**TITLE**

A practical guide to filtering in the era of large genomic data sets

**RATIONALE AND CONTENT/ABSTRACT**

Recent and rapid advances in both short (e.g., *Illumina*) and long (e.g., *PacBio*) read sequencing have resulted in the proliferation of genome-wide data. These sequencing methods all result in large numbers of reads that, importantly, are not entirely free of sequencing errors (Pompanon et al., 2005; Stolar & Nekrutenko 2021). These sequences are typically aligned back to a reference genome (reference guided assembly) or to each other (de-novo assembly), both of which can further increase error rates when genetic variants are called (e.g., SNPs, see Fountain et al., 2016; O’Leary et al., 2018). Most researchers, therefore, perform some or multiple types of filtering on minor allele frequencies or counts, missing data, deviations from Hardy-Weinberg Proportions (HWP), and other factors. This filtering may seem trivial, but is in reality of paramount importance because: 1.) all genomic data must proceed through some form of filtering, 2.) filtering choices can be confusing, subjective, and lack standard guidelines (Nazareno & Knowles, 2021), and 3.) the same data set filtered in different ways can produce entirely different results in all downstream analyses (Larson et al., 2021). We provide here a badly needed comprehensive review of different filtering strategies and their downstream effects.

**ARTICLE STRUCTURE and CONTENT**

In this review, we will first summarize the different types of filtering approaches and the rationale behind these filtering strategies. We will also provide an extensive literature review of the most-commonly used filtering approaches and filtering values used across various subdisciplines. We will next highlight the consequences and trade-offs associated with various filtering decisions. For example, using a minor allele frequency filter can minimize the occurrence of including incorrectly genotyped SNPs, improve inferences of population structuring (Linck & Battey, 2019), and reduce the number of false positives during association testing (Ahrens et al., 2021), but can drastically bias estimates of genetic diversity (e.g., or observed heterozygosity, see Cubry et al., 2017, for example) and sequence-based estimators (e.g., Tajima’s D). We will recommend that different sets of analyses require the same underlying data set to be filtered in different ways and conclude with a comprehensive flow chart and series of recommendations to standardize the filtering used across genomic data sets. Our review content will include a table listing key studies and research questions, their filtering decisions, and effects of these decisions on downstream results. We will list pipelines or software that facilitate re-filtering to quantify effects of filtering on genotyping error rates on downstream statistics or estimators.

**SECTIONS**

***What is filtering and why is it important?***

We will provide a conceptual overview of the different types of filtering and why they are commonly applied. In this section, we will use figures to clearly explain what filtering is, why it is necessary with modern sequencing technologies, and the different filtering choices that can be made with a large genomics data set. Examples of the filtering choices we will discuss include:

* Minor Allele Frequency Filtering (MAF): filtering out loci where the less common allele (i.e., the minor allele) occurs below a certain frequency are removed from the data set. This type of filtering is commonly performed based on the assumption that alleles that occur in less than ~5% of the population (frequency < 0.05) must be genotyping errors. Notably, Minor Allele Count (MAC) filtering is an alternative to this approach wherein loci are removed based on the counts of their minor alleles rather than their frequency, allowing for more consistent filtering across samples of different sizes.
* Missing Data: filtering out loci or individuals with a high amount of missing data, which can indicate that something went awry during genomic library preparation or alignment. Missing data can also obscure patterns between samples or populations. However, similar to MAF filtering, choices used vary widely among studies and the downstream consequences are rarely considered.
* HWP: Filtering out loci based on deviations from Hardy-Weinberg Proportions (HWP). HWP is a common assumption of many downstream analytical tools, so removing loci that violate it can help ensure unbiased results. Furthermore, HWP deviations can often reflect sequencing or assembly errors (such as allelic dropout which can cause a heterozygote deficit or paralogous regions that can cause heterozygote excesses, see (Gautier et al., 2013; Günther & Nettelblad, 2019). However, removing loci out of HWP inappropriately can remove loci that, for example, are indicative of selection or population sub-structuring.
* Linkage disequilibrium (LD): pruning clusters of loci that are in substantial LD with each other down to a single locus to ensure independence among loci; an assumption made by many tools. For example, methods based on site frequency spectra (SFS) may be biased if genetic diversity is higher than average in an area that also has many SNPs with substantial pairwise LD. However, diversity estimates themselves (such as the number of segregating sites across genomic regions) may be strongly biased by LD pruning. We’ll also discuss effects of LD filtering on datasets of unmapped loci (Waples et al. 2022) as many non-model species will lack reference genomes long into the future.

A review of some of these filtering issues is available from O’Leary et al., 2018, although they do not demonstrate the effects of filtering choices on down-stream inferences. In addition, while exciting and high-profile, studies with ancient DNA, environmental DNA, non-invasive sampling, and other approaches which tend to yield low quantity or quality DNA, are especially susceptible to filtering issues, which we will discuss.

***Effects of filtering***

In this section we will explain the effects of filtering choices on downstream analyses. For example, genome wide estimates of Tajima’s D can be very positive with a standard MAF of 0.05, yet very negative with no MAF (see figure below). Note that with a positive genome wide Tajima’s D, researchers would conclude that there has been a recent, large reduction in population sizes. When a MAF of 0.02, is used for the same data set, the genome-wide estimates of Tajima’s D are 0, which would suggest no notable demographic changes. Lastly, when no MAF filter is applied to the data set, Tajima’s D is negative and would lead researchers to conclude that there has been a population expansion within the last 100-1000 generations. Therefore, depending on the filtering choices made, researchers can reach two entirely different and opposing conclusions that could have large management and conservation implications. In this example, researchers should not apply a MAF filter when estimating Tajima’s D (low-frequency sites are central to the estimator), such that the correct biological interpretation is that the population has recently expanded. Alternatively, researchers might apply only weak (low-stringency) filters when estimating Tajima’s D to quantify effects of removing only singletons for example, which are sometimes/often sequencing or genotyping errors. The effects of other filtering choices on other genetic metrics can also be significant, as shown for example, by Shafer et al. (2017) for several diversity estimators and demographic inferences and by Díaz-Arce & Rodríguez-Ezpeleta (2019) for FST estimates and allele frequency spectra.

Background pattern

Description automatically generated with low confidence

**Figure** illustrating the changes in Tajima’s D from North American Monarch Butterflies with three different levels of minor allele frequency filters applied (0, 0.02, and 0.05, resulting in 368,672, 119,539, and 69,339 SNPs, respectively (data from Hemstrom et al., 2022)).

Throughout the review, we will illustrate the effects of filtering MAF, HWP, missing data, and LD on summary statistics and estimators like below:

* 1. Nucleotide diversity (π); / Observed heterozygosity (), observed heterozygosity ()
  2. Tajima’s D and Site Frequency Spectra (SFS)
  3. Additional sequence-based estimators
     1. dxy, McDonald Kreitman, dn/ds
  4. Effective population size ( (contemporary local estimates via LDNe; Long-term global estimates via coalescent approaches)
  5. Runs of homozygosity (ROH) and individual inbreeding coefficients (FH)
  6. *FST* (population-wide and outlier detection)
  7. *FIS*
  8. LD
  9. Common analyses and software:
     1. STRUCTURE/PCA
     2. Population assignment tests and gene flow estimates (Nm)
     3. GEA and GWAS
     4. Phylogenies and dendrograms
     5. Admixture and hybridization
     6. Parentage and relatedness

To illustrate and quantify the effects of filtering on these parameters, we will use multiple pre-existing published datasets (whole-genome sequencing and RAD-seq) and ground-truthed simulated data.

***Filtering across populations versus within populations***

Sometimes researchers apply filters across all individuals in their data set at the same time (regardless of population origin) and sometimes researchers apply filters within every population separately. For example, Pearman et al. (2022) reviewed 219 publications that calculated and filtered on HWP using RAD-seq data and found a wide range of population pooling approaches (with approximately ~20% of studies likely using a pooled filtering approach). This filtering choice can have large effects.

* 1. Conceptual issues (with many populations a very common—even fixed—allele in one population could be removed due to being absent everywhere else, for example).
  2. Estimates of genetic diversity (e.g., number of SNPs/kb; total number of SNPs) are greatly affected by within vs across population filtering.
  3. Per SNP *FST* analyses (e.g., outlier analysis)
  4. Removing loci out of HWP in a pooled sample of multiple populations will remove loci that are out of HWP due to the Wahlund effect (De Meeûs, 2018). This is problematic since these loci are also particularly useful for population differentiation and assignment (Waples, 2015).
  5. Whole genome resequencing vs. reduced representation (RAD-seq).

The importance of population-specific filtering schemes has been recently demonstrated for HWP filtering (Pearman et al., 2022), but not extensively for other filtering choices.

***Trade-offs***

Different filtering choices result in different trade-offs, many of which can be couched in terms of type I and type II errors (i.e., erroneous acceptance and erroneous removal errors). At a single locus, a type I error occurs when an incorrectly called genotype and/or sequencing errors are accepted for use (Halvorsen et al., 2023).This error occurs most frequently when no MAF filtering is performed and/or when the read depth at a locus is low and many incorrectly genotyped loci are thus allowed to proceed into downstream analyses. By contrast, type II error occurs when a correctly called genotype is incorrectly filtered out. This error will be more likely when a high MAF filter is used, since low MAF sites are assumed to be errors (and thus removed) even though many of those sites represent real, correctly called genotypes. In this section we will illustrate the trade-offs with varying filtering choices and the consequences on inferences that are likely to occur. Furthermore, we will illustrate that for any given study, it is likely that the same underlying empirical data set will have to be filtered in different ways to use different tools.

***Solutions and best practices***

In this section, we will present a clear path forward considering the challenges and trade-offs listed above. We will: 1.) Recommend that the same empirical data set be filtered in different ways (resulting in several filtered data sets) depending on the questions being asked and the analyses being performed, 2.) Review and recommend a standardized nomenclature based on terms commonly used in the literature to describe the various stages or types of filtered data sets (such as “raw” for an unfiltered dataset, “permissive” for lightly filtered data such as the resulting from the removal of singletons alone, and “strict” for stringently filtered data at a MAF of 0.01, 0.05, etc., see Hendricks et al., 2018), and 3.) Include a detailed flow chart to guide the specific types of filtering to be performed to answer specific questions. For example, no MAF filter (and thus a “raw” dataset) should be applied for estimating Tajima’s D or for calculating the number of SNPs in a population whereas a high MAF filter (and thus a “strict” dataset) should be applied for the estimation of population substructure, population assignment, individual inbreeding (F), or parentage analysis, and relatedness (Díaz-Arce & Rodríguez-Ezpeleta, 2019; Linck & Battey, 2019).

**POTENTIAL DISPLAY ITEMS:**

**FIGURES**

**Figure 1**. Conceptual schematic illustrating what filtering is and why it is fundamental to all genomic data sets (e.g., reads aligned back to genomes, how filtering removes both real and artefactual genetic variants).

**Figure 2**. Summary of frequency of various types of filters used through time (quantitatively estimated from the literature); illustrating in particular that minor allele frequency filtering is now (and relatively recently) ubiquitous, but other forms of filtering (HWP) have been used for longer. Will be included in Box 1.

**Figure 3**. Illustration of how different filtering choices can substantially bias results (e.g., Tajima’s D, genetic diversity, effective population size, population structure).

**Figure 4**. Illustration of the trade-offs associated with various filtering decisions (e.g., minor allele frequency filtering removes spurious genotypes (or alleles) but removes real low-frequency genotypes; essentially type I versus type II error using both simulated and real data sets). May also illustrate the sensitivity of various analyses to filtering choices (some robust (e.g., FST), while others are less or not robust (e.g., SFS, LD, assignment tests)).

**Figure 5**. Flow chart illustrating what types of filtering should be performed for certain analyses and introducing a standardized nomenclature to easily reference differently filtered data sets.

**Figure 6**. Forward looking vision of filtering. Reduced sequencing errors and longer reads will require fewer total reads (probably 1 or more decades away); better error imputation methods, genomes etc. will (perhaps) reduce the need for some types of filtering (MAF). Estimators will account for genotyping errors (e.g. by accounting for typical genotyping errors resulting from low coverage sequencing in a Bayesian genotype-likelihood framework (Korneliussen et al., 2014; Skotte et al., 2013)). Ideally, future analytical pipelines will either automatically conduct weak and strong filtering and quantify effects on downstream statistics or make doing so simple and efficient (as is planned, for example, for the next major release of the snpR analytical toolkit, see Hemstrom & Jones, 2022). Will be included in Box 3.

**TABLES**

**Table 1**: Different filtering methods – main advantages, limitations, and possible downstream effects; standardized nomenclature introduced to complement Figure 4.

**Table 2**: Different studies and research questions, filtering decisions (choices), effects of decisions on results and conclusions.

**BOXES**

**Box 1 | Prevalence of minor allele frequency filtering (MAFs)**

We will summarize papers and the MAF(s) used, how many (and %) of papers use only a single MAF (vs. two or more) MAFs, and if any effect of MAF choice was reported. MAF’s in the literature range from 0.1 to 0.005 and below

**Box 2 | Trade-offs associated with filtering**

We will explain the different errors that can occur based on the different types of filtering choices that can be made. We will include a conceptual figure illustrating type I versus type II errors:

* Type I = call a genotype as correct when it is, in reality, incorrect.
* Type II = call a genotype as incorrect when it is, in reality, correct.

For most genomic data sets the power is high (1- type II error) because the majority of genotypes are correctly called. Nevertheless, even small amounts of type I and type II errors can have large effects (on certain estimators) and filtering choices directly affect the error rates. Many researchers may think that being conservative with their filtering choices is the best or safest strategy (e.g., removing “spurious” loci), however, this strategy also often removes real loci (increases the type II error rate), and the downstream consequences can result in incorrect conclusions depending on the analyses being performed. We will further discuss the interactions between error, power, and filtering.

Diagram

Description automatically generated with low confidence

**Box 3 | Low Coverage and Low Quality**

We will discuss the filtering approaches when using particularly low coverage or low quality sequencing data, where the risks of Type I errors during genotyping are highest. We will review suggestions and demonstrate the results of filtering approaches in such instances. We will highlight the possible filtering considerations when using environmental or ancient DNA (eDNA and aDNA, respectively) using a recent, high profile study of ancient eDNA analyzed by Vernot et al. (2021), which is a particularly apt example because it combines these two particularly error prone approaches.

Neanderthals

**References Cited**

Ahrens, C. W., Jordan, R., Bragg, J., Harrison, P. A., Hopley, T., Bothwell, H., Murray, K., Steane, D. A., Whale, J. W., Byrne, M., Andrew, R., & Rymer, P. D. (2021). Regarding the F-word: The effects of data filtering on inferred genotype-environment associations. *Molecular Ecology Resources*, *21*(5), 1460–1474. https://doi.org/10.1111/1755-0998.13351

Cubry, P., Vigouroux, Y., & François, O. (2017). The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences. *Frontiers in Genetics*, *8*. https://www.frontiersin.org/articles/10.3389/fgene.2017.00139

De Meeûs, T. (2018). Revisiting FIS, FST, Wahlund Effects, and Null Alleles. *Journal of Heredity*, *109*(4), 446–456. https://doi.org/10.1093/jhered/esx106

Díaz-Arce, N., & Rodríguez-Ezpeleta, N. (2019). Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better? *Frontiers in Genetics*, *10*. https://www.frontiersin.org/articles/10.3389/fgene.2019.00533

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., & Peery, M. Z. (2016). Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Molecular Ecology Resources*, *16*(4), 966–978. https://doi.org/10.1111/1755-0998.12519

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M., & Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*(11), 3165–3178. https://doi.org/10.1111/mec.12089

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, *15*(7), e1008302. https://doi.org/10.1371/journal.pgen.1008302

Halvorsen, S., Korslund, L., Mattingsdal, M., & Slettan, A. (2023). Estimating number of European eel (Anguilla anguilla) individuals using environmental DNA and haplotype count in small rivers. *Ecology and Evolution*, *13*(2), e9785. https://doi.org/10.1002/ece3.9785

Hemstrom, W., & Jones, M. (2022). SnpR: User friendly population genomics for SNP data sets with categorical metadata. *Molecular Ecology Resources*, *n/a*(n/a). https://doi.org/10.1111/1755-0998.13721

Hendricks, S., Anderson, E. C., Antao, T., Bernatchez, L., Forester, B. R., Garner, B., Hand, B. K., Hohenlohe, P. A., Kardos, M., Koop, B., Sethuraman, A., Waples, R. S., & Luikart, G. (2018). Recent advances in conservation and population genomics data analysis. *Evolutionary Applications*, *11*(8), 1197–1211. https://doi.org/10.1111/eva.12659

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Larson, W. A., Isermann, D. A., & Feiner, Z. S. (2021). Incomplete bioinformatic filtering and inadequate age and growth analysis lead to an incorrect inference of harvested-induced changes. *Evolutionary Applications*, *14*(2), 278–289. https://doi.org/10.1111/eva.13122

Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639–647. https://doi.org/10.1111/1755-0998.12995

Nazareno, A. G., & Knowles, L. L. (2021). There Is No ‘Rule of Thumb’: Genomic Filter Settings for a Small Plant Population to Obtain Unbiased Gene Flow Estimates. *Frontiers in Plant Science*, *12*. https://www.frontiersin.org/articles/10.3389/fpls.2021.677009

O’Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren’t the loci you’e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, *27*(16), 3193–3206. https://doi.org/10.1111/mec.14792

Pearman, W. S., Urban, L., & Alexander, A. (2022). Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Molecular Ecology Resources*, *22*(7), 2599–2613. https://doi.org/10.1111/1755-0998.13646

Pompanon, F., Bonin, A., Bellemain, E., & Taberlet, P. (2005). Genotyping errors: Causes, consequences and solutions. *Nature Reviews Genetics*, *6*(11), 847–859. https://doi.org/10.1038/nrg1707

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917. https://doi.org/10.1111/2041-210X.12700

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics*, *195*(3), 693–702. https://doi.org/10.1534/genetics.113.154138

Vernot, B., Zavala, E. I., Gómez-Olivencia, A., Jacobs, Z., Slon, V., Mafessoni, F., Romagné, F., Pearson, A., Petr, M., Sala, N., Pablos, A., Aranburu, A., de Castro, J. M. B., Carbonell, E., Li, B., Krajcarz, M. T., Krivoshapkin, A. I., Kolobova, K. A., Kozlikin, M. B., … Meyer, M. (2021). Unearthing Neanderthal population history using nuclear and mitochondrial DNA from cave sediments. *Science*, *372*(6542), eabf1667. https://doi.org/10.1126/science.abf1667

Waples, R. S. (2015). Testing for Hardy–Weinberg Proportions: Have We Lost the Plot? *Journal of Heredity*, *106*(1), 1–19. https://doi.org/10.1093/jhered/esu062