**Next-generation data filtering in the genomics era**

William Hemstrom1,4†, Jared A. Grummer2,4, Gordon Luikart2, Mark R. Christie1,3†

**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

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

† Corresponding authors:

William X. Hemstrom; Department of Biological Sciences, Purdue University; 915 W. State St., West Lafayette, Indiana 47907; [whemostro@purdue.edu](mailto:whemostro@purdue.edu), 765-496-0062

Mark R. Christie; Department of Biological Sciences & Department of Forestry and Natural Resources, Purdue University; 915 W. State St., West Lafayette, Indiana 47907; [christ99@purdue.edu](mailto:christ99@purdue.edu), 765-494-2070

**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. We use simulated and empirical datasets to illustrate the large effects of these filtering thresholds on statistics such as Tajima’s D, *FST*, *π*, and effective population size (*N*e). Varying filtering thresholds creates similar trajectories in parameter estimates across datasets, however, their scale and rates of change are highly variable. Thus, we emphasise that there is no single best strategy for filtering all genomic datasets. To help usher in the next generation of data filtering, we review best-practices for investigators to apply when analysing genomic data and provide practical filtering threshold recommendations and strategies for improving the implementation, reproducibility, and reporting standards for filtering approaches.

**[H1] Introduction**

Recent and rapid advances in both short- and long-read sequencing technologies (*e.g.*, Illumina, PacBio, and Nanopore) have resulted in the proliferation of large genomic datasets1–4. All high-throughput (‘next-generation’) sequencing methods result in large numbers of reads that, importantly, have variable error rates5,6. In addition to the inherent error rate of the sequencing technology, errors can arise when these sequences are aligned to a ***reference genome*** or transcriptome, particularly if the reference is incomplete, highly divergent, or assembled ***de novo*** from the reads themselves7–9. 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 or guiding principles for data filtering.

Correctly filtering genomic data is not a trivial task. 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, individual genotypes, 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 dataset 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 can yield entirely different results12,13 and 3) filtering choices can be confusing and subjective, currently lack consistent or agreed-upon guidelines, and may result in unknown downstream consequences14.

Filtering approaches are often inadequately described and vary widely among studies. It is frequently impossible to tell whether certain filtering steps were conducted and not described or skipped entirely. Even when filters are mentioned, methodological details (such as the thresholds applied) are infrequently reported. Among papers that *do* record filtering thresholds, the values used for a given filter can vary several-fold, suggesting limited standardisation 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 comment15. Given the lack of comprehensive genomic data filtering principles 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 often not sufficient for reproduction.

To help investigators improve their filtering approaches, we review different strategies for filtering genomic datasets and examine their downstream effects by reporting the effects of a range of filtering decisions on both empirical and simulated datasets. We also provide practical filtering threshold recommendations and a comprehensive suite of resources including a detailed flowchart, 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 alike. We illustrate a typical filtering workflow by providing coding notebooks that contain examples of how to automate and easily examine the output from multiple filtering parameter values. Lastly, we stress the importance of knowing your study system and population genetics theory16–18, both of which will help researchers make informed choices for setting filtering thresholds and facilitate the interpretation of different results inevitably produced with different filtering thresholds.

**[H1] Common filters: complexity, and importance of study system knowledge**

Today, investigators have many different instrumentation options for obtaining DNA, RNA, and epigenetic sequencing data19,20, which subsequently yield reads of different lengths, configurations (*e.g.,* single- or paired-end), and qualities21. 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 of reads) for subsequent variant calling or analyses22–25. Without some form of reference-guided alignment, some downstream analyses can be limited because it can be difficult to estimate relevant parameters (*e.g.*, runs of homozygosity26, recombination points) or determine the genomic context for loci of interest (*e.g.*, linkage, ***haplotype phase***)27. Lacking a quality reference genome can therefore be a challenge for researchers working in non-model systems where ***de novo alignments***, alignments to low-quality reference genome assemblies (*e.g*., many contigs and low N/L50 scores28), or alignments to a different, distantly related species29,30 may be required. 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 minimise downstream biases and maximise data quality.

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 phenomena31–34 and thereby suggest the need for further filtering steps. Thus, the process of filtering begins immediately after sequence data collection and may not end until all analyses are completed and researchers are confident that their filtering choices were optimal or did not systematically influence conclusions. A comprehensive list of filters, their descriptions, and the typical stage in the workflow where they are applied can be found in Supplementary Table 1.

Below, 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 Supplementary 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* assemblyor alignment to a reference, data are usually filtered via the removal of low-quality reads (see Supplementary Table 1). This initial filtering may h ave 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 535,36 to 4037 (between 69% and 0.01% maximum allowable error rates). Reads that have too many bases below that quality threshold are subject to removal; the exact number of allowable low-quality bases varies widely across studies. Low-quality reads can also be removed during the process of alignment itself, because different alignment algorithms can prevent sequences from mapping back to a reference if their quality scores are below user-defined thresholds38–41.

Conversely, high-quality reads may also be excluded if the reference does not contain the sequence, is highly repetitive (*e.g.*, transposable elements), or found in multiple genomic locations (*e.g.*, paralogs; see Figure 1). 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 a soft filter based on a statistical model40). Below, we provide further recommendations on reporting across all filtering steps.

Reporting these statistics can help reviewers assess the quality of the data underlying the study and allow future investigators to determine whether alternative filtering strategies 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***42,43, genome assembly errors44, ***structural variants***45, and challenging alignments (*e.g.*, transposable elements46, PCR duplicates47, and copy-number variants48,49) 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), andstructuralvariantsare algorithmically identified with software such as GATK50, ANGSD51, STACKS9, ipyRAD52, or LUMPY53, or others. During this process, the read depth (or coverage) of each locus must be considered, since a greater depth of coverage allows for more confidence in genotyping (and subsequently downstream inferences). Very high read depths (in relation to the study-wide mean), by contrast, can be indicative of paralogs, highly repetitive regions, copy number variants (CNVs), mitochondrial DNA loci, or technical (*e.g.*, PCR) duplicates (see Figure 1, Supplementary Table 1). Variant calling algorithms typically either mark genotypes as missing if below a certain read depth or if they have poor quality scores54. The depth and/or quality filters used at this step vary substantially between studies. Filtering loci sequenced at a low depth is not without risks, however; since calling heterozygotes requires higher depth than homozogytes, stringent depth filtering can skew observed heterozygosities and site frequency spectra55,56. However, well-developed approaches to make use of low coverage sites and mitigate such biases do exist51,56, and so filtering out such loci is not always necessary.

While many different ways to filter genotypic data exist (Supplementary 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. In Box 1, we examine simulated and empirical datasets to show the systematic effects of these filters, and in the “Solutions and Best Practices” section we suggest how to best implement each filter for different research questions.

*[H3] Missing data*

Missing data can result from either the absence of reads covering a locus in an individual or from upstream read or genotype quality, depth, mapping, or other filtering methods. Loci and/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 potentially important variation57. The filtering choices used for missing data vary *widely* among studies and the downstream consequences are rarely evaluated.

*[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 due to genotyping errors. MAF filtering is often performed at a specific threshold of 0.05 to reflect this, 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 separately 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). For example, in a sample of 30 diploid individuals a MAF of 0.05 would remove SNPs where the minor allele occurs three times or fewer, whereas in a sample of 60 diploid individuals the same MAF of 0.05 would remove SNPs where the minor allele occurs six times or fewer. By contrast, a MAC of two would remove SNPs where the allele occurs two times or fewer regardless of the sample size. MAC filters may be particularly useful when sample sizes are small or highly variable because a typical MAF filter (*e.g.*, 0.05) will never remove loci (including singletons) in small samples (*n* < 10 diploid individuals).

*[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.,* STRUCTURE58), and removing loci that violate HWP can help ensure unbiased results for downstream analyses in randomly mating popuations59. HWP deviations often reflect sequencing, assembly, or alignment errors (such as a heterozygote deficit caused by allelic dropout or a heterozygote excess caused by paralogous regions; see43,60–62). However, loci out of HWP often indicate real biological phenomena, such as cryptic population sub-structuring (Figure 2), or balancing selection. As a result, it is crucial to filter for HWP within sample groups (*e.g.*, populations) rather than study-wide (*e.g.*, globally on all samples; see *Study-wide versus within sample-group filtering*, below)63 and to do so with a low stringency if loci under selection or those that differ between populations are of interest.

The thresholds used for HWP testing differ from the other filters because HWP testing is *hypothesis testing*, and, as such, produces *p-*values upon which filtering thresholds are set. Because thousands of loci may be considered at once, multiple testing corrections should be considered; however, this is rarely performed15. That said, HWP *p-*value correction differs from typical testing approaches in that, from a broader perspective, not doing corrections *is the conservative approach,* since it will result in the removal of more potentially problematic loci, not less. 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–Hochberg64 to sequential or simple/stringent Bonferroni65), and, ultimately, different approaches and alpha thresholds should be applied depending on the questions being asked and the tolerance for including problematic loci16 (Box 2). For an in-depth discussion on implementing and interpreting HWP tests, see16,59,66).

*[H3] Linkage disequilibrium (LD)*

Pruning clusters of loci that are in substantial LD with each other down to a single locus ensures statistical independence among loci — a common assumption made by many downstream methods. For example, methods based on the ***site frequency spectrum*** (SFS) of a population may be biased if correlated allele frequencies in a variant-rich region differ from the genome-wide average, and failure to remove non-independent (linked) loci can bias estimates of parameters like effective population size (*N*e)67. However, filtering out SNPs based on LD could also strongly influence diversity estimates (such as the number of segregating sites across genomic regions) or inadvertently cause investigators 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 contigs68, which can be accomplished through pairwise correlation measures such as Pearson’s *r*2. Alternatively, many investigators working with *de novo*-assembled datasets often simply extract a single SNP from each contig to mitigate the effects of linkage (though this assumes distinct stacks/contigs are themselves unlinked, which may not be true). Corrections for multiple testing are also important for LD filtering, but seldom reported15.

**[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 (potentially multiple times) with different thresholds and at different data processing stages (Supplementary 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. Doing so may be alluring, 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 crucial 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 reach entirely different biological conclusions.

To illustrate this principle, we systematically filtered ten empirical and three simulated datasets by changing filtering thresholds for three key filters (MAF, missing data, and HWP) in Box 1. While MAF filters are often applied to remove **singletons or other rare variants** as described above (*e.g.,* MAF < 0.05), these variants are critical to 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 selection69, 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 D70, researchers should apply both *no* MAF filter and a *very minor* one (such as a singleton filter) and compare the results when using the statistic70. The effects of additional filters (including MAF) can be substantial for diversity estimators71, demographic inference12,72, *F*ST60, gene flow14, population structure estimates73,74, estimating the distribution of locus effects on phenotypes74, and allele frequency spectra60,75–77. Other filtering choices therefore require similar levels of care (Box 1).

***[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 within-group MAF filter of 0.01 was applied to a yellow perch (*Perca flavescens*) ***whole-genome sequencing (WGS)*** dataset, the number of SNPs within each population varied by a factor of 3.3 (ranging from 670,578 to 2,275,935; Figure 2). However, when the same 0.01 MAF filter was applied globally, each sample group was constrained to 714,000 SNPs78. In this case, some populations in the study had radically different SFS, likely caused by recent population expansions in some but not all populations69,70. Study-wide filtering can therefore lead to the removal of critically informative, globally rare, but locally common alleles. Filtering MAF globally (a common practice) instead of within study groups should generally be expected to have serious effects whenever SFS vary between sample groups, such as when demographic histories differ 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 with the threshold dictated by sample size (which can often be quite large, particularly in human or agricultural work79). The implications of these standardised 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 ancestry80. 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 HWP63 and LD81 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 points22. Tests for HWP, for example, should be always conducted on each sample group separately, because pooling genetically-distinct groups will result in 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)59,82. If a specific locus shows a consistent HWP deviation greater than the genome-wide trend in multiple different groups, it may indicate genotyping error (*e.g.*, allelic drop-out) or alignment or genome assembly errors (beyond natural/biological processes66).

**[H1] The importance of study system knowledge**

Due to the potential impacts of filtering, we strongly recommend that researchers have an excellent understanding of both their study system and population genetics theory before planning filtering strategies and interpreting results16. Critically, knowing a species’ ecology, demography, and pre-existing genetic results (*e.g.*, from microsatellites) can provide important *a priori* expectations for analytical results that can suggest sources of filtering (and other) error. At every step, researchers should ask themselves - given their knowledge of their study species, theory, and model assumptions - “Do these results and filtering choices make sense?”

For example, little geographic structuring might be expected in high gene flow species such as many migratory birds or marine organisms, and thus detection of relatively strong population subdivision (or high *F*ST values) could be an artefact of data filtering60,73. Similarly, knowing the mating system, degree of population isolation, and dispersal propensity of the species can help researchers determine if, for example, high inbreeding and/or effective sizes (*N*e) are biological in origin or produced artificially by the filtering choices. Strong genetic signals of recent population bottlenecks in populations known to have undergone demographic expansions might also suggest filtering issues (Box 1). A solid understanding of a species’ biology is also useful for model species and taxa, where known recombination rates, genomic organisation83, and different histories of captive breeding84 can help predict and interpret filtering results.

**[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, labeling, 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 population85, and failing to account for them may curtail power to identify causal associations or even lead to incorrect inferences54.

Because different filtering choices can result in different downstream inferences, we recommend that ***distinctively-filtered versions of the same dataset should be used to quantify the effects of filtering and to address specific research questions*.** A *minimum* of two 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.). Creating two datasets using extreme filtering values (*e.g.,* low and high stringency) allows a researcher to test if distinct filtering thresholds affect their dataset and downstream conclusions; if the effects are small, no further filtering may be needed.

Investigators should also remember that different questions or approaches may require different sets of filters, reflecting the specificities of the study. For example, researchers should consider using low-stringency MAF filters for several demographic inferences (*e.g*., Tajima’s D; SFS) but relatively stringent MAF filters for delineating populations73, planning genomically-informed breeding strategies86, or for estimating parentage or individual relatedness87–89. 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 accuracy90,91.

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. Investigators should report the effects of their filters 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. Some stand-out papers exist that already use and report the effects of different filters14,74,92,93, although they are in the minority. Note that we are not the first to suggest comparing outcomes from different filtering strategies12,14,60, and we suspect that this will become more common, and more commonly followed, over time.

The concept of using multiple, differently filtered datasets requires a fundamental shift in the way genomics data are analysed: investigators must realise that no single “best” filtering strategy or filtered dataset exists. No filtering method will remove all errors, but re-filtering with different thresholds can provide higher certainty that there is not a substantial bias from error (Box 2). We provide recommendations for initial filtering thresholds for many types of questions and analyses in Table 1 and justify these recommendations in Supplementary Table 4. These thresholds can and should be modified given the characteristics (such as sample sizes and general quality) of the data at-hand.

To assist investigators in matching research questions with methods and filters, we have created a detailed flow chart that can be applied generally across disciplines and study systems (Figure 3). We recognize that some alignment-free methods exist that use high throughput sequencing data, particularly for metagenomics and other phylogenomic analyses94, and they are not considered extensively here. Likewise, filtering for RNA-seq is not thoroughly reviewed here; while many concepts hold true (especially for SNPs called from RNA-seq data), many of the specific filters may not apply (see95,96). Thus, we do not provide extensive specific filtering guidelines for RNAseq, microbiomes, eDNA, and metagenomic/metabarcoding datasets as these guidelines are provided elsewhere and are beyond the scope of this review28,97–106. Some methods have different, specific filtering requirements, such as phylogenetics107 and germ-line specific medical genetics108. For most other studies we here highlight the salient features of a genomics study workflow.

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

First, we note that most genomic workflows differ depending on the research questions and data types. The documentation of filtering decisions is therefore paramount for reproducibility and research efficiency. As a first step, we recommend that raw data be immediately archived (privately or publicly) in independent, non-local repositories created for genomics data (*e.g.*, the NCBI Short-read Archive, the European Variation Archive, or the DNA DataBank of Japan) prior to any analysis (see109,110 for reviews on genomics data management best 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 to111 for information on dataset and study organisation.

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 useful112,113. 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 in92,114, 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-studied24,25,115,116.

After alignment, the data should be filtered for technical (*e.g.*, PCR) duplicates. While removing PCR duplicates has been suggested to be of little consequence117,118, this is unlikely true for every study, such as those with low-coverage data11,47,119. The remaining reads should then be filtered for mapping and read quality, and researchers should ensure that they record and report the number of reads that passed these pre-variant filters (see Table 2 as an example). We have provided an R notebook and a small test dataset, available in the Supplementary Information and referenced in Figure 3, which walks through an entire pre-variant filtering workflow - from raw reads to called genotypes - using a variety of commonly implemented tools, which provides an example of how to easily change, and importantly record, filtering parameters with minimal effort (Notebook S1; see also Table 2).

***[H2] Best practices for post-variant calling workflows***

Following pre-variant filtering, the next steps are to call variants, filter the resulting dataset 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 the order in which filters are applied is important: 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 individuals8. For example, filtering out loci first without strong missing data filters can result in the subsequent removal of all or most individuals if many loci have high proportions of missing data (see the *Daphnia* lines in Figure S8 which truncate below missing data requirements of 40-50% for loci, below which no individuals are subsequently retained after filtering due to high missingness). 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 versa8. Similarly, if a MAF filter is used to remove loci after variant discovery and then individuals with too much missing data are subsequently removed, a second round of MAF filtering could be considered to remove loci that now fall below the MAF threshold. As with pre-variant filters, the percentage of reads, sites, and individuals retained at each post-variant filtering step should be reported (Table 2; Box 3).

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, mis-classified into an incorrect sample group, contaminated during sample preparation, or closely related *(e.g.*, full-siblings). Such underlying causes should be carefully investigated and problematic samples reported and possibly removed prior to the re-calculation of downstream statistics. 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.

After analyses are completed, the resulting data should again be archived and/or recorded, including all relevant meta-data and the exact filtering decisions. Given that recreating distinctly filtered datasets requires a considerable amount of resources (and may actually be impossible given limited data, computational limitations, improper archival, unmet dependencies, or restricted access to old or out-of-date software), 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 (see120 for a detailed description of VCF files).

We strongly suggest that authors and journals require supplementary tables that describe 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 Tables 1 and 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.; see15). If reasonable, we also suggest that the downstream statistical effects of any different filtering thresholds used be reported (either in the main or supplementary text) to help build a better community understand of filtering thresholds. Lastly, we suggest that coding notebooks or scripts containing the exact software employed, the specific commands used in that software, and the flags and parameter values chosen should be submitted alongside the reporting tables. We provide an example notebook containing a post-variant filtering workflow in SI Notebook 2.

***[H2] The benefits of re-filtering and reproducible research***

We recognize that a proper and thorough examination of filtering (and re-filtering) will necessitate extra time, computational resources, and work from researchers. However, these changes to current workflows (*i.e.,* re-filtering) 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 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, see73,121,122.

Using workflows like this, researchers can plan ahead and design their analyses to minimise the time needed to re-run filtering steps (Figure 3) if required. Journal reviewers should, in kind, be reasonable when asking authors to reanalyse their data with different filtering parameters. For example, if authors have adequately justified their filtering choices and demonstrated that filters are unlikely to have biased their findings by running and quantifying the effects of two or more sets of filtering thresholds, reviewers generally should not suggest additional filters. The importance of communicating expert knowledge of a study system early and often while choosing filtering parameters and assessing their impacts could go a long way to minimising unnecessary filtering for students and early-career biologists. Having knowledge of a study system, previous data (*e.g.*, microsatellite *FST*), and population genetics theory while choosing filtering parameters will also help researchers justify their filtering choices.

**[H1] Conclusions and future directions**

Advancements in genomic sequencing technologies, improvements in reference quality123–125, and the burgeoning field of pangenomics126,127, will increase the accuracy and power of genomic data analyses. Nonetheless, filtering will remain a central part of all genomic analyses for decades to come because no genomic dataset will ever be error-free. Investigators should strive to filter with a focus on reproducibility and aim to match the filters they employ to their study species (*e.g.*, demography, life history, etc.) and the questions they intend to answer. Critically, we highlight that for the same dataset: 1) different filtering thresholds can create different downstream results and conclusions, 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 to improve certainty in the conclusions drawn from analyses. As more papers quantify filtering effects, the scientific community will better understand effects of filtering choices on downstream inferences, which will advance genomics applications across disciplines from ecology and evolution to human health, agriculture, and biodiversity conservation.

**Acknowledgements**

We thank E.Anderson, A. Leaché, M. Kardos, and three anonymous reviewers for their helpful comments that greatly improved this manuscript. We also thank M. Exposito-Alonso and the 1001 Genomes Consortium, the 1000 Genomes Project, C. Kessler, M. Freedman, M. Kardos, M. Lynch, R. Malison, D. Martchenko, M. Miller, R. Schweizer, A.B.A. Shafer, and X. Yin for allowing us to review and re-filter their datasets. MRC was funded, in part, by NSF DEB-1856710 and OCE-1924505. GL was funded in part by NSF-DOB-M66230.

**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. 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).

13. 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).

14. 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).

15. 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).

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

17. Gervais, L. *et al.* RAD-sequencing for estimating genomic relatedness matrix-based heritability in the wild: A case study in roe deer. *Mol. Ecol. Resour.* **19**, 1205–1217 (2019).

18. Crow, J. F. *An Introduction to Population Genetics Theory*. (Scientific Publishers, 2017).

19. 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).

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

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

22. 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).

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

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

25. 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).

26. 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).

27. 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).

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

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

30. 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).

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

32. 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).

33. 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).

34. Benestan, L. *et al.* Sex matters in massive parallel sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. *Mol. Ecol.* **26**, 6767–6783 (2017).

35. 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).

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

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

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

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

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

41. Lefouili, M. & Nam, K. The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. *Sci. Rep.* **12**, 11331 (2022).

42. 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).

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

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

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

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

47. 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).

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

49. Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R. & Portnoy, D. S. Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Mol. Ecol. Resour.* **17**, 955–965 (2017).

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

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

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

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

54. 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.

55. Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLOS ONE* **7**, e37558 (2012).

56. Warmuth, V. M. & Ellegren, H. Genotype-free estimation of allele frequencies reduces bias and improves demographic inference from RADSeq data. *Mol. Ecol. Resour.* **19**, 586–596 (2019).

57. 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).

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

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

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

61. 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).

62. 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).

63. 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).

64. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).

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

66. 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).

67. 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).

68. 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**, 233–240 (2016).

69. 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).

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

71. Arantes, L. S. *et al.* Scaling-up RADseq methods for large datasets of non-invasive samples: Lessons for library construction and data preprocessing. *Mol. Ecol. Resour.* **n/a**, (2023).

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

73. 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).

74. 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).

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

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

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

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

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

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

81. 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).

82. De Meeûs, T. Revisiting FIS, FST, Wahlund Effects, and Null Alleles. *J. Hered.* **109**, 446–456 (2018).

83. Lynch, M. & Walsh, B. *The Origins of Genome Architecture*. vol. 98 (Sinauer associates Sunderland, MA, 2007).

84. Lynch, M. & O’Hely, M. Captive breeding and the genetic fitness of natural populations. *Conserv. Genet.* **2**, 363–378 (2001).

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

86. 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).

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

88. 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).

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

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

91. 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).

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. 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).

94. 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.

95. 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).

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

97. Olofsson, D., Preußner, M., Kowar, A., Heyd, F. & Neumann, A. One pipeline to predict them all? On the prediction of alternative splicing from RNA-Seq data. *Biochem. Biophys. Res. Commun.* **653**, 31–37 (2023).

98. Upton, R. N. *et al.* Design, execution, and interpretation of plant RNA-seq analyses. *Front. Plant Sci.* **14**, (2023).

99. Rehn, J. *et al.* RaScALL: Rapid (Ra) screening (Sc) of RNA-seq data for prognostically significant genomic alterations in acute lymphoblastic leukaemia (ALL). *PLOS Genet.* **18**, e1010300 (2022).

100. Boshuizen, H. C. & te Beest, D. E. Pitfalls in the statistical analysis of microbiome amplicon sequencing data. *Mol. Ecol. Resour.* **23**, 539–548 (2023).

101. Combrink, L. *et al.* Best practice for wildlife gut microbiome research: A comprehensive review of methodology for 16S rRNA gene investigations. *Front. Microbiol.* **14**, (2023).

102. Cheng, Z. *et al.* Transcriptomic Analysis of Circulating Leukocytes Obtained during the Recovery from Clinical Mastitis Caused by Escherichia coli in Holstein Dairy Cows. *Animals* **12**, (2022).

103. Yang, L. & Chen, J. Benchmarking differential abundance analysis methods for correlated microbiome sequencing data. *Brief. Bioinform.* **24**, bbac607 (2023).

104. Patin, N. V. & Goodwin, K. D. Capturing marine microbiomes and environmental DNA: A field sampling guide. *Front. Microbiol.* **13**, (2023).

105. 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).

106. Deyneko, I. V. *et al.* Modeling and cleaning RNA-seq data significantly improve detection of differentially expressed genes. *BMC Bioinformatics* **23**, 488 (2022).

107. Zhao, M. *et al.* Exploring Conflicts in Whole Genome Phylogenetics: A Case Study Within Manakins (Aves: Pipridae). *Syst. Biol.* **72**, 161–178 (2023).

108. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 91 (2020).

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

110. 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).

111. 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).

112. 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).

113. 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).

114. 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).

115. 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).

116. 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).

117. 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).

118. 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).

119. 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).

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

121. 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).

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

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

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

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

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

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

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

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

130. Maruki, T., Ye, Z. & Lynch, M. Evolutionary Genomics of a Subdivided Species. *Mol. Biol. Evol.* **39**, msac152 (2022).

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

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

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

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

135. 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).

136. 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).

# **Figure Legends:**

**Figure 1:** Challenges and potential solutions related to filtering that occur prior to variant calling (“pre-variant” filtering). In the first row, individual 1 has both sufficient read depth (here illustrated with 9 reads, although a higher depth is preferred in practice) 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. Additional sequencing and 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 coverage 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, highly repetitive regions, 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 butterflies92 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 is much lower if performed across all samples (19-56%), the latter of which can obscure the differences in data quality between populations. Panel B uses data from a study on yellow perch78 to show that the number of SNPs varies 3-fold among populations if a MAF filter of 0.01 is applied within sample groups, which would be missed if the same filter was applied study-wide. Panel C illustrates how loci can be out of Hardy-Weinberg Proportions (HWP) due to homozygote or heterozygote excesses. We show that combining two divergent sample-groups results in higher *F*IS and many more loci out of HWP. Filtering study-wide would therefore cause the erroneous removal of loci due to Wahlund effects (see22). Lastly, in panel D, we examine linkage disequilibrium (LD) patterns with three-spined stickleback haplotypes128 that are highly correlated across individuals and illustrate that LD thinning (orange points) obscures an inversion (blue points).

**Figure 3:** Flow chart to facilitate thoughtful, systematic, and reproducible filtering for representative studies and questions utilising genomic DNA. Typical filtering proceeds through raw sequence QC filtering, alignment, mapped-read filtering, and variant discovery. After variant discovery, investigators must decide whether to apply filters study-wide or within sample-groups and whether to filter by locus or individuals first (see main text for recommendations). Regardless of the study objectives, multiple datasets should be constructed to examine the effects of various filtering decisions. Suggested filtering thresholds for per locus and individual filtering are provided for each question and objective (see also Table 1). Researchers should use a reproducible workflow to help them more easily repeat steps during analysis and the review process. Reproducible workflows can aid not only others in their lab or research group if more data will be produced in the future (*e.g.*, by students or postdocs) but also researchers in other labs. Data should be carefully archived before filtering, and all filtering methods and results carefully reported. See Supplementary Table 1 for a complete list of filters, Tables 2-3for a simplified example of how to report filtering results, Box 3 for a filtering and reporting checklist, and the pre- and post-variant filtering R notebooks for example reproducible workflows (Notebook S1-2).

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. Most alignment tools will return map quality (“mapQ”) scores, the derivation and meaning of which varies by program. Filtering thresholds based on this score must consider the specific aligner used.

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.

CONTIG

A contiguous sequence of DNA assembled from many overlapping sequence reads, representing a fragment of a chromosome.

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, bases, 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 specifically focuses on the *removal* (or “filtering out”) of data 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.

HAPLOTYPE PHASE

The complete sequence of variants that occur along a single homologous chromosome.

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 or samples. Imputation can use reference panels of well-described haplotypes to improve performance when available, usually in well-studied model organisms.

LINKAGE DISEQUILIBRIUM (LD)

Non-random association of alleles at different loci within a population or sample group. Also known as gametic phase disequilibrium. This association can be caused either by physical linkage, when alleles are co-inherited due to non-independent assortment caused by close physical proximity, or can occur across chromosomes when inbreeding, paralogy, genetic drift or other factors make 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 across 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 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 duplication of 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.

PHASING

The reconstruction or inference of two distinct sequences of variants which co-occur on the same chromosome or DNA strand.

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 limited migration and geneflow including spatial, temporal, or behavioral separation. Population structure is characterized by divergent allele frequencies across loci.

READ DEPTH

The number of reads which cover a given or fixed genomic position. Also referred to as “coverage”.

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 GENOME

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.

REFERENCE BIAS

The propensity for reads containing the non-reference allele (the allele not noted in the reference genome) to have lower mapping quality scores or map to the wrong location than those containing allele noted in the reference genome.

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. Often presented as a genome-wide average.

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 spectrum which describe the frequency distribution of derived alleles or a “folded” or “unpolarized” minorallele frequency spectrum which describes the frequency distribution of the minor alleles. Also known as the allele frequency distribution. Can provide evidence for different population demographic histories.

STRUCTURAL VARIANTS

Genetic variants which are caused by underlying variation in the order, number, and/or arrangement of loci.

STUDY-WIDE FILTERING

Applying a filtering threshold globally (simultaneously across all samples in the entire dataset without respect to their provenance or treatment) rather than separately within each sample group.

VCF FILE

A file in the Variant Call Format, which contains genotype calls (or likelihoods, posteriors, etc) alongside a flexible suite of metadata such as filtering and processing history and quality information.

WAHLUND EFFECT

A reduction in observed homozygosity () 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 HWP (causing an elevated *F*IS).

WITHIN-GROUP FILTERING

Applying a filtering threshold within each sample group separately rather than across all individuals simultaneously (*e.g.,* study-wide or globally).

WHOLE-GENOME SEQUENCING

Sequencing the entire genome without any attempt to target specific regions (in contrast to reduced-representation sequencing). Usually entails “shotgun” sequencing and results 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 sample collections, populations, or statistical methods. We used ten empirical datasets (*Arabidopsis*129, *Daphnia*130, White-tailed deer131, Mountain goats132, Humans, Deer mice133, Killer whales, Monarch butterflies92, Stoneflies134, and Yellow perch78 – see Supplementary Methods and Supplementary Table 2) along with three simulated datasets (see Supplementary Methods and Supplementary Table 3) to demonstrate how filters can impact a wide range of commonly calculated population-genetic parameters. We first illustrate the effects of varying a single filter; minor allele frequency (MAF) on four commonly used parameters (panel A). Given the differences in scales across studies, parameter estimates were standardized such that the values shown represent the change in a statistic as filtering thresholds increase in stringency (See Figures S4, S6, S8, S10, and S12 for absolute values). We also show how multiple filters (MAF, HWP, and missing data) and the higher-order interaction of MAF and HWP influence a single parameter (the number of segregating sites, panel B). A wide range of additional filtering effects are also presented in the Supplementary Information (Figures S1-S17).

Changing the filtering threshold for a single filter can result in a large change in *F*ST, *F*IS, HO/H­e, and Tajima’s D estimates. Increasing MAF thresholds reliably increases average *F*ST, HO (per segregating site), and Tajima’s D (see also60,71). *F*IS, however, is impacted differently and unpredictably among datasets (panel A). Filtering with the “standard” (most commonly used) MAF threshold of 0.05 can often flip the sign of Tajima’s D from negative to positive, thus completely changing its interpretation from a population expansion to a bottleneck (Panel A, top-right, diamonds indicate change in sign; see also Figure S4). MAF filtering is also of particular note in that it can dramatically change effective population size (*Ne*) estimates derived from LD-based approaches113 ­(see Figure S3).

Varying filtering thresholds across multiple filters also results in substantial changes to a single parameter (panel B). The number of segregating sites is, predictably, always reduced by stringent filtering, although the degree to which that occurs varies across datasets and populations. MAF filters, which are perhaps the most widely used, can have a particularly strong effect that, while constant in direction, can widely vary in magnitude depending, primarily, upon the shape of the underlying site frequency spectra across populations. HWP and missing data filtering (here “Required % Genotyped” such that any loci or individuals with less than the noted percentage of called genotypes are removed) also impact different datasets in different ways–datasets where more loci are out of HWP lose more segregating sites with higher HWP filters, as do those with more missing data. Higher-order interactions between filters may also be important—substantial change in the average impact of HWP filtering on the proportion of retained segregating depending on the MAF filter used (panel B, right). Beyond this, higher-order interactions among filters have still not been explored to any substantial degree and are worthy of more study.

The demographic context of the study system (including model species) and their underlying genomic architectures can also affect filtering. Populations which have undergone recent population expansions, for example, will lose far more rare alleles during MAF filtering than will those which have undergone population bottlenecks (Figure S1). The removal of regions of elevated F­ST caused by selection occurring in areas with reduced recombination rates will have an larger estimate on genome-wide PCA results than if those regions were in areas with larger recombination rates (Figure S2). Many parameters such as *F*ST, *F*IS, and LD will also be influenced by positional factors such as the presence of conserved exons, introns, centromeres, telomeres, or sex-linked loci.

A group of graphs showing different colors

Description automatically generated with medium confidence

**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 to equal the proportion (or percentage) of correctly called genotypes that are retained in a dataset. Within this framework, a false-positive occurs when an incorrectly called genotype (at a single locus) is retained within a dataset and a false-negative occurs when a correctly called genotype is incorrectly filtered out of the dataset (see diagram below)135. False-positives occur most frequently when filters are not stringent (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 stringent filters are used (such as a high MAF filter), because 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). This procedure, however, will invariably allow more false-positives (false variants) into the dataset affecting the precision of the estimator. Alternatively, low-frequency sites are often removed when performing GWAS or outlier *FST* detection79,136, which creates datasets 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.

A diagram of a scientific experiment

Description automatically generated

**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, knowledge of the system, and population genetics theory (see Figure 3) |
|  | Create filter recording/reporting tables (see Table 1 and Table 2) |
|  | Filter on raw sequences (for example, 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 by filter type |
|  | 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 (some variant callers will do so automatically if these are marked) |
|  | 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. |
|  | Report the # SNPs remaining after these steps to understand which steps remove the most loci |
|  | Individual filtering |
|  | Missing data; mislabelling/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 (see Tables 12) |
|  | 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 |

**Table 1**: Threshold recommendations for low and high stringency datasets used to assess the impact of filtering decisions for different questions and objectives (see Figure 3, main text). Please note that these values are *suggestions* and should not necessarily supplant any existing knowledge of the study system or details of the base dataset. These values represent suggested extreme values to examine effects; moderate values may be preferable for final analysis. The suggested MAC values may actually be stricter than the suggested MAF values for datasets with low sample sizes and should be adjusted accordingly. Furthermore, some methods for questions may require different filter values that differ from these suggestions. See Supplementary Table 4 for brief justifications for many of these suggestions. Columns: 1) Question/Objective: Objective of study or the question being asked; 2) MAF/MAC: Minor allele frequency or minor allele count filtering threshold below which loci are removed; 3) Individuals missing data: required % genotyped: individuals called at less than the threshold percentage of loci are removed; 4) Loci missing data: required % genotyped: loci called at less than the threshold percentage of individuals are removed; 5) HWP: loci with a p-value for Hardy-Weinberg proportions less than the threshold are removed; 6) LD: should loci be purged to reduce the degree of linkage disequilibrium between pairs of loci.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Question/Objective** | **MAF/MAC** | **Individuals missing data: required % genotyped** | **Loci missing data: required % genotyped** | **HWP** | **LD** |
| Population Structure | MAC > 1;  MAF > 0.05 | 50; 95% | 50; 95% | 1e-6; 0.01 | No; Yes |
| Demography | MAC > 0;  MAC = 1 | 75; 95% | 50; 95% | 1e-6; 0.01 | Yes |
| Relatedness/  Pedigree Construction | MAC > 2;  MAF > 0.05 | >5%, provided sufficient remaining loci for power | 80%; 95% | 0.001; 0.01 | No, although CI’s can be affected |
| Selection | MAC > 1;  MAF > 0.05 | 25; 75% | 25; 75% | 1e-6; 0.001 | No |
| Genetic Diversity | MAC > 1;  MAF > 0.05 | 50; 75% | 50; 75% | 1e-6; 0.01 | No, usually |
| Phylogenetics | MAC > 1;  MAF > 0.05 | 50; 80% | 50; 80% | 0.001; 0.01 | Yes |
| GWAS | MAF > 0.05, unless large *n* | 50; 80% | 50; 80% | 1e-6; 0.01 | No, but correct p-values |
| Mutation Detection | No filter | Parents: 100%,  Offspring: 0% | Parents: 100%,  Offspring: 0% | 1e-6 | No |
| Metagenomics/eDNA | MAC > 4 | - | - | - | - |

**Table 2:** Representative table demonstrating pre-variant filtering reporting standards. All accompanying code and filtering steps should be reported in the Prevariant Notebook; See Figure 3 for a detailed Flow Chart and Supplementary Notebook 1 for the accompanying Prevariant Notebook. The sequence of filtering events can affect downstream results, so rows should be arranged chronologically.

|  |  |  |
| --- | --- | --- |
| **Filtering Step** | **Results to Report** | **Values** |
| Sample selection and Sequencing | Number of individuals/samples collected | *N*=250 individuals |
| Sample selection and Sequencing | Number of samples initially sequenced | *N*=210 individuals |
| Sequence QC | Number of individuals that were successfully sequenced\* | *N*=200 individuals |
| Sequence QC | Total number of reads prior to any filtering\* | 1.5 x 108 reads |
| Sequence QC | Number of reads remaining after filtering for read quality\*\* | 0.8 x 108 reads |
| Mapped read filtering | Number of reads that mapped | 6.8 x 107 reads |
| Mapped read filtering | Number of reads remaining after filtering for mapping quality | 5.1 x 107 reads |
| Mapped read filtering | Number of reads remaining after filtering for improperly paired reads | 4.8 x 107 reads |
| Mapped read filtering | Number of reads remaining after filtering for PCR duplicates\*\*\* | * 1. x 106 reads |

\*The parameters and parameter values used to characterize “success” should be clearly described. Samples filtered out at this stage typically include those with large deviations from the average number of reads per individual and/or a large percentage of reads with low base-quality scores.

\*\*Any additional filters used during sequence QC, such as filtering for poly-G tails or adapter mismatches should be given their own row.

\*\*\*Any additional filters used during filtering of mapped reads should be given their own row. Notice in this heuristic example that alignment (number of reads that mapped) and filtering for PCR duplicates had large effects. By using a table such as the one presented here, readers will be able to determine which filtering steps had the largest effects.

**Table 3:** Representative table demonstrating post-variant reporting standards. All accompanying code and filtering steps should be reported in the Postvariant Notebook; See Figure 3 for a detailed Flow Chart and Supplementary Notebook 2 for an example Post-variant Notebook. The sequence of filtering events can affect downstream results, so rows should be arranged chronologically. All filters should be recorded as a separate row; even if a particular filter is not mentioned or default values are used for that type of filter. Depending on the objectives of the study, different numbers of datasets may need to be created. In this example, we assume a targeted read depth of 20x coverage per individual and 200 sequenced individuals (following from Table 1).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Filtering Step** | **Dataset 1** | **Dataset 2** | **Dataset 3** | **Dataset 4** |
| Objective | Population structure | GWAS-stringent | GWAS-relaxed | Demography |
| Minimum Genotype Quality | 40 | 50 | 40 | 40 |
| Minimum Genotype Coverage | 10x | 20x | 15x | 10x |
| Maximum Genotype Coverage | 30x | 30x | 50x | 30x |
| Study-wide or within sample-group | Study-wide | Study-wide | Within sample-group | Within sample-group |
| Maximum missing per individual | 15% (190/200 individuals retained) | 5% (180/200 individuals retained) | 10% (188/200 individuals retained) | 15% (190/200 individuals retained) |
| Maximum missing per locus\*\* | 15% | 5% | 10% | 5% |
| MAF/MAC | MAF=0.05 | MAF=0.05 | MAF=0.01 | MAC=2 |
| Linkage Disequilibrium | No filter used | No filter used | No filter used | *r2 >= 0.25* |
| Hardy-Weinberg deviations | P < 0.05; Bonferroni\* | P < 0.05; Bonferroni\* | No filter used | P < 0.05; no FDR correction |
| Results | Figure 1b-c | Figure 2, Figure 3 | Supplementary Fig. 1 | Figure 4 |
| VCF md5sum | 8d3d627940ee2a77b4770db1fd710459 | 3dcccbf8d3fb869c3cf5de291c0fe893 | 0b0681ad8b5bdab39e7b76afc190d4c8 | 1f131fdc2ee6444e1b94071195a1acd2 |

\*Hardy-Weinberg filters should only be applied within each sample group, not study-wide; corrections for multiple comparisons should be reported

\*\* Missing data should be examined both study-wide and within sample-groups; different sample groups may contain different amounts of missing data