Next-generation data filtering in the genomics era

William Hemstrom[1,4†], Jared Grummer[2,4], Gordon Luikart[2], Mark Christie[1,3]

[1]Department of Biological Sciences, Purdue University; 915 W. State St., West Lafayette, IN, USA

[2]Flathead Lake Biological Station, Wildlife Biology Program and Division of Biological Sciences, University of Montana, Missoula, MT, USA

[3]Department of Forestry and Natural Resources, Purdue University; 715 W. State St., West Lafayette, IN, USA

[4]These authors contributed equally to this work: William Hemstrom, Jared Grummer

[†]e-mail: whemstro@purdue.edu

Abstract

Genomic data are now ubiquitous in studies across disciplines including human health, agriculture, biodiversity, molecular ecology, and evolutionary biology. A key step in genomic data analysis is filtering (the intentional removal of bases, reads, individuals and/or genetic variants from a genomic dataset with the explicit goal of improving data quality for downstream analyses). Researchers are confronted with a multitude of choices when filtering genomic data: they must choose which of many filters to apply and select appropriate thresholds at which to apply them. Here, we review commonly applied filter types and thresholds, including minor allele frequency, missing data per individual or per locus, linkage disequilibrium, and Hardy–Weinberg deviations, and illustrate large effects of filtering thresholds on statistics such as Tajima's D, F_{ST}, and effective population size

(N_e). To help usher in the next generation of data filtering, we propose key best-practice principles for investigators to apply when analysing genomic data and conclude with recommendations for standardizing filtering approaches.

[H1] Introduction

Recent and rapid advances in both short- and long-read sequencing technologies (*e.g.*, Illumina and PacBio) have resulted in the proliferation of large genomic datasets[1–4]. All high-throughput ('next-generation') sequencing methods result in large numbers of reads that, importantly, do not all have low error rates[5,6]. Errors can be further inflated when these sequences are aligned to a reference genome or transcriptome, particularly if the reference is incomplete or assembled ***de novo*** from the reads themselves[7–9]. Therefore, most investigators perform multiple types of data ***filtering*** for quality control prior to data analysis, including filtering on read quality, mapping quality, read depth/coverage, minor allele frequency or minor allele counts, missing data (across loci and/or individuals), and deviations from Hardy–Weinberg or linkage equilibrium. Despite the ubiquity of filtering in genomics, there are few established best practices and little standardization for data filtering.

Genomic filtering is not a trivial problem. Indeed, it has been aptly termed the "*F*-word"[10,11] by geneticists due to how challenging it can be to conduct and understand effects of filtering on downstream conclusions. Here, we define ***filtering*** as the intentional removal of bases, reads, individuals, and/or ***genetic variants*** from a genomic dataset with the explicit goal of improving data quality for downstream analyses. Filtering is distinct from other forms of data processing in that it centres on the removal of data (as

in “filtering out”), whereas other approaches, such as ***imputation***, modify the data directly without removal. Filtering is an issue of paramount importance because: 1) every genomic dataset must be filtered, often repeatedly, 2) the same dataset filtered in different ways often yields entirely different results[12], and 3) filtering choices can be confusing, are often subjective, and currently lack standard or agreed-upon guidelines[13].

Filtering approaches are often inadequately described and vary widely among studies. Often, it is impossible to tell whether certain filtering steps were conducted and not described or if they were skipped entirely. Furthermore, even when filters are mentioned, the methodological details used are infrequently reported. Among papers that *do* record filtering thresholds, the values for used for a given filter can vary several-fold, suggesting limited standardization of filtering efforts. Furthermore, the vast majority of studies provide no justification for the thresholds they use, and researchers often use the default program settings with no further comment. Given the lack of comprehensive sequence data filtering standards or best practices and that the impact of filtering is rarely quantified (and thus is poorly understood), the methodological details provided by many published studies are ultimately not sufficient for reproduction.

To help investigators improve their filtering methodologies, we review different strategies for filtering genomic datasets, examine their downstream effects, and provide practical recommendations including a detailed flowchart illustrating a typical order of operations. We also provide example tables demonstrating how to document filtering steps and a checklist to consult before publication, which should be useful for both investigators and scientific journals.

[H1] Common filters: complexity and importance

Today, investigators have many different instrumentation options for obtaining DNA, RNA, and epigenetic sequencing data[14,15], which subsequently yield reads of different lengths, configurations (*e.g.,* single- or paired-end), and qualities[16]. These sequences are then either aligned back to a ***reference*** (such as a reference genome or transcriptome), or *de novo* assembled and aligned to consensus sequences (*i.e.*, contigs or stacks) for subsequent variant calling or analyses[17–20]. Without some form of reference-guided alignment, downstream analyses can be limited because it can be difficult to estimate relevant parameters (*e.g.*, runs of homozygosity[21]) or determine the genomic context for loci of interest (*e.g.*, linkage)[22] (but see[23] and [24] for best practices in metagenomics and eDNA, respectively). Lacking a quality reference genome therefore represents a challenge for researchers working in non-model systems, where alignments are often made to low-quality reference genome assemblies (*e.g*., high contig counts with low N/L50 scores[25]) or even to different, often distantly related species[26,27]. Given these many challenges and technical limitations, errors are always introduced to some degree during both sequencing and alignment, and these errors must be accommodated – usually via filtering – to avoid downstream biases.

Many types of population genetic or statistical analyses should be conducted only after applying specific filters, and the results from certain analyses can suggest the need for additional or modified filtering strategies. For example, unexpected results from exploratory methods such as a principal component analysis (PCA) can be indicative of experimental or laboratory errors (*e.g.*, mislabelling), sequencing bias, sex-linked loci, selection, or other phenomena[28–30] and thereby suggest the need for further filtering

steps. Thus, the process of filtering begins immediately after sequence data collection and typically may not end until all analyses are completed. An exhaustive list of filters, their descriptions, and the typical stage in the workflow where they are applied can be found in Table 1.

In this section, we introduce and review several of the filters that often have the largest downstream effects. Specifically, we discuss filters for: read and mapping quality, ***read depth***, ***missing data***, ***minor allele frequency (MAF)/minor allele count (MAC)***, ***Hardy–Weinberg proportions***, and ***linkage disequilibrium***. These filters can be generally divided into two sequential categories: pre-variant filtering (*e.g.*, read/mapping quality and read depth) and post-variant filtering (*e.g.*, MAF/MAC, missing data; see Table 1). Pre- and post-variant stages generally coincide with pre- and post-VCF-file stages (or genotyping).

[H2] Pre-variant filtering

Prior to ***de novo assembly*** or alignment to a ***reference***, data are usually filtered via the removal of low-quality reads (see Table 1). This initial filtering may have downstream effects, and researchers across disciplines often use different and potentially arbitrary definitions of what constitutes a low-quality read. For example, a wide range of read quality thresholds are used, ranging from a Phred scaled quality score of 5[31,32] to 40[33], or between 31% and 99.99% calling confidence. Low-quality reads can also be removed during the process of alignment itself, since different alignment algorithms can prevent sequences from mapping back to a reference if their quality scores are below user-defined thresholds[34–36].

Conversely, high-quality reads may also be excluded if the reference does not contain the sequence or if their sequence is highly repetitive (*e.g.*, transposable elements) or found in multiple genomic locations (*e.g.*, paralogs; see Figure 1A). To aid reviewers and other researchers in understanding the impacts of these filtering decisions, investigators should always report: 1) the percentage of reads that were removed prior to alignment and 2) the percentage of reads that mapped successfully and uniquely (to one location) on the reference. Researchers should also report the methodology used to remove low-quality reads (e.g., a hard filter like read length or soft filter based on a statistical model[36]).

Reporting these statistics can help reviewers assess the quality of the data underlying the study and allow future investigators to determine whether alternative filtering could address additional questions (*e.g.*, re-filtering to not remove reads that map to multiple locations could identify paralogous loci or transposable elements). Note, however, that even a ***strong alignment*** of putatively high-quality reads does not always mean a *correct alignment,* since reference bias[37], genome assembly errors[38], ***structural variants***[39], and challenging alignments (*e.g.*, transposable elements[40], PCR duplicates[41], and copy-number variants[42]) are present in most genomic datasets (Figure 1).

[H2] Post-variant filtering

After pre-variant filtering, genomic work-flows typically proceed with **genotyping**, whereby genetic variants such as **single nucleotide polymorphisms** (SNPs), insertions or deletions (indels), and structural variants are algorithmically identified with software such as GATK[43], ANGSD[44], STACKS[9], ipyRAD[45], or LUMPY[46]. During this process, the

read depth (or coverage) of each locus must be considered, since high depth of coverage allows for more confidence in genotyping (and subsequently downstream inferences). These algorithms typically either mark genotypes as missing if below a certain read depth or if they have poor quality scores caused by low read depths[47]. The depth and/or quality filters used at this step vary substantially across studies. Note, however, that well-developed approaches to make use of low coverage sites do exist[44], and so filtering out such loci is not *always* necessary.

While many different ways to filter genotypic data exist (Table 1), we focus here on four widely used filters (Figure 2): missing data, minor allele frequency/count (MAF/MAC), Hardy-Weinberg proportions, and linkage disequilibrium. Below, in the "Solutions and Best Practices" section, we suggest how to best implement these filters.

[H3] Missing data

Loci or individuals with more than a user-defined amount (or proportion) of missing data are often filtered out. An excess of missing data can indicate something awry with sample collection or preservation, genomic library preparation, or alignment, all of which can obscure patterns of variation[48]. The filtering choices used for missing data vary *widely* among studies and the downstream consequences are rarely evaluated (Box 1).

[H3] Minor allele frequency (MAF)

Loci (typically SNPs) for which the less frequent allele (*i.e.*, the minor allele) occurs below a certain frequency are also often filtered out. MAF filtering is often based on the assumption that **singletons** or **other rare variants** that occur at a frequency of less than

~5% are the product of genotyping errors. MAF filtering is therefore typically done to this specific threshold (MAF = 0.05), although threshold values across published studies can vary by orders of magnitude (e.g. from 0.001 to 0.10). Depending on the analysis and objectives, this filter can be applied study-wide (*e.g.,* globally across populations) or within each population.

[H3] Minor allele count (MAC)

MAC filtering is an alternative to MAF filtering wherein loci are removed based on the absolute count of the minor allele rather than its frequency, allowing for more consistent filtering across samples of different sizes (although arguably producing an uneven MAF filter across those same samples). A MAC may be preferable over a MAF to remove only singletons (or doubletons, *etc.*) or when small samples of individuals exist while filtering within populations or sample groups.

[H3] Hardy–Weinberg proportions (HWP)

It is often desirable to filter out loci based on statistically significant (for a given alpha/*p*-value) deviations from Hardy–Weinberg proportions. HWP is a common assumption of many downstream analytical tools (*e.g.,* STRUCTURE[49]), and removing loci that violate HWP can help ensure unbiased results for downstream analyses[50]. Correcting *p*-values for multiple tests in datasets with many loci is rarely performed[51]. Researchers who do correct for multiple tests should explicitly report the reasoning behind the correction method they use, which can vary in stringency (from Benjamini–Hochberg[52] to sequential or simple/stringent Bonferroni[53]). Note, however, that multiple testing correction for HWP

182 can paradoxically *decrease* filtering stringency overall, since more loci will be found to be
183 non-significant and thus retained.

184 Ultimately, different approaches and alpha thresholds should be applied
185 depending on the questions being asked and the tolerance for including problematic loci[54]
186 (Box 2). HWP deviations can often reflect sequencing, assembly, or alignment errors
187 (such as a heterozygote deficit caused by allelic dropout or a heterozygote excess caused
188 by paralogous regions; see[55–57]). However, loci out of HWP can be indicative of real,
189 biological phenomena, such as cryptic population sub-structuring (Figure 2), or balancing
190 selection. As a result, it is crucial to filter for HWP within sample groups (*e.g.*, populations)
191 rather than study-wide (*e.g.*, globally on all samples; see *Study-wide versus within*
192 *sample-group filtering*, below)[58] and to do so with a low stringency if loci under selection
193 or those that differ between populations are of interest.

194

195 *[H3] Linkage disequilibrium (LD)*

196 Pruning clusters of loci that are in substantial LD with each other down to a single locus
197 ensures statistical independence among loci — a common assumption made by many
198 downstream methods. For example, methods based on the site frequency spectrum
199 (SFS) of a population may be biased if allele frequencies in a variant-rich region differ
200 from the genome-wide average. Similarly, failure to remove non-independent (linked) loci
201 can bias estimates of parameters like effective population size (N_e)[59]. However, filtering
202 out SNPs based on LD could also strongly influence diversity estimates (such as the
203 number of segregating sites across genomic regions) or inadvertently cause investigators
204 to overlook important structural variants (Figure 2). Studies that lack a high-quality

reference may also require LD filters to ensure independence of loci or contigs[60], which can be accomplished through pairwise correlation measures such as Pearson's r^2. Alternatively, many investigators working with *de novo*-assembled datasets simply extract a single SNP from each contig to mitigate the effects of linkage (though this assumes distinct stacks/contigs are themselves unlinked). Corrections for multiple testing are also important for LD filtering, but seldom reported[51].

[H1] Effects of filtering

The effects of filtering are often unappreciated and unknown in genomic studies. While concerning, this is not particularly surprising, given that many different filtering approaches exist; filtering requires non-trivial time and computational resources to perform; and many individual filters can be applied with different thresholds and at different data processing stages (Table 1). Indeed, circumstances exist where the same type of filter can, and possibly should, be applied at multiple stages of the bioinformatic process (Table 1). Furthermore, many types of filtering occur during the 'black box' of alignment and genotyping, leading many investigators to use default settings and not think about the downstream consequences. This strategy occurs because the added complexity of filtering can be overwhelming, time-consuming to properly address, and seemingly distract from the main goals of the study. However, properly considering filtering choices and their impacts is nonetheless *critical*, because different filtering choices can lead to completely different downstream results such that two researchers

who made different decisions but analysed the same data, could, as we will show here, reach entirely different biological conclusions.

To illustrate this, we re-filtered three pre-existing empirical datasets multiple times by changing filtering thresholds for two key filters (MAF and missing data) in Box 1. While MAF filters are often applied to remove **singletons or other rare variants** (e.g., MAF < 0.05), which can reflect sequencing or genotyping errors, these rare variants are critical in several analyses including demographic estimation and tests for selection. Most notably, ***Tajima's D***, a commonly used indicator of both demographic history and response to selection[61], is *substantially* biased by a MAF filter choice, leading to widely differing biological inferences depending on filtering stringency (Box 1). In this case, our recommendations are straightforward: because low-frequency alleles heavily influence Tajima's D[62], researchers should apply both *no* MAF filter and a *very minor* one (such as a singleton filter) and compare the results when using the statistic[62]. The effects of additional filters (including MAF) can be substantial for diversity estimators, demographic inference[63,64], F_{ST} and population structure estimates[65,66], estimating the distribution of locus effects on phenotypes[66], and allele frequency spectra[67–70]. Other filtering choices therefore require similar levels of care.

[H2] Study-wide versus within sample-group filtering:

Many filtering methods can be applied to all individuals in the study or separately within each sample group, which can represent different populations, geographic or temporal

sampling units, or experimental treatments. When filtering occurs across all samples (*e.g.*, all individuals) within a study jointly and simultaneously, we refer to this process as ***study-wide*** (or "global") ***filtering***. When filtering occurs within each sample group separately, we refer to this process as ***within-group filtering***. The effects of this filtering decision can be surprisingly large. For example, when a study-wide MAF filter of 0.01 was applied globally to a yellow perch (*Perca flavescens*) ***whole-genome sequencing (WGS)*** dataset, each sample group was constrained to 714,000 SNPS (Figure 2). However, when the same 0.01 MAF filter was applied separately within each sample group, the number of SNPs per sample group varied by a factor of 3.3[71]. In this case, some populations in the study had radically different SFS, likely caused by recent population expansions in some but not all populations[61,62]. Study-wide filtering therefore led in this case to the removal of critically informative, globally rare but locally common alleles. While filtering MAF globally instead of within study groups is common, it should generally be expected to have serious effects whenever SFS vary between study groups, such as when demographic histories differ (like above) or when local adaptation has occurred.

Study-wide versus within-group filtering will also impact ***genome-wide association studies (GWAS)***, where it is common to perform study-wide MAF filtering where the MAF threshold is dictated by sample size (which can often be quite large, particularly in human or agricultural work[72]). The implications of these standardized pipelines are often not given much consideration, but the effects may be non-trivial. When populations with different SFS are compared, for example, a study-wide MAF filter can introduce ascertainment bias by removing more segregating loci from specific study groups. Human populations (and those of other species with complex biogeographic

histories) may be prone to this bias, since populations with African ancestry tend to have more sites with low-frequency alleles than those with European ancestry[73]. Using a study-wide MAF filter will therefore remove more segregating loci from the African ancestry sample group and could result in the preferential detection of large-effect loci in European populations. While we have focused on MAF filtering here due to its near universal implementation, other filtering approaches can be similarly biased by study-wide versus within-group filtering. Differences in downstream outcomes from filtering HWP[58] and LD[74] within-groups vs. study-wide, for example, have been previously documented.

In light of these findings, it is crucial to consider *why* results may differ when applying filters globally or within groups, particularly if sample groups include individuals from different populations, locations, or time points[75]. Tests for HWP should be conducted on each local group or population (deme) separately. If genetically-distinct groups are pooled, there will be an excess of homozygotes (positive F_{IS}) across loci genome-wide (i.e., a ***Wahlund effect***) and their removal can mask population structure (see Figure 2)[50,76]. Genome-wide heterozygote deficiency can also result from low sequencing coverage and allelic dropout, which can be difficult to differentiate from Wahlund effects (although the latter influences all loci independent of read depth[54,68]). If a specific locus shows a consistent HWP deviation (*e.g.,* positive or negative F_{IS}) in multiple different groups, it may indicate genotyping error (*e.g.*, allelic drop-out) or alignment or genome assembly errors (beyond natural/biological processes[77]); whereas consistent, multilocus within-group deviations may instead indicate inbreeding, underlying cryptic population structure, and/or assortative mating.

[H1] Solutions and best practices

Filtering is a powerful tool that should be applied thoughtfully, early, and often throughout genomic dataset construction and analysis to test for the effects of broad-ranging problems (*e.g.*, experimental, sample collection, labelling, or library preparation errors; batch-effect sequencing and genome assembly errors), and to detect interesting biological phenomena (*e.g.*, natural selection, structural variants, Wahlund effects). Even in highly-studied species, such as humans, thoughtful and multi-faceted filtering is important as novel structural and genetic variants occur within every population[78], and failing to account for them may curtail power to identify causal associations or even lead to incorrect inferences[47].

Our core recommendation to researchers follows from this. Because different filtering choices can result in different downstream inferences, we recommend that *distinctively- filtered versions of the same dataset should be used quantify the effects of filtering and to address specific research questions*. Investigators should filter their data in multiple ways, report the results of such filtering on downstream analyses, and think critically to ensure that the filtered datasets used to answer specific questions are appropriate and do not themselves create a significant source of bias. Investigators should become more comfortable using distinctively- filtered datasets to answer different questions even within the same study. For example, researchers should consider using low-stringency MAF filters for several demographic inferences (e.g., Tajima's D; SFS) but a relatively stringent MAF filter for delineating populations[65], planning genomically informed breeding strategies[79], or for estimating individual relatedness[80–82]. This concept

316 requires a fundamental shift in the way genetics data are analysed: investigators must
317 realize that no single “best” filtering strategy or filtered dataset exists for every question,
318 method, or objective.

319 To assist investigators in matching research questions with methods and filters we
320 have created a detailed flow chart that can be applied generally across disciplines and
321 study systems (Figure 3; see also Table 1 and Box 3). We recognize that some alignment-
322 free methods exist that use high throughput sequencing data, particularly for
323 metagenomics and other phylogenomic analyses[83], and they are not considered
324 extensively here. Likewise, filtering for RNA-seq is not thoroughly reviewed here: while
325 many concepts hold true (especially for SNPs called from RNA-seq data), many of the
326 specific filters may not apply (see [84] and [85]). For all other studies, we here highlight the
327 salient features of a genomics study workflow.

328

329

330 *[H2] Best practices for pre-variant calling workflows*

331 First, we note that most genomic workflows differ depending on the research
332 questions and data types. The documentation of filtering decisions is therefore paramount
333 for reproducibility and research efficiency. As a first step, we recommend that raw data
334 be immediately archived in independent, non-local repositories created for genomics data
335 (*e.g.*, the NCBI Short-read Archive, the European Variation Archive, the DNA DataBank
336 of Japan) prior to any analysis (see[86,87] for reviews on genomics data management best
337 practices). Given that filtering, by definition, requires manipulating data, the importance

of archiving raw data cannot be understated. To this point, we refer the reader to[88] for information on dataset and study organization.

After archival, reads should be filtered for general quality control (base quality, adapter removal, poly-G tails, sequencing artefacts, *etc.*) and trimmed when appropriate and useful[89,90]. For most workflows, the alignment of reads to a reference or *de novo* assembly is the next step (Figure 3). Depending on the goals of the study, it may be useful to create multiple datasets with different filters and/or filtering thresholds at this stage for downstream analysis (as in [91,92], for example). This practice is particularly relevant to *de novo* reference assembly, since assembly decisions can result in very different references and thus very different filtering and analytical outcomes. The *m* and *M* STACKS parameters and their impacts on *de novo* reference construction, for example, have been well-studied[19,20,93,94].

After alignment, the data should be filtered for technical (*e.g.*, PCR) duplicates. Note that while removing PCR duplicates has been suggested to often be of little consequence[95,96], this is unlikely true for every study, such as those with low coverage[11,41,97] (in conflict with common practice, see Supplementary Materials). The remaining reads should then be filtered for mapping and read quality, and researchers should ensure that they record and eventually report the number of reads that passed these pre-variant filters.

[H2] Best practices for post-variant calling workflows

Researchers can then call variants, filter the resulting data set to remove potentially problematic loci (for MAF/HWP/LD/*etc*.), and then remove poorly sequenced

individuals (and/or samples with other quality or analytical concerns; see Figure 3). Note that it may be beneficial to reverse the last two steps and filter across individuals first (and loci second) in instances where retaining as many loci as possible is needed or where data quality varies widely among individuals[8]. An iterative approach, where individuals and loci are first removed with low stringency filters and then subjected to additional rounds of filtering stringencies may also improve data quality by removing poor individuals that reduce overall call-rates in high-quality loci and vice versa[8]. As with pre-variant filters, the percentage of reads, sites, and individuals retained at each post-variant filtering step should be reported (Box 3).

We generally recommend at this stage that a *minimum* of two (divergently-filtered) datasets be created — one with low filtering stringency (*e.g.*, allowing more missing data, a low/permissive MAF threshold, and few loci removed due to HWP and LD deviations), and one with high filtering stringency (*e.g.*, many loci/individuals removed due to missing data, and a higher/restrictive MAF threshold, etc.). The precise filters and thresholds used should reflect the specificities of the study: for example, studies interested in transposable elements may want to vary alignment thresholds (uniquely versus multiply mapped reads) but keep other filters stringent to strike a balance between sensitivity and accuracy[98,99]. After the initial filtered datasets are created, investigators should proceed with their parameter estimation, statistical analyses, and modelling with these datasets in parallel to answer their key questions of interest. Note that we are not the first to suggest comparing outcomes from different filtering strategies[13,64,68], and we suspect that this recommendation will become more common over time. Some stand-out papers exist that already follow this recommendation[13,66,92,100], although they are in the minority.

As data analysis proceeds, we suggest that re-filtering should be part of most genomics workflows. For example, PCA may reveal individuals that were mislabelled, or mis-classified into an incorrect sample group or were contaminated during sample preparation and should be removed. Similarly, the addition of new analyses (*e.g.*, Tajima's D, transposable element annotation) that were not initially considered may also require a careful re-filtering of the data. We suggest that authors and journals implement a supplementary table that describes the final datasets, the specific filters and thresholds employed, the name of the final VCF files, and the specific analyses for which each distinctly filtered dataset was used. We provide an example of such in Table 2. Researchers should also explain if they corrected for multiple testing along with a (brief) justification for the correction method used (*e.g.*, Bonferroni, FDR, etc.; see [51]).

After analyses are completed, the resulting data and analytical tools should again be archived and/or recorded, including all relevant meta-data and the exact filtering decisions and genotyping pipelines (*e.g.*, bioinformatic scripts, software versions). Given that recreating the exact filtering and genotyping pipelines requires a considerable amount of resources (and may actually be impossible given limited data or analytical tool availability), we strongly recommend that post-project archival include all filtered genotypic/variant data in the form of carefully annotated ***VCF files*** that include detailed filtering descriptions in the header (see [101] for a detailed description of VCF files).

We recognize that a proper and thorough examination of filtering will necessitate extra time, computational resources, and work from researchers. However, these changes to current workflows are generally necessary to achieve high-quality, reproducible research and a better understanding and quantification of filtering effects.

Following reproducible research guidelines may help; reproducible research is reproducible not just for other researchers, but also for the primary investigators themselves. A reproduction-friendly pipeline script that runs a suite of analyses when given a dataset and a set of filtering parameters is also easy to re-run a second time with a new (re-filtered) dataset. For examples of studies with well-documented methods, easily accessible data, and that would be relatively straightforward to reproduce with new filters and thresholds, see[65,102,103].

[H1] Conclusions and future directions

In this new and exciting era of genomics, a systematic, thoughtful, and transparent approach to filtering sequence data should be an integral part of every publication and data analysis pipeline. Investigators should strive to filter with a focus on reproducibility and aim to match the filters they employ to the questions they intend to answer. While technological advances have exponentially increased the amount sequencing data produced, advances in filtering and its subsequent documentation or reporting have not kept pace. Instead, many investigators are unfamiliar with or ignore filtering issues during the quality control and analysis of large genomic datasets, for example, using default program settings without critical thought or explanation of their filtering decisions.

Here, we reviewed the different types of data filtering, illustrated the effects of divergent filtering choices and thresholds, and presented a flow chart and checklist to simplify, streamline, and potentially standardize and improve the filtering process. Critically, we highlight that for the same dataset 1) different filtering thresholds can create different downstream results for the same dataset, and 2) most analyses should be run

on multiple datasets produced with different filters and thresholds to allow for the quantification of filtering effects on results and improve conclusion certainty. As the technology behind genomic sequencing continues to advance, we will likely see longer, higher-quality reads, which, alongside improvements in reference quality[104–106] and the burgeoning field of pangenomics[107,108], will increase the accuracy and power of genomic data analyses. Nonetheless, no genomic dataset will be error-free, and filtering will undoubtedly remain a central part of all genomic analyses for decades to come. We hope this review ushers a new era of next-generation filtering in genomic analyses that sparks improvements in our understanding and applications of filtering, data interpretation, reproducibility, and drives production of new data analysis tools to make it easier to re-filter and quantify filtering effects on questions across disciplines.

Acknowledgements

We thank Xiaoshen Yin, Claire Schraidt, Micah Freedman, Helen Neville, and Michael Miller, for allowing us to review and re-filter their data. Mark R. Christie was funded, in part, by NSF DEB-1856710 and OCE-1924505.

Competing interests

The authors declare no competing interests.

References

1. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nat. Rev. Genet.* **11**, 697–709 (2010).

2. Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C. & Scorilas, A. Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life* **12**, (2022).

3. Fiedler, P. L. *et al.* Seizing the moment: The opportunity and relevance of the California Conservation Genomics Project to state and federal conservation policy. *J. Hered.* **113**, 589–596 (2022).

4. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: An overview. *Gener. Seq. Its Appl. Med. Lab. Immunol.* **82**, 801–811 (2021).

5. Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.* **6**, 847–859 (2005).

6. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).

7. Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J. & Peery, M. Z. Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Mol. Ecol. Resour.* **16**, 966–978 (2016).

8. O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M. & Portnoy, D. S. These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* **27**, 3193–3206 (2018).

9. Rochette, N. C., Rivera-Colón, A. G. & Catchen, J. M. Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.* **28**, 4737–4754 (2019).

10. Ahrens, C. W. *et al.* Regarding the F-word: The effects of data filtering on inferred genotype-environment associations. *Mol. Ecol. Resour.* **21**, 1460–1474 (2021).

11. Andrews, K. R. & Luikart, G. Recent novel approaches for population genomics data analysis. *Mol. Ecol.* **23**, 1661–1667 (2014).

12. Larson, W. A., Isermann, D. A. & Feiner, Z. S. Incomplete bioinformatic filtering and inadequate age and growth analysis lead to an incorrect inference of harvested-induced changes. *Evol. Appl.* **14**, 278–289 (2021).

13. Nazareno, A. G. & Knowles, L. L. There Is No 'Rule of Thumb': Genomic Filter Settings for a Small Plant Population to Obtain Unbiased Gene Flow Estimates. *Front. Plant Sci.* **12**, (2021).

14. Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* **18**, 9–19 (2020).

15. Kumar, K. R. ; C., Mark J. ;. Davis, Ryan L. Next-Generation Sequencing and Emerging Technologies. *Semin. Thromb. Hemost.* **45**, 661–673 (2019).

16. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).

17. Lou, R. N., Jacobs, A., Wilder, A. P. & Therkildsen, N. O. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol. Ecol.* **30**, 5966–5993 (2021).

18. Olson, N. D. *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* **24**, 464–483 (2023).

19. Rochette, N. C. & Catchen, J. M. Deriving genotypes from RAD-seq short-read data using Stacks. *Nat. Protoc.* **12**, 2640–2659 (2017).

20. Paris, J. R., Stevens, J. R. & Catchen, J. M. Lost in parameter space: a road map for stacks. *Methods Ecol. Evol.* **8**, 1360–1373 (2017).

21. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).

22. Heller, R. *et al.* A reference-free approach to analyse RADseq data using standard next generation sequencing toolkits. *Mol. Ecol. Resour.* **21**, 1085–1097 (2021).

23. Gihawi, A., Cardenas, R., Hurst, R. & Brewer, D. S. Quality Control in Metagenomics Data. in *Metagenomic Data Analysis* (ed. Mitra, S.) 21–54 (Springer US, 2023). doi:10.1007/978-1-0716-3072-3_2.

24. Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Glob. Ecol. Conserv.* **17**, e00547 (2019).

25. da Fonseca, R. R. *et al.* Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Mar. Genomics* **30**, 3–13 (2016).

26. Bohling, J. Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecol. Evol.* **10**, 7585–7601 (2020).

27. Valiente-Mullor, C. *et al.* One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLOS Comput. Biol.* **17**, e1008678 (2021).

28. Vaux, F., Dutoit, L., Fraser, C. I. & Waters, J. M. Genotyping-by-sequencing for biogeography. *J. Biogeogr.* **50**, 262–281 (2023).

29. Jackson, B. C., Campos, J. L. & Zeng, K. The effects of purifying selection on patterns of genetic differentiation between Drosophila melanogaster populations. *Heredity* **114**, 163–174 (2015).

30. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* **4**, 981–994 (2003).

31. Yang, Z. *et al.* Multi-omics provides new insights into the domestication and improvement of dark jute (Corchorus olitorius). *Plant J.* **112**, 812–829 (2022).

32. Zeng, L. *et al.* Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* **20**, 79 (2019).

33. Zhernakova, D. V. *et al.* Genome-wide sequence analyses of ethnic populations across Russia. *Genomics* **112**, 442–458 (2020).

34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

36. Pfeifer, S. P. From next-generation resequencing reads to a high-quality variant data set. *Heredity* **118**, 111–124 (2017).

37. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* **22**, 8 (2021).

38. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).

39. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).

40. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).

41. Rochette, N. C. *et al.* On the causes, consequences, and avoidance of PCR duplicates: Towards a theory of library complexity. *Mol. Ecol. Resour.* **n/a**, (2023).

42. Singh, A. K. *et al.* Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Med. Genomics* **14**, 214 (2021).

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

44. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).

45. Eaton, D. A. R. & Overcast, I. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* **36**, 2592–2594 (2020).

46. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).

47. Mona, S., Benazzo, A., Delrieu-Trottin, E. & Lesturgie, P. Population genetics using low coverage RADseq data in non-model organisms: biases and solutions. (2023) doi:10.22541/au.168252801.19878064/v1.

48. Wright, B. *et al.* From reference genomes to population genomics: comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. *BMC Genomics* **20**, 453 (2019).

49. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945–959 (2000).

50. Waples, R. S. Testing for Hardy–Weinberg Proportions: Have We Lost the Plot? *J. Hered.* **106**, 1–19 (2015).

51. Sethuraman, A. *et al.* Continued misuse of multiple testing correction methods in population genetics-A wake-up call? *Mol. Ecol. Resour.* **19**, 23–26 (2019).

52. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

53. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **6**, 65–70 (1979).

54. Allendorf, F. W. *et al. Conservation and the Genomics of Populations*. (Oxford University Press, 2022). doi:10.1093/oso/9780198856566.001.0001.

55. Gautier, M. *et al.* The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* **22**, 3165–3178 (2013).

56. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genet.* **15**, e1008302 (2019).

57. McKinney, G. J., Waples, R. K., Seeb, L. W. & Seeb, J. E. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol. Ecol. Resour.* **17**, 656–669 (2017).

58. Pearman, W. S., Urban, L. & Alexander, A. Commonly used Hardy–Weinberg equilibrium filtering schemes impact population structure inferences using RADseq data. *Mol. Ecol. Resour.* **22**, 2599–2613 (2022).

59. Larson, W. A. *et al.* Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (Oncorhynchus tshawytscha). *Evol. Appl.* **7**, 355–369 (2014).

60. Waples, R. K., Larson, W. A. & Waples, R. S. Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity* **117**, 233 (2016).

61. Gattepaille, L. M., Jakobsson, M. & Blum, M. G. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity* **110**, 409–419 (2013).

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

63. Cubry, P., Vigouroux, Y. & François, O. The Empirical Distribution of Singletons for Geographic Samples of DNA Sequences. *Front. Genet.* **8**, (2017).

64. Shafer, A. B. A. *et al.* Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods Ecol. Evol.* **8**, 907–917 (2017).

65. Linck, E. & Battey, C. J. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol. Ecol. Resour.* **19**, 639–647 (2019).

66. Andersson, B. A., Zhao, W., Haller, B. C., Brännström, Å. & Wang, X.-R. Inference of the distribution of fitness effects of mutations is affected by single nucleotide polymorphism filtering methods, sample size and population structure. *Mol. Ecol. Resour.* **n/a**, (2023).

67. Díaz-Arce, N. & Rodríguez-Ezpeleta, N. Selecting RAD-Seq Data Analysis Parameters for Population Genetics: The More the Better? *Front. Genet.* **10**, (2019).

68. Hendricks, S. *et al.* Recent advances in conservation and population genomics data analysis. *Evol. Appl.* **11**, 1197–1211 (2018).

69. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat. Rev. Genet.* **10**, 639–650 (2009).

70. Roesti, M., Salzburger, W. & Berner, D. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* **12**, 94 (2012).

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

72. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

73. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64–69 (2012).

74. Dementieva, N. V. *et al.* Assessing the effects of rare alleles and linkage disequilibrium on estimates of genetic diversity in the chicken populations. *Animal* **15**, 100171 (2021).

75. Lou, R. N., Jacobs, A., Wilder, A. P. & Therkildsen, N. O. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol. Ecol.* **30**, 5966–5993 (2021).

76. De Meeûs, T. Revisiting F IS, F ST, Wahlund effects, and null alleles. *J. Hered.* **109**, 446–456 (2018).

77. Graffelman, J., Jain, D. & Weir, B. A genome-wide study of Hardy–Weinberg equilibrium with next generation sequence data. *Hum. Genet.* **136**, 727–741 (2017).

78. Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).

79. Zhang, H., Yin, L., Wang, M., Yuan, X. & Liu, X. Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. *Front. Genet.* **10**, (2019).

80. Anderson, E. C. & Garza, J. C. The Power of Single-Nucleotide Polymorphisms for Large-Scale Parentage Inference. *Genetics* **172**, 2567–2582 (2006).

81. Dussault, F. M. & Boulding, E. G. Effect of minor allele frequency on the number of single nucleotide polymorphisms needed for accurate parentage assignment: A methodology illustrated using Atlantic salmon. *Aquac. Res.* **49**, 1368–1372 (2018).

82. Thompson, E. The estimation of pairwise relationships. *Ann Hum Genet* **39**, 173–188 (1975).

83. Van Etten, J., Stephens, T. G. & Bhattacharya, D. A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data. *Syst. Biol.* syad037 (2023) doi:10.1093/sysbio/syad037.

84. Todd, E. V., Black, M. A. & Gemmell, N. J. The power and promise of RNA-seq in ecology and evolution. *Mol. Ecol.* **25**, 1224–1241 (2016).

85. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).

86. Brown, A. V. *et al.* Ten quick tips for sharing open genomic data. *PLOS Comput. Biol.* **14**, e1006472 (2018).

87. Zhang, D. *et al.* PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Mol. Ecol. Resour.* **20**, 348–355 (2020).

88. Tanjo, T., Kawai, Y., Tokunaga, K., Ogasawara, O. & Nagasaki, M. Practical guide for managing large-scale human genome data in research. *J. Hum. Genet.* **66**, 39–52 (2021).

89. Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F. M. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. *PLOS ONE* **8**, e85024 (2013).

90. Yang, S.-F., Lu, C.-W., Yao, C.-T. & Hung, C.-M. To Trim or Not to Trim: Effects of Read Trimming on the De Novo Genome Assembly of a Widespread East Asian Passerine, the Rufous-Capped Babbler (Cyanoderma ruficeps Blyth). *Genes* **10**, (2019).

91. Hotaling, S. *et al.* Demographic modelling reveals a history of divergence with gene flow for a glacially tied stonefly in a changing post-Pleistocene landscape. *J. Biogeogr.* **45**, 304–317 (2018).

92. 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. *Mol. Ecol.* **31**, 4544–4557 (2022).

93. Cumer, T. *et al.* Double-digest RAD-sequencing: do pre- and post-sequencing protocol parameters impact biological results? *Mol. Genet. Genomics* **296**, 457–471 (2021).

94. Mastretta-Yanes, A. *et al.* Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* **15**, 28–41 (2015).

95. Ebbert, M. T. W. *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* **17**, 239 (2016).

96. Euclide, P. T. *et al.* Attack of the PCR clones: Rates of clonality have little effect on RAD-seq genotype calls. *Mol. Ecol. Resour.* **20**, 66–78 (2020).

97. Flanagan, S. P. & Jones, A. G. Substantial differences in bias between single-digest and double-digest RAD-seq libraries: A case study. *Mol. Ecol. Resour.* **18**, 264–280 (2018).

98. Goubert, C. *et al.* A beginner's guide to manual curation of transposable elements. *Mob. DNA* **13**, 7 (2022).

99. Storer, J. M., Hubley, R., Rosen, J. & Smit, A. F. A. Curation Guidelines for de novo Generated Transposable Element Families. *Curr. Protoc.* **1**, e154 (2021).

100. Escoda, L., González-Esteban, J., Gómez, A. & Castresana, J. Using relatedness networks to infer contemporary dispersal: Application to the endangered mammal Galemys pyrenaicus. *Mol. Ecol.* **26**, 3343–3357 (2017).

101. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

102. Peñalba, J. V., Peters, J. L. & Joseph, L. Sustained plumage divergence despite weak genomic differentiation and broad sympatry in sister species of Australian woodswallows (Artamus spp.). *Mol. Ecol.* **31**, 5060–5073 (2022).

103. Thompson, N. F. *et al.* A complex phenotype in salmon controlled by a simple change in migratory timing. *Science* **370**, 609–613 (2020).

104. Howe, K. *et al.* Significantly improving the quality of genome assemblies through curation. *GigaScience* **10**, giaa153 (2021).

105. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

106. Michael, T. P. & VanBuren, R. Building near-complete plant genomes. *Genome Stud. Mol. Genet.* **54**, 26–33 (2020).

107. Tettelin, H. & Medini, D. The pangenome: Diversity, dynamics and evolution of genomes. (2020).

108. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).

109. Hemstrom, W. Thirty-four Kilometers and Fifteen Years: Rapid Adaptation at a Novel Chromosomal Inversion in Recently Introduced Deschutes River Three-spined Stickleback. (2016).

110. Hemstrom, W. *et al.* Population genomic monitoring provides insight into conservation status but no correlation with demographic estimates of extinction risk in a threatened trout. *Evol. Appl.* **15**, 1449–1468 (2022).

111. Hemstrom, W. & Jones, M. snpR: User friendly population genomics for SNP data sets with categorical metadata. *Mol. Ecol. Resour.* **23**, 962–973 (2023).

112. Francis, R. M. pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32 (2017).

113. Waples, R. S. & Do, C. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evol. Appl.* **3**, 244–262 (2010).

114. Whitlock, M. C. & Lotterhos, K. E. Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST. *Am. Nat.* **186**, S24–S36 (2015).

115. Asif, H. *et al.* GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Mol. Psychiatry* **26**, 2048–2055 (2021).

116. Roorkiwal, M. *et al.* Genome-wide association mapping of nutritional traits for designing superior chickpea varieties. *Front. Plant Sci.* **13**, (2022).

117. Halvorsen, S., Korslund, L., Mattingsdal, M. & Slettan, A. Estimating number of European eel (Anguilla anguilla) individuals using environmental DNA and haplotype count in small rivers. *Ecol. Evol.* **13**, e9785 (2023).

Figure Legends:

Figure 1: Challenges and potential solutions related to filtering that occurs prior to variant calling ("pre-variant" filtering). In the first row, individual 1 has both sufficient read depth (here illustrated with 9 reads, but a higher read depth would be better) and read alignment quality ("quality scores") to allow for successful variant calling (a SNP is circled in orange). Despite this seemingly successful read alignment, challenges still exist; genome assembly errors (such as misplaced scaffolds on reference genomes), and structural variants (such as inversions) can cause issues downstream. Filtering for linkage disequilibrium can resolve some of these concerns. Individual 2 has a weak alignment both across an entire read and at a single base pair position across all reads, both of which should be filtered out prior to variant calling. Individual 3 has too few reads; this individual should be removed, re-sequenced, or, if this low-quality is expected and occurs across all individuals in a study, genotype likelihoods could be calculated. Individual 4 has too many reads, which can be caused by paralogs, copy number variants (CNVs), or technical (*e.g.*, PCR) duplicates. These excess reads should be filtered out and then analysed carefully to determine the underlying causes and to facilitate answering questions of interest.

Figure 2: Challenges associated with four common filters that can occur after variant discovery ("post-variant" filtering). Panel A illustrates missing data, which can occur across loci and individuals. Data from monarch butterflies[92] are used to show that the percentage of missing data can be high (21-100% per locus) when missing data filtering occurs within sample-groups, but if much lower if performed across all samples (19-56%), obscuring the data quality differences between populations. In panel B, we illustrate different minor allele frequencies and use a study on yellow perch[71] to show that the number of SNPs can vary 3-fold among populations if a MAF filter of 0.01 is applied within sample groups instead of study-wide. In panel C, we illustrate how loci can be out of Hardy-Weinberg Equilibrium (HWE) due to homozygote or heterozygote excesses. With the same monarch dataset, we show that combining two divergent sample-groups results in higher F_{IS} and many more loci out of HWE. Filtering study-wide would therefore cause

754 the erroneous removal of loci due to Wahlund effects (see[75]). Lastly, in panel D, we
755 illustrate linkage disequilibrium (LD) with haplotypes that are perfectly correlated across
756 individuals. We use a three-spined stickleback dataset[109] to illustrate that LD thinning
757 (orange points) obscures an inversion (blue points).
758
759 **Figure 3:** Flow chart to facilitate thoughtful, systematic, and reproducible filtering. Typical
760 filtering proceeds through raw sequence QC filtering, alignment, mapped-read filtering,
761 and variant discovery. After variant discovery, investigators must decide whether to apply
762 filters study-wide or within sample-groups and whether to filter by locus or individuals first
763 (see main text for recommendations). Regardless of the study objectives, multiple
764 datasets should be constructed to examine the effects of various filtering decisions. Data
765 should be carefully archived before filtering, and all filtering methods and results carefully
766 reported. See Table 1 for a complete list of filters, Table 2 for a simplified example of how
767 to report filtering results, and Box 3 for a checklist.

Glossary

ALIGNMENT

Mapping of sequencing reads and/or contigs to either each other (pairwise/multiple alignment) or to a **reference**. Alignments can vary in the strength of the evidence that supports them. Strong alignments, for example, usually have a **base (Phred) quality score** of > 20 (99% certainty); weak alignments may have a Phred score of 10 or less (<90% certainty).

BASE QUALITY SCORE

Value in a logarithmic, Phred scale given to each base on a sequencing read that indicates a quantitative degree of confidence in a nucleotide called from the sequencing instrument.

DE NOVO ASSEMBLY

Reference free alignment of reads into overlapping stacks or contigs for subsequent use in variant discovery and genotyping.

DISCORDANT READ-PAIRS

Paired-end reads that do not match either the expected relative directions (5' to 3') or the physical distance (base pairs) between reads.

FAMILY STRUCTURE

Non-independence among individuals in a study caused by direct recent shared ancestry (*e.g.*, parentage, siblings, cousins) between groups of individuals.

FILTERING

The use of quality control steps to remove errors, reads, individuals, loci, or genotypes from a dataset to improve data quality for specific analyses. This procedure differs from other forms of data processing in that it focuses on the *removal* (or "filtering out") of data specifically rather than other forms of quality improvement such as **imputation**.

GENETIC VARIANT

A polymorphism or change in a DNA (or RNA) sequence. Includes both sequence variants (such as single-nucleotide polymorphisms) and structural variants (such as chromosomal inversions, indels, and copy-number variations). Used interchangeably with "loci" in this review.

GENOME WIDE ASSOCIATION STUDY (GWAS)

A test for statistical relationships between a phenotype (including disease) and the allelic/genotypic state of individuals across the entire set of sequenced loci. GWAS is "genome wide" in that associations are tested at many loci spread throughout the entire genome.

GENOTYPING

Calling the allelic states at a locus (*e.g.*, A/A, A/C, or C/C at a biallelic SNP in a diploid organism) or at many loci based on the underlying sequence data. Genotyping algorithms often consist of multiple steps during which filtering can occur.

HARDY-WEINBERG PROPORTIONS (HWP)

The expected frequencies of the genotypes at a given locus under Hardy-Weinberg Equilibrium (HWE). Filtering on HWP is often executed via an exact test, with loci that deviate significantly from HWP removed from subsequent analyses.

IMPUTATION

The filling in of missing data for specific genotypes and/or loci by leveraging linkage disequilibrium between missing genotypes and the called genotypes at other loci. Imputation can use reference panels of well-described haplotypes to improve performance when available, usually in well-studied model organisms.

LINKAGE DISEQUILIBRIUM (LD)

When the genotypic state of individuals in a dataset/sample group at one locus are predictive of the genotypic state of individuals at other loci. This can be caused either by physical linkage, when alleles are co-inherited due to non-independent assortment driven by physical proximity on a genome or other factors such as inversions, or gametic phase disequilibrium**,** when underlying population or family structure makes certain alleles at different loci more likely to co-occur.

MAPPING QUALITY

Score given to a read or other DNA sequence indicating the uniqueness of the alignment to a reference sequence; mapping quality algorithms vary between alignment programs.

MINOR ALLELE COUNT (MAC)

The number of gene copies or individuals carrying the minor (*i.e.* least frequent) allele at a locus.

MINOR ALLELE FREQUENCY (MAF)

The proportion (frequency) of the least common allele at a locus across a study or sample group. For example, if one diploid individual out of 50 had one copy of a unique allele (*e.g.*, an A instead of a T), the MAF would equal 0.01 In this review, we often refer to filtering out loci with MAFs below a given threshold as “MAF filtering”.

MISSING DATA

Missing genotype calls at a specific locus or individual. Missing data can be caused by many reasons, most commonly the absence of a sufficient number of reads covering a locus to call a genotype in an individual with any degree of confidence.

PARALOG

A duplicated genomic region that has arisen via either the duplication of that specific region or the entire genome (*e.g.,* genome duplication events). A type of homolog (loci identical by descent) distinct from orthologs, which arise instead due to speciation events.

PCR DUPLICATE

A technical duplicate resulting in spurious, usually identical read copies caused by repeatedly sequencing the same piece of template DNA multiple times.

POPULATION STRUCTURE

Also known as population subdivision. Non-independence among individuals in a study area/region caused by within-group biases in reproduction, usually created by spatial, temporal, or behavioral separation and often responsible for creating allele frequency differences systematically across loci.

REDUCED-REPRESENTATION SEQUENCING

The sequencing of either random or targeted subsets of a genome. Common examples include Restriction-Associated Digest (RAD) sequencing, Genotyping-By-Sequencing, and targeted sequence/exon capture.

REFERENCE

A known, often well-annotated sequence to which reads can be aligned. Can be either a published reference genome or transcriptome for a species or a ***de novo* reference**.

SAMPLE GROUPS

Groups of samples that are not independent of each other created by natural populations, geographic or temporal variation in sampling, and/or experimental treatments.

SINGLE-NUCLEOTIDE POLYMORPHISM (SNP):

A genetic variant where the allelic state of the population varies at a single base-pair.

SEQUENCING DEPTH/COVERAGE

The number of reads that were aligned to and overlap a particular genomic locus. More simply, the number of times a locus was sequenced.

SINGLETON

An allele sequenced only one time across all individuals. Sometimes alternatively defined as an allele sequenced in only one individual (which may be homozygous for that allele).

SITE-FREQUENCY SPECTRUM (SFS)

The distribution of allele frequencies across loci within a study or sample group. Can be either an “unfolded” or “polarized” derived allele frequency spectra which describe the frequency distribution of derived alleles or a “folded” or “unpolarized” minor allele frequency spectra which describes the frequency distribution of the minor alleles. Also known as the allele frequency distribution.

STRUCTURAL VARIANTS

Genetic variants which are caused by underlying variation in the order, number, and/or arrangement of loci. Copy number variations, paralogs, and indels all fall into this category. Sequence variations, where individual or series base-pairs change without changing genomic positions or counts, do not.

STUDY-WIDE FILTERING

Applying a filtering threshold globally (across the entire data set) rather than separately within each sample group.

WAHLUND EFFECT

A reduction in observed homozygosity ($Ho > He$) at many/most loci caused by underlying population structure. When multiple (sub)populations are included in a sample, any differences in allele frequency between (sub)populations will cause there to be considerably more homozygous individuals at those loci than would be expected under HWE (causing an elevated F_{IS}, the reduction in heterozygosity within individuals relative expected heterozygosity). Note that this observation is what underlies the differentiation

889 measure F_{ST} when populations are split and could thus be described as F_{ST} improperly
890 calculated as F_{IS}.

891 WITHIN-GROUP FILTERING
892 Applying a filtering threshold within groups separately rather than globally.

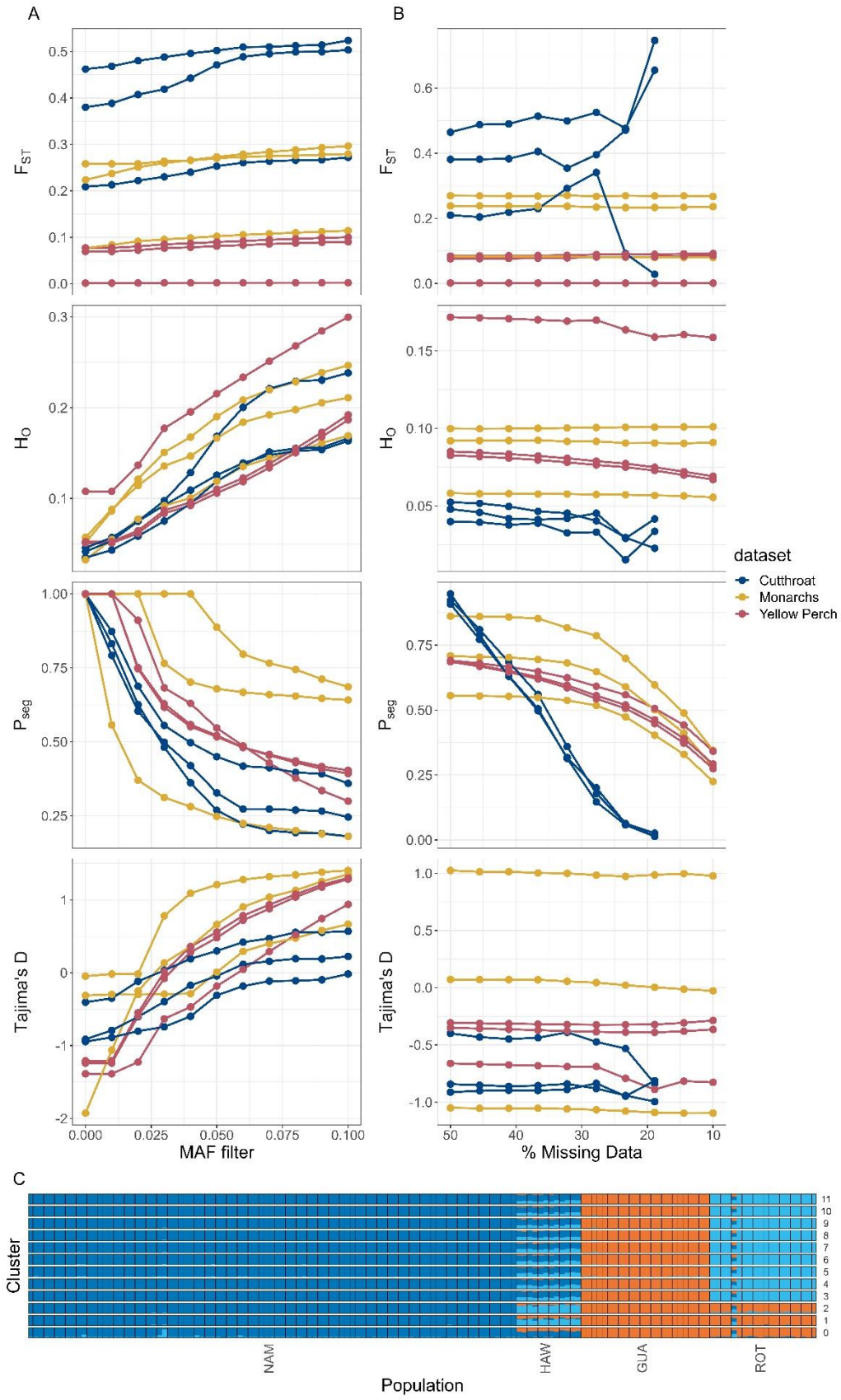
893 WHOLE-GENOME SEQUENCING
894 Sequencing the entire genome without any attempt to target specific regions (in contrast
895 to reduced-representation sequencing). Usually entails “shotgun” sequencing and results
896 in many more sequenced loci than reduced representation sequencing.

Box 1: Effects of Filtering

Genomic filtering choices can have substantial effects on downstream analyses that are not necessarily consistent across samples, populations, or statistical methods. We used three existing datasets (monarch butterflies[92]; cutthroat trout[110]; yellow perch[71] to demonstrate how different MAF (panel A) or missing data (panel B) filters can affect estimates of pairwise F_{ST}[68–70], observed heterozygosity, the number of segregating sites (polymorphic loci; P_{seg}), and Tajima's D[62]. We also show how a MAC filter can affect population structure estimates using the same monarch data filtered with a MAC of between 0 and 11(reproducing methods from[65], see panel C)[111,112].

Our principal finding is that changing MAF and missing data thresholds can impact different datasets and different populations within datasets unequally. For example, while the results monarch and perch data were not strongly affected by increased missing data stringency, outcomes from the cutthroat data changed unpredictably past a certain threshold. The impact of MAF threshold likewise differed between studies and populations: note that the number of segregating sites plummeted with even a mild MAF filter (MAF > 0.0125) in one specific population of monarch butterflies. Tajima's D deserves particular attention—even a relatively low MAF filter (e.g., < 0.01) caused every population in every dataset to flip from an "expansion" signal (D < 0) to a "bottleneck" signal (D > 0)[61,62]. While not shown here, MAF filtering is also of particular note in that it almost uniformly decreases estimates of effective population size (N_e) derived from LD based approaches[113].

In contrast, population structure resolution (panel C) increased with a higher MAF filter, consistent with previous findings[65,68,91]. Some other methods, such as F_{ST} outlier detection[114], GWAS [115,116], and genomic selection[79] can also see improvements with higher MAF thresholds, although not without risk of bias.


A
B
F_{ST}
H_O
P_{seg}
Tajima's D
MAF filter
% Missing Data
dataset
Cutthroat
Monarchs
Yellow Perch
C
Cluster
Population
NAM
HAW
GUA
ROT

Box 2: Filtering Trade-offs

Different filtering choices result in trade-offs described here as false-positives and false-negatives (Type-I, α, and Type-II, β, errors, respectively; see diagram below). In the case of variant calling (*i.e.*, genotyping) we here hold that an incorrectly called genotype is the null hypothesis. While this is not the only reasonable interpretation, it simultaneously allows for a more conservative philosophical approach towards genotyping and allows for power $(1-\beta)$ to equal the proportion (or percentage) of correctly called genotypes that are retained in a data set. Within this framework, a false positive occurs when an incorrectly called genotype (at a single locus) is retained within a data set and a false-negative occurs when a correctly called genotype is incorrectly filtered out of the dataset (see diagram below)[117]. False-positives occur most frequently when filters are not strict (for example, no MAF filtering is performed) and/or when the read depth at a locus is low and thus incorrectly genotyped loci are allowed to proceed into downstream analyses. By contrast, false-negatives will be more likely when strict filters are used (such as a high MAF filter), since more loci are assumed to be errors (and thus removed) even though many of those sites may represent real, correctly called genotypes.

To answer specific questions, certain trade-offs will invariably arise. For example, when calculating Tajima's D, many or most low-frequency sites must be retained (see Box 1), however, this procedure will invariably allow more false-positives into the data set (affecting the precision of the estimator). Alternatively, when performing GWAS or outlier F_{ST} detection low frequency sites are often removed[72,114], which creates data sets with few genotyping errors but that may exclude real, causal variants that segregate at a low frequency. Investigators should be cognizant of the very real trade-offs associated with

certain filtering choices and should consider practical solutions such as creating two or more datasets with different filtering thresholds, sequencing loci of interest to higher depth/confidence, and re-sequencing select samples.

Null hypothesis H_O = Genotype is called incorrectly

Decision about Null hypothesis H_O is :	Null hypothesis H_O is : True	Null hypothesis H_O is : False
Don't reject	Correct inference: An incorrectly called genotype is filtered out of the data set	Type II error: A correctly called genotype is incorrectly filtered out of the data set (false negative)
Reject	Type I error: An incorrectly called genotype is kept in the data set after filtering (false positive)	Correct inference: A correctly called genotype is kept in the data set after filtering

Box 3: Filtering checklist

Throughout dataset assembly (*e.g.*, from raw sequencing reads to genotypes), researchers should perform various steps to explore the effects of alternative filtering strategies on downstream analyses and aid in reproducibility. To ensure a robust examination of filtering effects and study reproducibility, the example checklist below should be consulted before and during a research project and checked-off before submitting a manuscript for peer-review.

☐	Data archival
☐	Decide on filtered data sets given *a priori* study questions
	☐ Create filter recording/reporting table (see Table 2)
☐	Filter on raw sequences (*e.g.*, read quality, poly-G tails; see Table 1)
	☐ Report exact filters used for filtering on raw sequences
	☐ Report total number of reads in study
	☐ Report total number of reads filtered out
☐	Perform sequence alignment; report alignment parameters
	☐ Report total number of reads that aligned successfully
	☐ Report total number of reads that mapped uniquely
	☐ Report total number of reads that were filtered out
☐	Perform filtering on successfully mapped reads
	☐ Filter on mapping quality, PCR duplicates, read pairs
	☐ Report number of reads retained/filtered at each step

☐	Variant discovery
☐	Begin or continue creation of multiple data sets
	☐ Decide on study-wide versus within-sample group filters
	☐ Decide on filter values to employ and order of filters
☐	Locus filtering (see text for when individual filtering should go first)
	☐ Filter for MAF, HWP, paralogs, coverage etc. (see Table 1)
☐	Individual filtering
	☐ Missing data; Mislabeling/contamination
☐	Dataset construction; continue reporting of all filters and reads filtered
☐	Data analysis and parameter estimation
	☐ Report effects of filters on parameters/questions of interest
☐	Perform re-filtering and/or re-sequencing if necessary
☐	Final filter recording (report reads, loci, individuals lost at each step)
☐	Archive all filtered data sets as VCF files

955

956 **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 spurious guanines (“G”s) that can be added to 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 mismatches in the 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 poor mapping scores.
Improperly paired reads (orientation and distance)	*ii*	Removal of paired-reads that are improperly paired (too 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 for computational efficiency, but also to remove potential errors.
Low coverage/Quality by depth	*iii*	Removal of individual called genotypes that had poor coverage. Joint "Quality-by-depth" often alternatively used.
Genotype Quality/Confidence	*iii*	Removal of individually called genotypes with poor genotyping confidence. Joint "Quality-by-depth" often alternatively or additionally used.
High coverage	*iii*	Removal of individual called genotypes with very high coverage (usually indicating errors in the reference, paralogs, or copy-number variants, all of which require additional investigation).
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 (*e.g.*, monomorphic or tri-allelic SNPs); required by many down-stream applications.
Allow/deny listed variants	*iii*	Removal or inclusion of variants from pre-defined loci. Common for methods that target specific loci but may also sequence some additional off-target 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 too few loci.

Missing data - per locus	*iii, iv*	Removal of loci with called genotypes in too few individuals.
Minor allele frequency	*iii, iv*	Removal of loci with low minor allele frequencies.
Minor allele count	*iii, iv*	Removal of loci with a low count of the minor allele across samples.
Hardy-Weinberg proportions	*iii, iv*	Removal of loci significantly out of Hardy-Weinberg proportions.
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 have biased statistical outcomes due to sex-specific sampling or behave in unexpected ways.
Paralogs - allelic imbalance/depth/heterozygosity	*ii, iii, iv*	Removal of reads aligned to paralagous loci, where for recent paralogs it can be unclear from which of the gene copies the read was sequenced. Additional analyses are required.
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 (*e.g.*, 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.

957 **Table 2: Example table demonstrating ideal filtering reporting standards.**

Questions	Level	Mapping Quality	Genotype Quality	MAF/ MAC	HWE*	LD	Missing Data - Individuals	Missing Data - Loci	F_{ST} Outliers	# SNPs **	# Individuals

Population structure	Study-wide	>20	>14 (95% confidence)	> 0.05	p > 0.01 *	Yes	< 20%	< 20%	Yes	400,000	190
Population structure	Study-wide	>20	>14 (95% confidence)	> 0.1	p > 0.01 *	Yes	< 20%	< 20%	Yes	300,000	190
GWAS	Study-wide	>30	>20 (99% confidence)	> 0.05	p > 0.01 *	No	< 20%	< 20%	No	200,000	100
Tajima's D; SFS based estimators	Within Sample-Group	>20	>14 (95% confidence)	MAC > 1	p > 1e-6	No	< 50%	< 50%	No	1,500,000**	200
Genetic diversity	Within Sample-Group	>20	>14 (95% confidence)	MAC > 5	p >1e-4	Yes	< 75%	< 75%	Yes	500,000**	170
Selection; Selective sweeps	Within Sample-Group	>20	>14 (95% confidence)	Any within group MAF > 0.05	p > 1e-4	No	< 75%	< 75%	No	700,000**	180

*Filters for loci out of HWE should only be applied at a within sample-group level (see Pearman *et al* 2022[58]).

**When filtering SNPs within sample-groups, different sample groups will contain different numbers of polymorphic SNPs. The numbers here reflect a hypothetical mean across all sample groups.

In this example we assume that all filtering was performed on the same dataset. Filtering should be first guided by the types of questions investigators wish to answer, but we also strongly recommend that different filtering thresholds should be used to address the same question for comparison. For population structure, for example, investigators may want to examine the effects of several different MAF thresholds, and therefore create several different similar datasets. In practice several thresholds should be used for most questions, and more questions should be tested at multiple filtering thresholds. Notice that different questions suggest that different levels of filtering, types of filters, and filter thresholds should be applied. As a final note, we also suggest that filtered VCF files be retained and archived for each dataset used in the analysis, and the names of those files clearly listed in this table or in the supplementary material for reproducibility purposes.