$–$1x10^{-2}$ and 0.05 in increments of factors of ten ($1x10^{-6}$, $1x10^{-5}$, and so on) using within-group filtering (*hwe_facets = "pop"*) such that loci were removed only if they were significantly out of HWP in any individual sample group. HWP was assessed using an exact test[16]. No corrections for multiple testing were conducted to ensure that identical filtering thresholds were used for all loci and to ensure conservative removal of loci out of HWP.

- Required % individuals genotyped (*min_ind*): 10–90% in increments of 10% such that loci were removed if they were not genotyped in at least the given percentage of individuals.

- Required % loci genotyped (*min_loci*): 10–90% in increments of 10% such that individuals were removed if they were not genotyped in at least the given percentage of individuals.

When testing different thresholds for a given parameter value, we generally held all other parameter values constant at these values:

- MAF = 0, MGC = 1. Note: MGC = 1 (*mgc = 1*) removes any loci sequenced in only one individual, regardless of the genotypic state of that individual such that loci observed in a single homozygous individual were still removed.
- HWP = $1x10^{-6}$
- Required % individuals genotyped: 70%
- Required % loci genotyped: 70%

In addition to the solitary filter variation iterations, we also varied required % individuals and loci genotyped together for their ranges (both values 10-90%) for all datasets and performed a full factorial comparison of our filter thresholds for HWP and MAF for the mountain goat (RAD) and stonefly datasets specifically.

Following filtering, we computed expected heterozygosity (H_E), observed heterozygosity (H_O), nucleotide diversity (π), F_{IS} and pairwise F_{ST} according to Weir and Cockerham[17], Tajima's D[18], Watterson's θ[19], Tajima's θ[18], a rarefaction-corrected measure of the number of segregating sites (Hemstrom and Christie, *in prep*), and a rarefaction-corrected estimate of the number of private alleles[20] for each population or pair of populations (where applicable). Lastly, we also estimated effective population sizes (N_e) for each population in each dataset using the LD method in the NeEstimator software[21] using only loci pairs on different chromosomes or scaffolds. All analyses were performed using the *snpR* R package[15]. Filtering R scripts, bash (shell) handling scripts, and parameter files are available at https://github.com/ChristieLab/filtering_simulation_paper.

Supplementary References:

1. Kardos, M. *et al.* Inbreeding depression explains killer whale population dynamics. *Nature Ecology & Evolution* **7**, 675–686 (2023).

2. Schweizer, R. M. *et al.* Broad Concordance in the Spatial Distribution of Adaptive and Neutral Genetic Variation across an Elevational Gradient in Deer Mice. *Molecular Biology and Evolution* **38**, 4286–4300 (2021).

3. Lowy-Gallego, E. *et al.* Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project [version 2; peer review: 2 approved]. *Wellcome Open Research* **4**, (2019).

4. Martchenko, D. & Shafer, A. B. A. Contrasting whole-genome and reduced representation sequencing for population demographic and adaptive inference: an alpine mammal case study. *Heredity* **131**, 273–281 (2023).

5. Kessler, C., Wootton, E. & Shafer, A. B. A. Speciation without gene-flow in hybridizing deer. *Molecular Ecology* **32**, 1117–1132 (2023).

6. Maruki, T., Ye, Z. & Lynch, M. Evolutionary Genomics of a Subdivided Species. *Molecular Biology and Evolution* **39**, msac152 (2022).

7. Malison, R. L. *et al.* Landscape Connectivity and Genetic Structure in a Mainstem and a Tributary Stonefly (Plecoptera) Species Using a Novel Reference Genome. *Journal of Heredity* **113**, 453–471 (2022).

8. Hemstrom, W. B., Freedman, M. G., Zalucki, M. P., Ramírez, S. R. & Miller, M. R. Population genetics of a recent range expansion and subsequent loss of migration in monarch butterflies. *Molecular Ecology* **31**, 4544–4557 (2022).

9. Schraidt, C. E. *et al.* Dispersive currents explain patterns of population connectivity in an ecologically and economically important fish. *Evolutionary Applications* **n/a**, (2023).

10. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).

11. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).

12. Staab, P. R., Zhu, S., Metzler, D. & Lunter, G. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* **31**, 1680–1682 (2015).

13. Staab, P. R. & Metzler, D. Coala: an R framework for coalescent simulation. *Bioinformatics* **32**, 1903–1904 (2016).

14. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065 (2010).

15. Hemstrom, W. & Jones, M. snpR: User friendly population genomics for SNP data sets with categorical metadata. *Molecular Ecology Resources* **23**, 962–973 (2023).

16. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics* **76**, 887–893 (2005).

17. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358–1370 (1984).

18. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585 LP – 595 (1989).

19. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276 (1975).

20. Kalinowski, S. T. Counting Alleles with Rarefaction: Private Alleles and Hierarchical Sampling Designs. *Conservation Genetics* **5**, 539–543 (2004).

21. Do, C. *et al.* NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources* **14**, 209–214 (2014).

22. Li, M.-X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics* **131**, 747–756 (2012).

Supplementary Tables:

Supplementary Table 1: Different types of filters available for genomic sequencing data.

Filter	Stage	Description
Base quality scores	*i*	Removal of reads with many poor-quality (likely mis-read) bases.
Poly-G tails	*i*	Removal of guanines ("G"s) erroneously called at the ends of reads on certain sequencing platforms.
Adapter/Barcode/Cut-site trimming	*i*	Removal of adapter, barcode, or cut-site sequences from the reads.
Adapter/Barcode/Cut-site mismatches	*i*	Removal of reads with sequences that do not match known adapter, barcode, or cut-site sequences.
Read K-mer distribution	*i, ii*	Removal of reads with too many very common or rare runs of base-pairs (K-mers).
Technical/PCR duplicates	*i, ii*	Thinning of technical or PCR duplicates down to a single representative read.
Alignment/Mapping scores	*ii*	Removal of reads that have mapping scores below a user-defined threshold.
Improperly paired reads (orientation and distance)	*ii*	Removal of paired-reads that are improperly paired (unexpectedly far apart or incorrectly oriented)
Stack depth of coverage	*ii*	Removal of loci "stacks" that have too low of a sequencing depth across samples; usually for reduced-representation sequencing.
Stack mismatches	*ii*	Removal of loci "stacks" that have too many mismatched base-pairs across samples; usually for reduced-representation sequencing.
Number of Alleles	*ii, iii*	Removal of genotypes, haplotypes, or "stacks" with too many possible alleles (usually > 2 for SNPs). Usually for computational efficiency, but also to remove potential errors.
Low coverage/Quality-by-depth	*iii*	Removal of individual called genotypes with coverage below a user-defined threshold. Joint "Quality-by-depth" often alternatively used.
Genotype Quality/Confidence	*iii*	Removal of individually called genotypes with genotyping confidence below a user-defined threshold. Joint "Quality-by-depth" often alternatively or additionally used.
High coverage	*iii*	Removal of individual called genotypes with coverage above a user-defined threshold (usually indicating errors in the reference, paralogs, or copy-number variants, all of which require additional investigation).

Supplementary Table 1: Continued from previous page.

Filter	Stage	Description
Insertion-deletions (Indels)	*iii*	Removal of insertions or deletions (indels), often required by many down-stream applications
Non-biallelic loci	*iii*	Removal of non-biallelic loci (for example, monomorphic or tri-allelic SNPs); required by many down-stream applications.
Allow/deny-listed variants	*iii*	Removal or inclusion of a set of user-defined loci. Common for methods that target specific loci or where specific variants are known to be problematic.
Variant Read Position	*iii*	Removal of variants that tend to occur in biased positions on shotgun-sequenced reads.
Missing data - per individual	*iii, iv*	Removal of individuals with called genotypes at fewer than a user-defined number of loci.
Missing data - per locus	*iii, iv*	Removal of loci with called genotypes at fewer than a user-defined number of individuals.
Minor allele frequency	*iii, iv*	Removal of loci with minor allele frequencies below a user-defined threshold.
Minor allele count	*iii, iv*	Removal of loci with a count of the minor allele below a user-defined threshold across samples.
Hardy-Weinberg proportions	*iii, iv*	Removal of loci out of Hardy-Weinberg proportions, typically below a user-defined p-value.
Strand Bias	*iii, iv*	Removal of loci where specific alleles are detected primarily on only the forward or reverse DNA strand.
Copy number variation	*iii, iv*	Removal of copy number variants. Often remain undiscovered.
Structural variants	*iii, iv*	Removal of structural variants, such as inversions. Often remain undiscovered.
Sex-linked loci	*iii, iv*	Removal of sex-linked loci, which may behave in unexpected ways or have biased statistical outcomes due to sex-specific sampling.
Paralogs - allelic imbalance/depth/heterozygosity	*ii, iii, iv*	Removal of reads aligned to paralogous genomic regions, where for recently diverged paralogs it can be unclear from which of the gene copies the read was sequenced. Additional analyses are required.

Supplementary Table 1: Continued from previous page.

Filter	Stage	Description
Mislabeling/Contamination	*iv*	Removal of individuals or loci that are likely mislabled, contaminated, or have similar issues. Can often be identified via PCA and other comparative analyses.
Transition-transversion bias	*iv*	Removal of loci from genomic regions with unexpected transition:transversion ratios.
F_{ST}/Selection Outliers	*iv*	Removal of outlier loci likely to be under selection. Useful for cases where putatively neutral processes specifically are of interest (for example, gene flow).

i = sequence QC (Quality control), *ii* = alignment to a reference, *iii* = variant discovery, and *iv* = data analysis. Note that stages *i* and *ii* constitute pre-variant filtering and stages *iii* and *iv* constitute post-variant filtering.

Supplementary Table 2: Empirical datasets used for filtering simulations.

Organism	Dataset Type	Number of SNPs	Number of Individuals	Reference
Arabidopsis	WGS	3,135,226	60	[10]
Monarch butterflies	RAD	238,368	54	[8]
Daphnia	WGS			[6]
Stoneflies	RAD	279,496	60	[7]
Yellow perch	WGS	6,586,547	57	[9]
Humans	WGS	17,458,468	60	[3]
Deer mice	Exome capture	5,373,633	55	[2]
Mountain goats	RADseq/ WGS	48,192/8,113,114	60/20	[4]
White-tailed deer	WGS	48,441,150	20	[5]
Killer whales	WGS	3,015,993	54	[1]

Supplementary Table 3: Simulated datasets used for filtering simulations.

Selection	Demographic History	Relevant Parameters	Number of SNPs	Number of Individuals
No Selection	Neutral/Static	$N_{curr} = N_{anc}$	237,698	60
No Selection	Large expansion in one population	$N_{curr} = 10N_{anc}$	247,922	60
No Selection	Large bottleneck in one population	$N_{curr} = 0.05N_{anc}$	198,783	60
Hard sweep at one locus in one population	Static	$s = 0.2$ $f_s = 1/2N$ $r = 0.1$ or 1 $t_s = 50$	21228/21744	60/60

N_{curr}: sample size at present
N_{anc}: ancestral sample size
s: selection coefficient against ancestral allele
f_s: frequency of the new mutation
r: recombination rate, average number per chromosome per generation
t_s: time (in generations) before present at which selection began

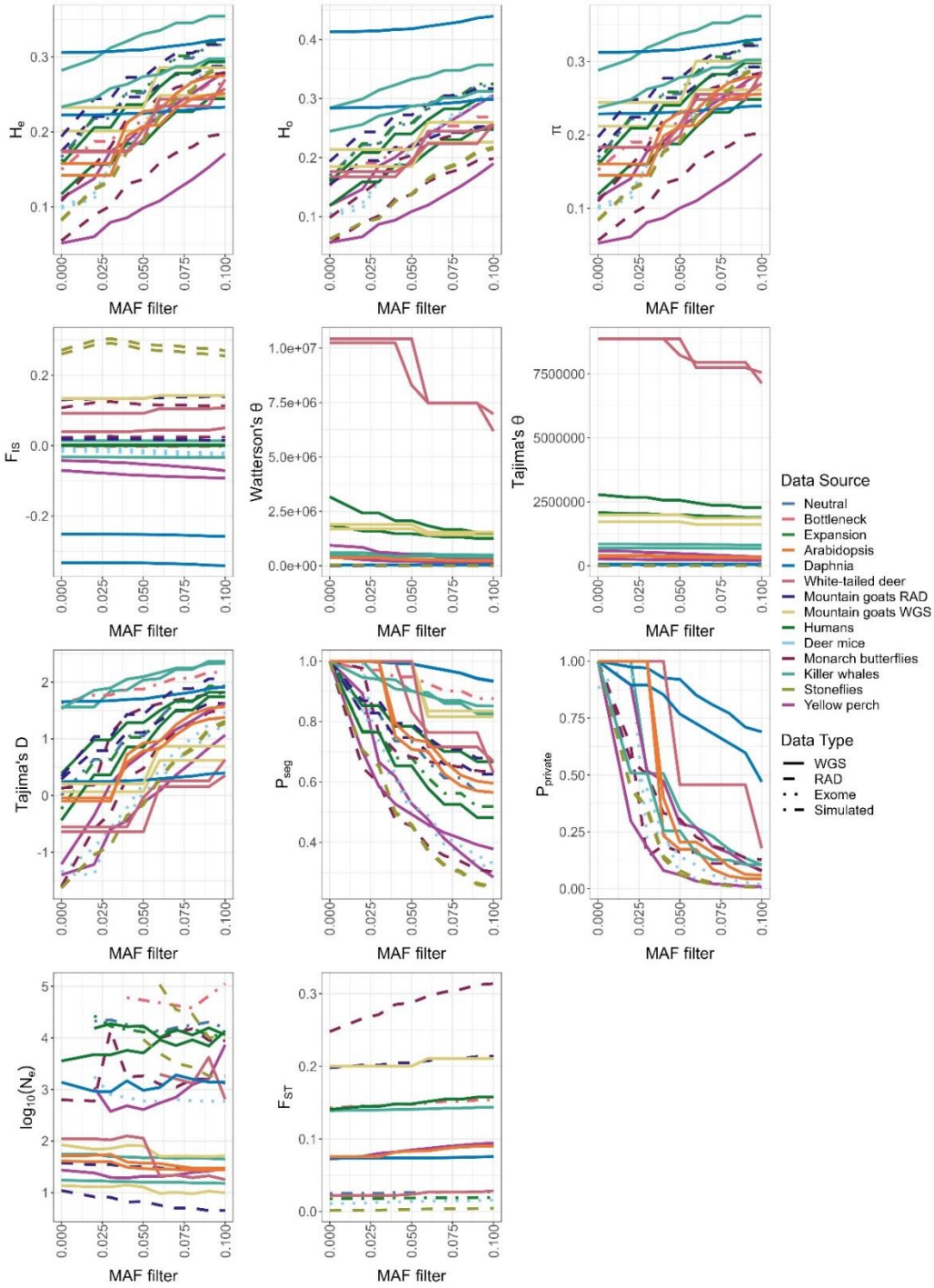
Supplementary Table 4: Justifications for filtering threshold recommendations in Table 1. Note: suggested MAC values may actually be stricter than the suggested MAF values for datasets with low sample sizes and should be adjusted accordingly. Some methods for specific questions may require filter values that differ from these suggestions.

Question/Approach	Individual missing data; <X% missing loci	Loci missing data; <X% missing individuals	MAC/MAF	LD	HWP
Population Structure	Poor quality individuals or loci may mask structure	Poor quality individuals or loci may mask structure	Higher MAF values can reveal additional population structure	Probably has no effect unless inversions or other factors drive clustering.	Need to keep a low pass filter for cryptic populations
Demography	Many demographic estimation approaches function well with small sample sizes but can be misled by poor quality individuals	Projection can reduce the impact of missing data.	Any measures which depend on SFS estimates are extremely biased by the removal of low frequency variants.	Non-independent loci can create misleading site frequency spectra. Not as essential for Tajima's D as it is for ABC, etc.	Paralog removal
Selection	Don't want to remove signal	Don't want to remove signal	Don't want to remove signal, but some methods require or recommend higher MAF (0.05)	Removing loci in LD can remove signals of hitchhiking/selective sweeps	Don't want to remove signal; less strict filtering to keep more loci
Genetic Diversity	Some metrics (# seg sites) can drop fast as you exclude too many individuals/loci (see Box 1 fig.)	Some metrics (# seg sites) can drop fast as you exclude too many individuals/loci (see Box 1 fig.)	Note that some metrics (like # seg sites) are impacted differently from others (like H_E)	Should not *usually* cause major impacts, but can if regions in high LD vary (chromosomal inversion). Can skew confidence intervals.	This range will usually capture most of the changes due to filtering (see Box 1 fig).

Supplementary Table 4: Continued from previous page

Question/Approach	Individual missing data; <X% missing loci	Loci missing data; <X% missing individuals	MAF/MAC	LD	HWP
Phylogenetics	High amounts of missing data within a locus can cause long branch attraction and incorrect topological inferences	For species-tree inference, must assure adequate representation of individuals in each species across gene trees	Autapomorphies/Singletons are not informative	Need independent evolutionary histories (e.g., un-linked)	Don't want anything with odd behavior throwing off signal
GWAS	If your method requires no missing data, use a high filter to avoid extra imputation	If your method requires no missing data, use a high filter to avoid extra imputation	Low frequency variants are typically uninformative unless sample sizes are very large.	Removal can cloud signals from linkage around causal genes. *p*-value correction is needed for multiple testing–see[22]	Selection on causal may rarely cause deviations. No filter to check, then do a permissive filter to keep most loci.
Mutation Detection	Cannot have missing data in parents, but skip over missing data in offspring (can't detect mutations but won't cause problems)	Cannot have missing data in parents, but skip over missing data in offspring (can't detect mutations but won't cause problems)	Many (new) mutations have very low frequencies	Irrelevant; don't want to remove potential mutations.	One test to get rid of paralogs, otherwise don't filter
Metagenomics/eDNA	NA	NA	Need multiple reads from a region for confidence (5+ or so)	Depends on context: no filter for a mix of species but consider otherwise	NA
Relatedness/ Pedigree Construction	Including all individuals is important	Only a few loci needed to infer a relationships but they need to be high quality.	Singletons don't help	Should not bias mean outcomes, but could change confidence intervals.	Only a few loci needed to infer a relationships but they need to be high quality.

Supplementary Figures:



Supplementary figure S1: Effects of **MAF filter** threshold on **non-standardized** change in F_{ST}, H_O, F_{IS}, Tajima's D, H_e, π, Watterson's θ, Tajima's θ, the proportion of the total (highest) number of segregating sites, the proportion of the total (highest) number of private alleles, and N_e, observed in each population and dataset. Note the drastic changes in F_{ST} in the monarch dataset, driven by strong differences in the site frequency spectra between populations. The removal of rare alleles substantially changes the distribution of allele frequencies and increases F_{ST} in one population which had undergone a substantial demographic expansion.

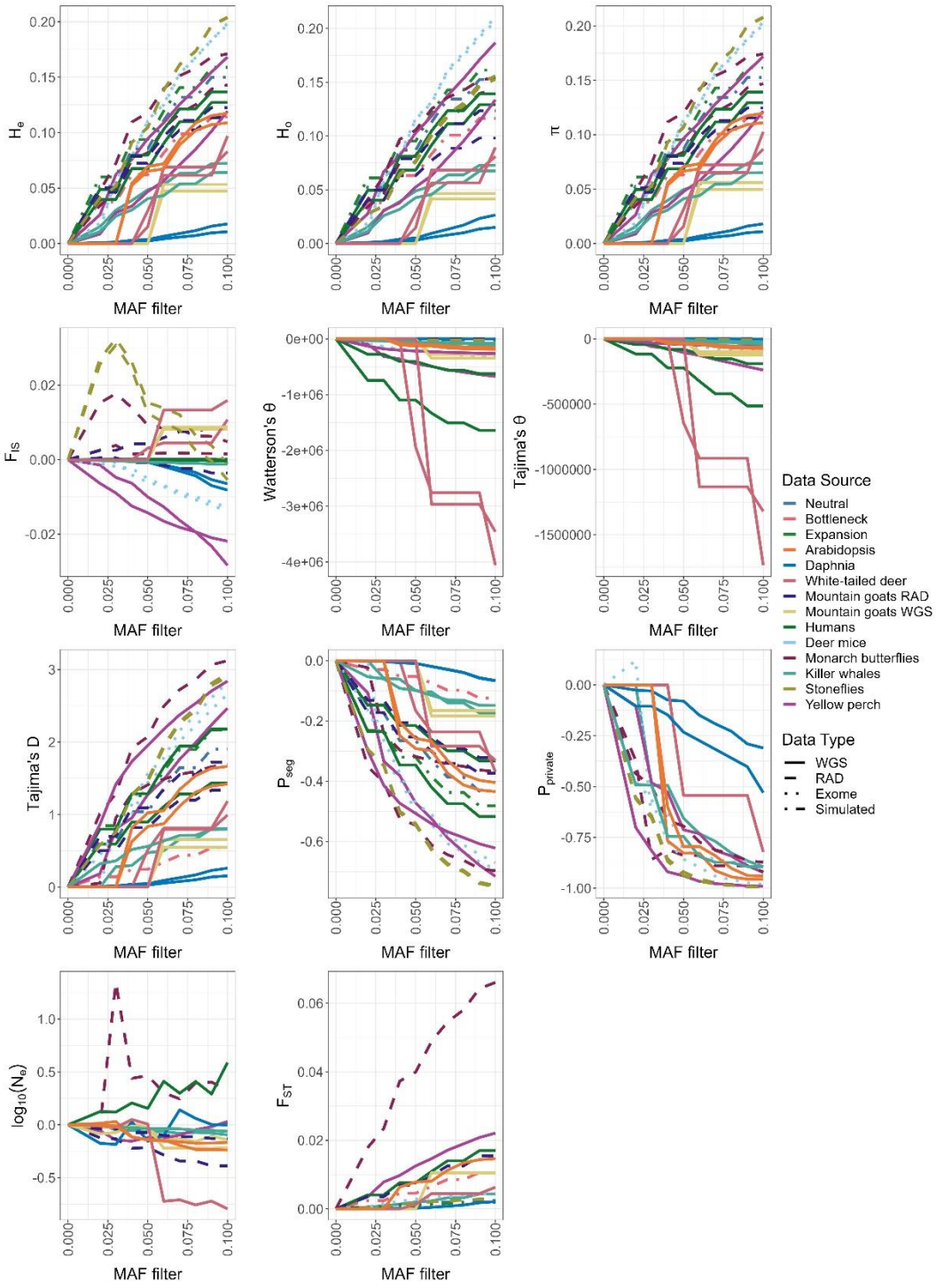


Figure S2: Effects of **MAF filter** threshold on **standardized** F_{ST}, H_O, F_{IS}, Tajima's D, H_e, π, Watterson's θ, Tajima's θ, the proportion of the total (highest) number of segregating sites, the proportion of the total (highest) number of private alleles, and N_e, observed in each population and dataset depending on MAF filtering stringency. Parameter values have been standardized to a value of 0 for the first MAF filter value (also 0) to highlight differences among studies. Some panels (H_O, F_{IS}, F_{ST}, Tajima's D, and P_{seg}) are also shown in Box 2 (main text), but are retained here for ease of comparison.

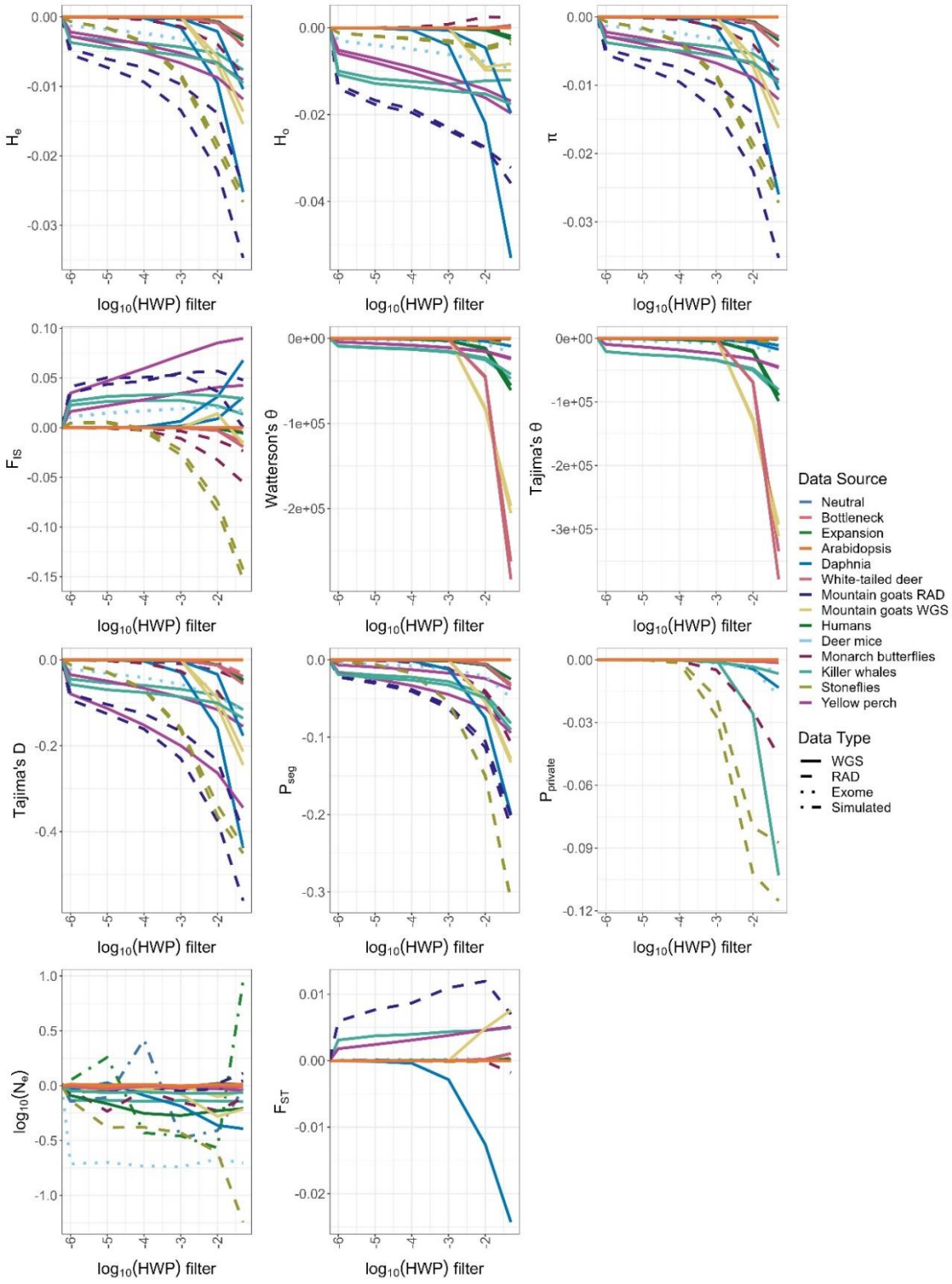


Figure S3: Effects of **HWP filtering** threshold on standardized F_{ST}, H_O, F_{IS}, Tajima's D, H_e, π, Watterson's θ, Tajima's θ, the proportion of the total (highest) number of segregating sites, the proportion of the total (highest) number of private alleles, and N_e. Parameter values have been normalised to show change by subtracting off the value observed with no filter. Note that RAD datasets tend to see more changes with increasing HWP filter stringency for many statistics, including substantial Tajima's D and F_{IS} changes in some cases.

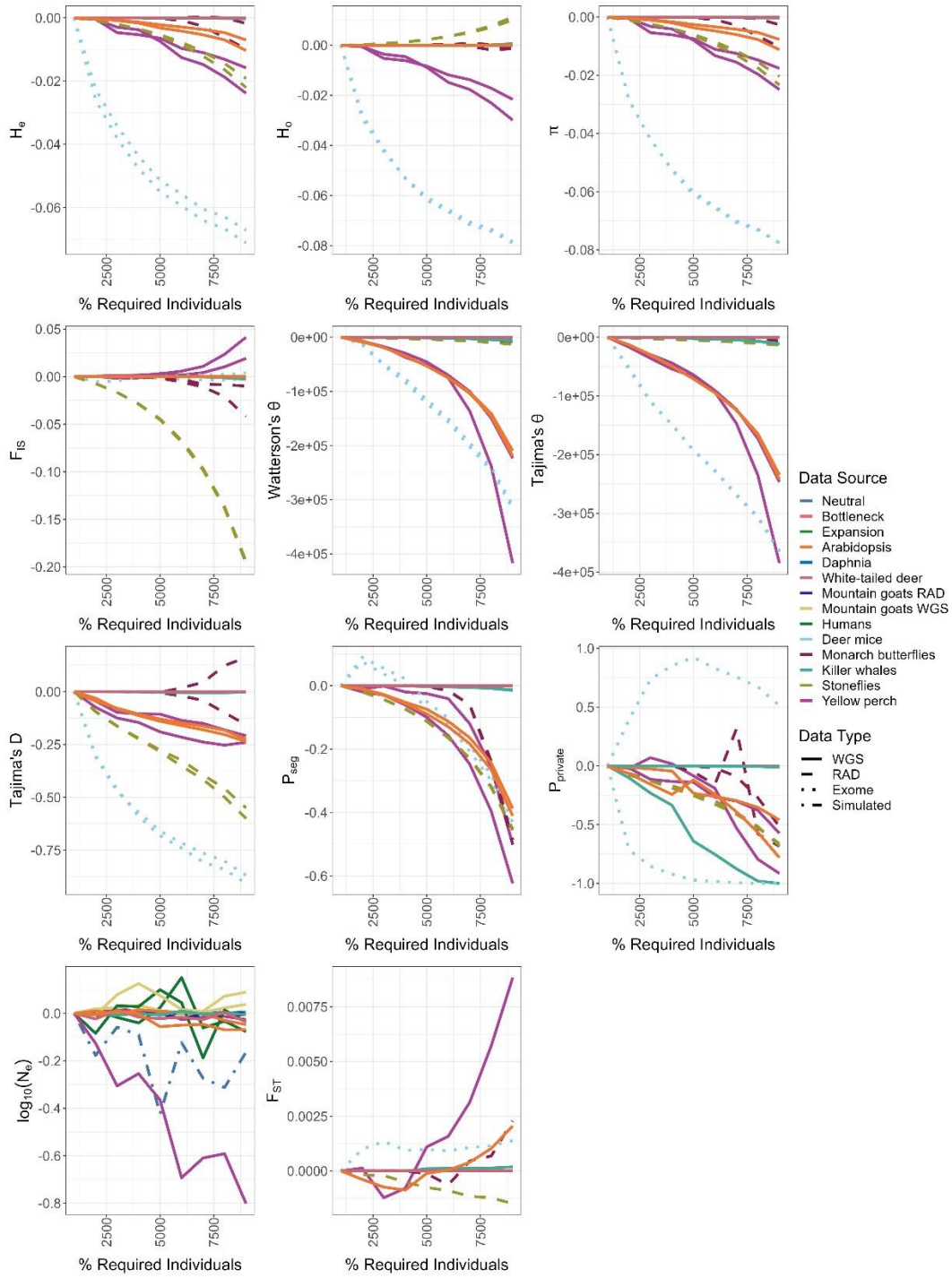


Figure S4: Effects of different thresholds for **filtering out loci based on missing data** on F_{ST}, H_O, F_{IS}, Tajima's D, H_e, π, Watterson's θ, Tajima's θ, the proportion of the total (highest) number of segregating sites, the proportion of the total (highest) number of private alleles, and N_e, observed in each population and dataset depending on missing data filtering stringency. Loci were removed if genotyped in too few individuals. Parameter values have been normalised to show change by subtracting off the value observed with no filter. Note that the deer mouse exome sequencing data in particular shows very large changes across most statistics with stricter missing data filters, including a very large decrease in Tajima's D.

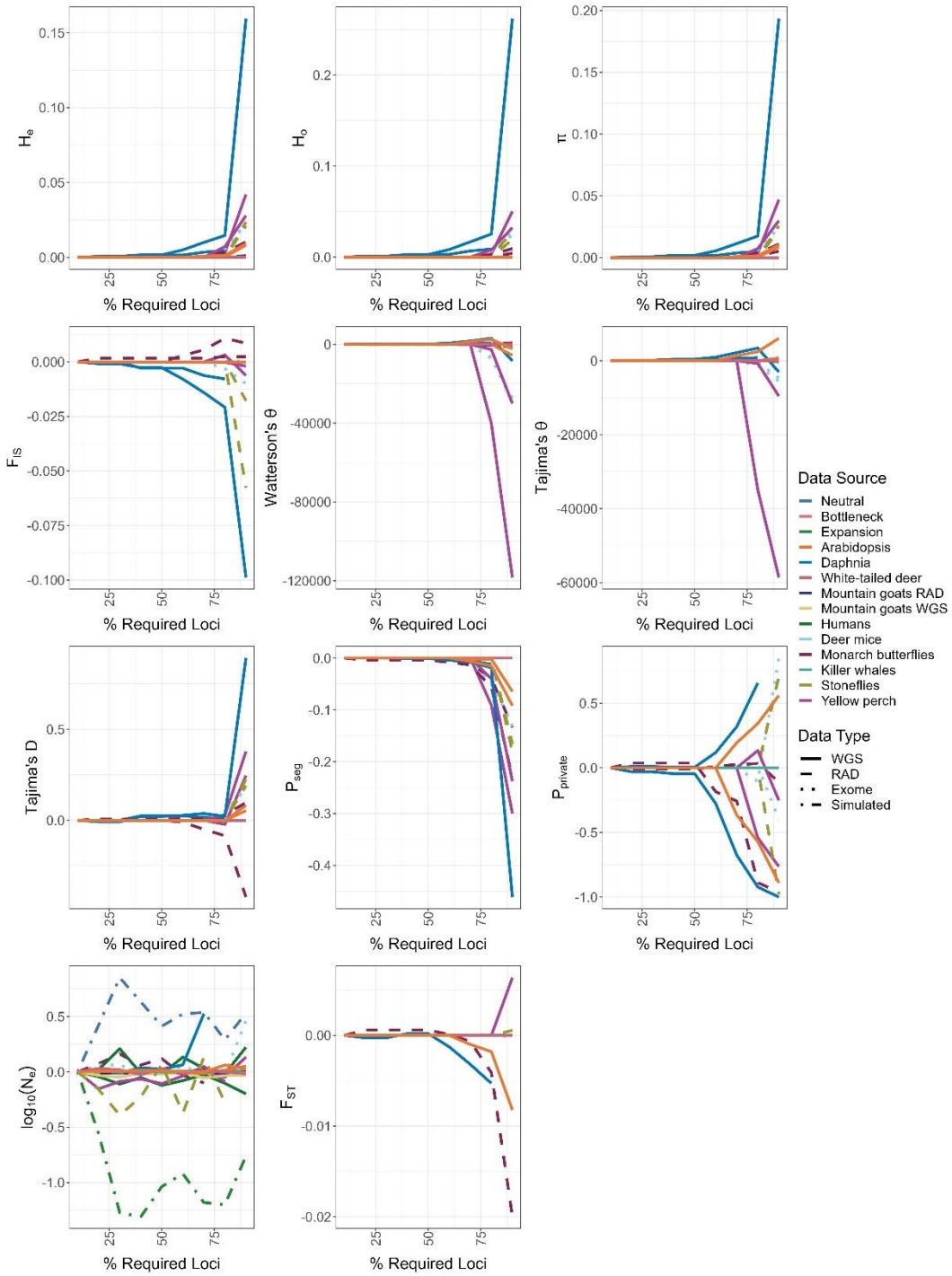


Figure S5: Effects of different thresholds for **filtering out individuals based on missing data** on F_{ST}, H_O, F_{IS}, Tajima's D, H_e, π, Watterson's θ, Tajima's θ, the proportion of the total (highest) number of segregating sites, the proportion of the total (highest) number of private alleles, and N_e, observed in each population and dataset depending on missing data filtering stringency. Individuals were removed if genotyped at too few loci. Parameter values have been normalised to show change by subtracting off the value observed with no filter. Note that the *Daphnia* and yellow perch data show the largest change for most statistics, including substantial changes in Tajima's D, F_{IS}, and H_o/H_e.

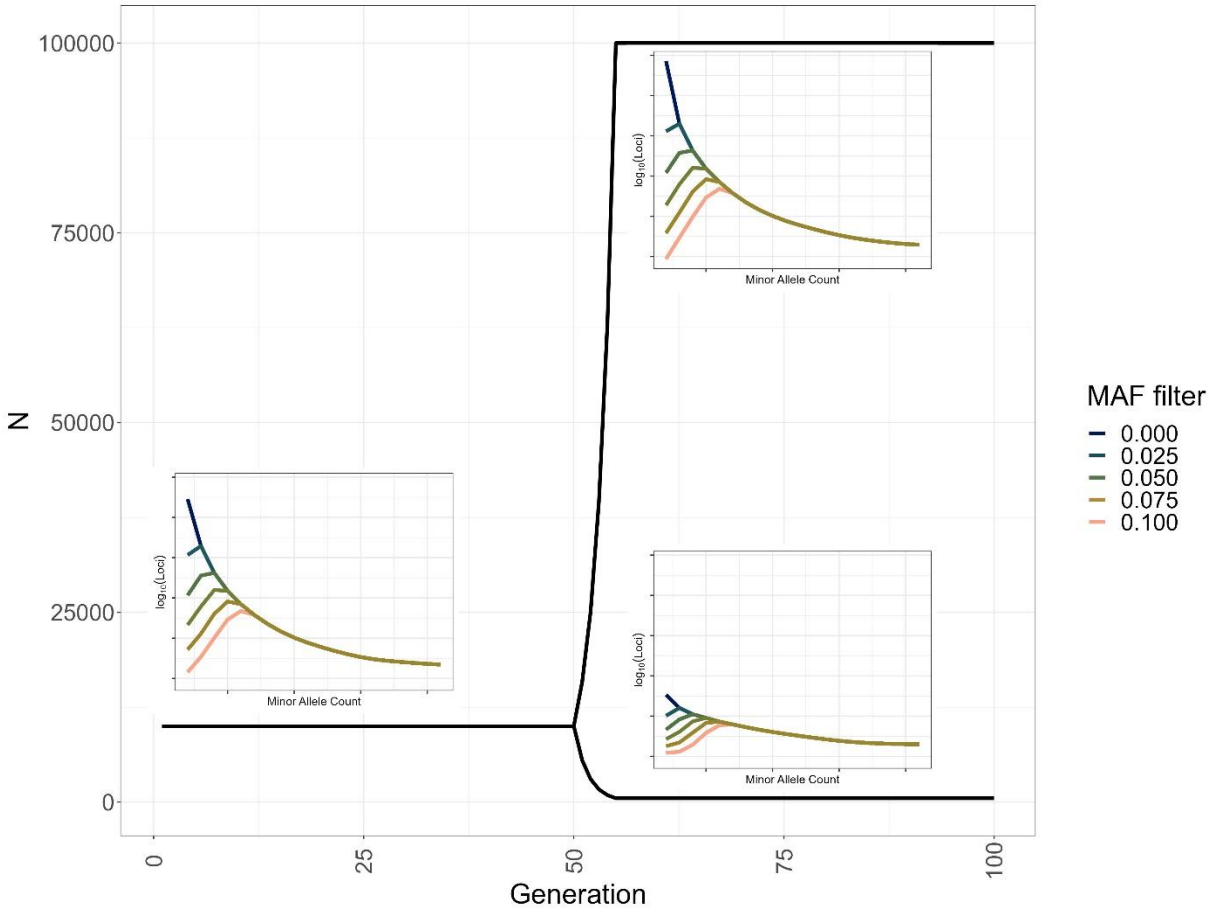


Figure S6: Effect of **MAF filtering** on the site frequency spectrum depending on demographic history. Primary plot displays demographic history, insets show the effect of different MAF filters on the site frequency spectra of neutral, historically expanded, and historically bottlenecked populations. Filtering for MAF impacts "flat" site frequency spectra, like those that have undergone bottlenecks, much less strongly than those with many rare alleles, like in those that have undergone expansions. Note that the y-axes in each spectra plot are scaled to the same minimum and maximum values. Full demographic model parameters are available in Supplementary Table 3.

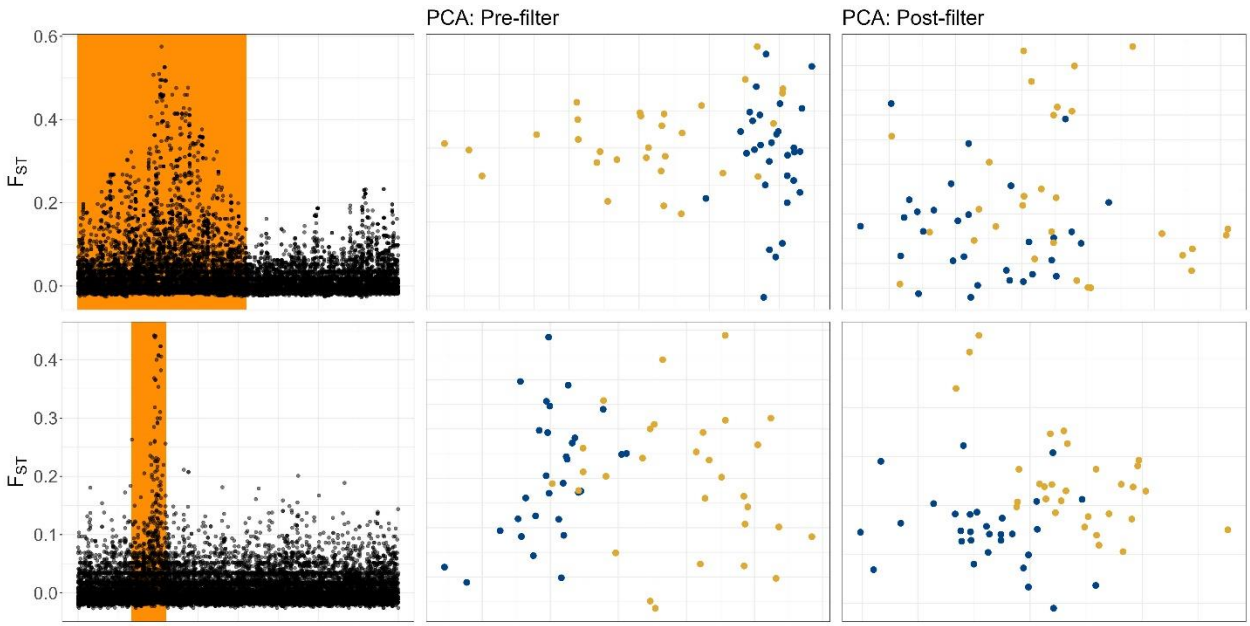


Figure S7: Effect of **filtering out high F_{ST} regions** on subsequent principal component analyses depending on local linkage patterns. Selection or other factors which generate localized sections of elevated F_{ST} can have a strong effect on population structure estimates, particularly in areas with lower recombination rates and/or higher rates of linkage. Blocks of elevated F_{ST} (left, marked in orange) were generated via simulated selection according to the parameters noted in Supplementary Table 3 using either a recombination rate of 0.1 (top) or 1 (bottom).