nature reviews genetics

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

Review article

Check for updates

Next-generation data filtering in the genomics era

William Hemstrom [1,4] ✉, Jared A. Grummer[2,4], Gordon Luikart[2] & Mark R. Christie [1,3] ✉

Abstract

Genomic data are ubiquitous across disciplines, from agriculture to biodiversity, ecology, evolution and human health. However, these datasets often contain noise or errors and are missing information that can affect the accuracy and reliability of subsequent computational analyses and conclusions. A key step in genomic data analysis is filtering — removing sequencing bases, reads, genetic variants and/or individuals from a dataset — to improve data quality for downstream analyses. Researchers are confronted with a multitude of choices when filtering genomic data; they must choose which filters to apply and select appropriate thresholds. To help usher in the next generation of genomic data filtering, we review and suggest best practices to improve the implementation, reproducibility and reporting standards for filter types and thresholds commonly applied to genomic datasets. We focus mainly on filters for minor allele frequency, missing data per individual or per locus, linkage disequilibrium and Hardy–Weinberg deviations. Using simulated and empirical datasets, we illustrate the large effects of different filtering thresholds on common population genetics statistics, such as Tajima's D value, population differentiation (F_{ST}), nucleotide diversity (π) and effective population size (N_e).

Sections

Introduction

Common filters: complexity and importance

Effects of filtering

Solutions and best practices

Conclusions and future directions

[1]Department of Biological Sciences, Purdue University, 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, West Lafayette, IN, USA. [4]These authors contributed equally: William Hemstrom, Jared A. Grummer. ✉e-mail: whemstro@purdue.edu; christ99@purdue.edu

Introduction

Rapid advances in both short-read and long-read sequencing technologies have resulted in the proliferation of large genomic datasets[1–4]. All high-throughput ('next-generation') sequencing methods yield large numbers of DNA sequences, known as reads, which have variable error rates[5,6]. In addition to the inherent errors introduced from sequencing, such as biases introduced during library preparation, polymerase errors during DNA replication or inaccurate base calling, errors can arise when sequences are aligned to a reference genome or transcriptome, particularly if the reference is incomplete, highly divergent or assembled de novo from the reads themselves[7–9]. Therefore, investigators must perform multiple types of data filtering for quality control (QC) prior to data analysis. Here, we define filtering as the intentional removal of sequencing bases or reads before genetic variant calling (pre-variant filtering) or the removal of genetic variants, genotypes and/or individuals from a genomic dataset after variant calling (post-variant filtering), with the explicit goal of improving data quality for downstream analyses. Pre-variant and post-variant filtering generally coincide with pre-VCF file and post-VCF file stages (or pre-genotyping and post-genotyping).

Pre-variant filtering includes filtering according to read quality, mapping quality and read depth. Filters on read quality assess the reliability of each base call within a sequencing read; mapping quality scores give an indication of the strength and uniqueness of read alignment to a single location in a reference genome; and thresholds for read depth (also known as coverage) define the minimum number of reads required to cover a genomic position. Post-variant filtering includes minor allele frequency (MAF) and minor allele count (MAC) filters, which remove variants based on their allele frequencies within a population, and missing data filters, which filter out loci and/or individuals with a user-defined proportion of missing calls. Other post-variant filtering approaches remove variants that substantially deviate from Hardy–Weinberg proportions (HWP) or linkage disequilibrium (LD).

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 to understand effects of filtering on downstream conclusions. 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 data removal. Filtering is an issue of paramount importance because every genomic dataset must be filtered, often repeatedly, and the same dataset filtered in different ways can yield entirely different results[12,13]. Furthermore, filtering choices can be confusing and subjective, currently lack consistent, agreed-upon guidelines and may result in unknown downstream consequences[14].

Filtering approaches vary widely among published studies, and the filters used are often not described or are done so inadequately. When specific filters are mentioned, the methodological details provided are often insufficient for reproduction. Among papers that sufficiently record filtering thresholds, the values used for a given filter can vary several-fold[15], even after accounting for sequencing depth and sample sizes. Furthermore, researchers often use default program settings, which is problematic given that default settings can sometimes be inappropriate and lead to problems, including the removal of deviations from HWP that are important for understanding population structure[11]. To help investigators improve their filtering approaches, we review best practices for filtering genomic datasets and illustrate the downstream effects of a range of filtering decisions on both empirical and simulated datasets. We provide practical filtering threshold recommendations and an extensive suite of resources to facilitate better filtering. We also stress the importance of knowing your study system and population genetics theory[16–18], both of which will help researchers to make informed choices for setting filtering thresholds and facilitate the interpretation of different results inevitably produced with different filtering thresholds.

We recognize that some alignment-free methods exist that use high-throughput sequencing data, particularly for metagenomics and phylogenomic analyses[19], and these are not considered extensively here. Similarly, filtering for RNA sequencing data is not thoroughly reviewed; although many concepts hold true (especially for single-nucleotide polymorphisms (SNPs) called from RNA sequencing data), many of the filters considered here may not apply[20,21]. Thus, we do not provide extensive, specific filtering guidelines for RNA sequencing, microbiomes, environmental DNA (eDNA) and metagenomic or metabarcoding datasets, as these guidelines are provided elsewhere[22–33]. Similarly, some analytical methods have specific filtering requirements, such as phylogenetic reconstruction (which at larger divergence scales often requires alignments across divergent species)[34] and germline-specific medical genetics approaches (which may use trio sequencing of parents–offspring, combine results from multiple variant callers and use benchmarking datasets with known, 'ground truth' sequence variants to improve data quality)[35], which are not covered here.

Common filters: complexity and importance

Today, investigators have many different options for obtaining DNA, RNA or epigenome sequencing data[36,37], which yield reads of different lengths, quality and configurations (for example, single-end or paired-end)[38]. 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 (that is, contigs or stacks of reads) for subsequent genotyping or variant calling analyses[39–42]. Without some form of reference-guided alignment, some downstream analyses can be limited as it can be difficult to estimate relevant parameters (for example, runs of homozygosity[43] or recombination points) or determine the genomic context for loci of interest (for example, linkage or haplotype phase)[44]. Lacking a quality reference genome can therefore be a challenge for researchers working in non-model systems where de novo alignment, alignment to a low-quality reference genome assembly (for example, many contigs and low N50 or L50 scores)[33] or alignment to a different, distantly related species[45,46] is required (but see ref. 47). The errors introduced throughout these steps must be accommodated, usually through filtering, to minimize downstream biases and maximize 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 principal component analysis (PCA), can be indicative of experimental or laboratory errors (for example, mislabelling), sequencing bias, sex-linked loci, selection or other phenomena[48–51], thereby suggesting 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 complete and researchers are confident that their filtering choices have not systematically influenced any conclusions. A comprehensive list of filters, their descriptions and the typical genomic workflow stage at which they are applied can be found in Supplementary Table 1. Below, we discuss commonly applied

pre-variant filters, including read/mapping quality and read depth, and post-variant filters, including MAF or MAC, missing data and deviations from HWP or LD.

Pre-variant filtering

Read or mapping quality. Prior to de novo assembly or alignment to a reference, data are usually filtered via the removal of low-quality reads (Supplementary Table 1). This initial filtering can have downstream effects, and researchers across disciplines often use different and, potentially, arbitrary definitions of what constitutes a low-quality read. For example, a wide range of read quality thresholds are used, ranging from a Phred-scaled base quality score of 5 (refs. 52,53) to 40 (ref. 54) (representing between ~32% 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 per read also 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 thresholds[55–58]. Conversely, high-quality reads may also be excluded if the reference does not contain the sequence, they are highly repetitive (as with transposable elements) or they are found in multiple genomic locations (for example, paralogues) (Fig. 1a,b).

To aid researchers in understanding the potential effects of these filtering decisions, investigators should always report the percentage of reads that were removed prior to alignment and 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 (such as a read length hard filter or a soft filter based on a statistical model)[57]. Reporting these statistics can help reviewers to assess the quality of the data underlying the study and allow future investigators to determine whether alternative filtering strategies could address additional questions (for example, re-filtering to not remove reads that map to multiple locations could identify paralogous loci or transposable elements). Of note, even strong alignments of putatively high-quality reads are not always correct, as reference bias[59,60], genome assembly errors[61], structural variation[62] such as copy number variants (CNVs)[63,64] and challenging alignments (for example, transposable elements[65] or PCR duplicates[66]) are present in most genomic datasets (Fig. 1).

Read depth or coverage. After initial pre-variant filtering, genomic workflows typically proceed with genotyping (Box 1), whereby genetic variants such as SNPs, insertions or deletions (indels) and structural variants are algorithmically identified with software such as GATK[67], ANGSD[68], STACKS[9], ipyrad[69], LUMPY[70], BCFtools[71] or others. During this process, the read depth (coverage) of each locus must be considered, as greater depth of coverage usually allows for more confidence in genotyping (and subsequently downstream inferences) (Fig. 1c). However, very high read depths (relative to the study-wide mean) can be indicative of paralogues, highly repetitive regions, CNVs, non-target DNA (for example, mitochondrial DNA) or technical (for example, PCR) duplicates (Fig. 1d and Supplementary Table 1). Variant calling algorithms typically mark genotypes as missing either if they are below or above certain read depths or if they have poor quality scores[72]. The depth and/or quality filters used at this step vary substantially between studies. However, filtering out loci sequenced at a low depth is not without risk given that calling heterozygotes requires higher depth than homozygotes and that stringent depth filtering can skew observed heterozygosities and, therefore, site-frequency spectra (SFS)[73,74]. Well-developed approaches to make use of low-coverage sites and mitigate such biases do exist[68,74], so filtering out such loci is not always necessary. Note that although many of the principles we cover here still apply, low-coverage whole-genome sequencing (WGS) data have their own filtering specificities[39].

Post-variant filtering

Missing data. Missing data can result either from the absence of reads covering a locus in an individual or from upstream filtering on read or genotype quality, depth, mapping or other filters (Fig. 2a). 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 that something went awry with sample collection or preservation, genomic library preparation or alignment, all of which can obscure patterns of potentially important variation[75]. The filtering choices used for missing data vary widely among studies, and the downstream consequences are rarely evaluated. The acceptable amount of missing data depends on the research question: for example, a lower amount of missing data might be allowed for some phylogenetic analyses for which missing data can be highly problematic and that require relatively high-quality sequence data for all individuals[14]. By contrast, if maximizing the number of loci is a priority, researchers might choose to keep loci with more missing data and acknowledge that filtering threshold choices might impact the types of loci retained[76].

Missing data can have serious effects on downstream biological conclusions. For example, missing data due to low sequencing coverage may hinder the detection of runs of homozygosity and, subsequently, downwardly bias estimates of individual inbreeding[77]. Recent work has suggested that declines in coverage across studies may have resulted in underestimates of inbreeding in some North American wolf populations[78] (Box 1). Missing data levels can also bias estimates of genome-wide heterozygosity in either predictable[79] or unpredictable[80] ways. Other methodologies seem to be more resistant to missing data; for example, one study found that the proportion of missing data did not seem to affect either gene flow or parentage results in a tropical plant[14].

MAF and MAC. Loci (typically SNPs) for which the less frequent allele (that is, the minor allele) occurs below a certain frequency are also often filtered out (Fig. 2b). MAF filtering is often based on the assumption that singletons or other rare variants that occur at a frequency of less than ~5% in a sample-group are due to genotyping errors. MAF filtering is often performed at the specific threshold of 0.05 to reflect this, although threshold values across published studies can vary by orders of magnitude. Depending on the analysis and objectives, this filter can be applied study-wide (for example, globally across populations) or separately within each sample-group.

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 sample-groups 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 3 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 6 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 can be particularly useful when sample sizes are small or highly variable because a typical MAF filter (for example, 0.05) will never

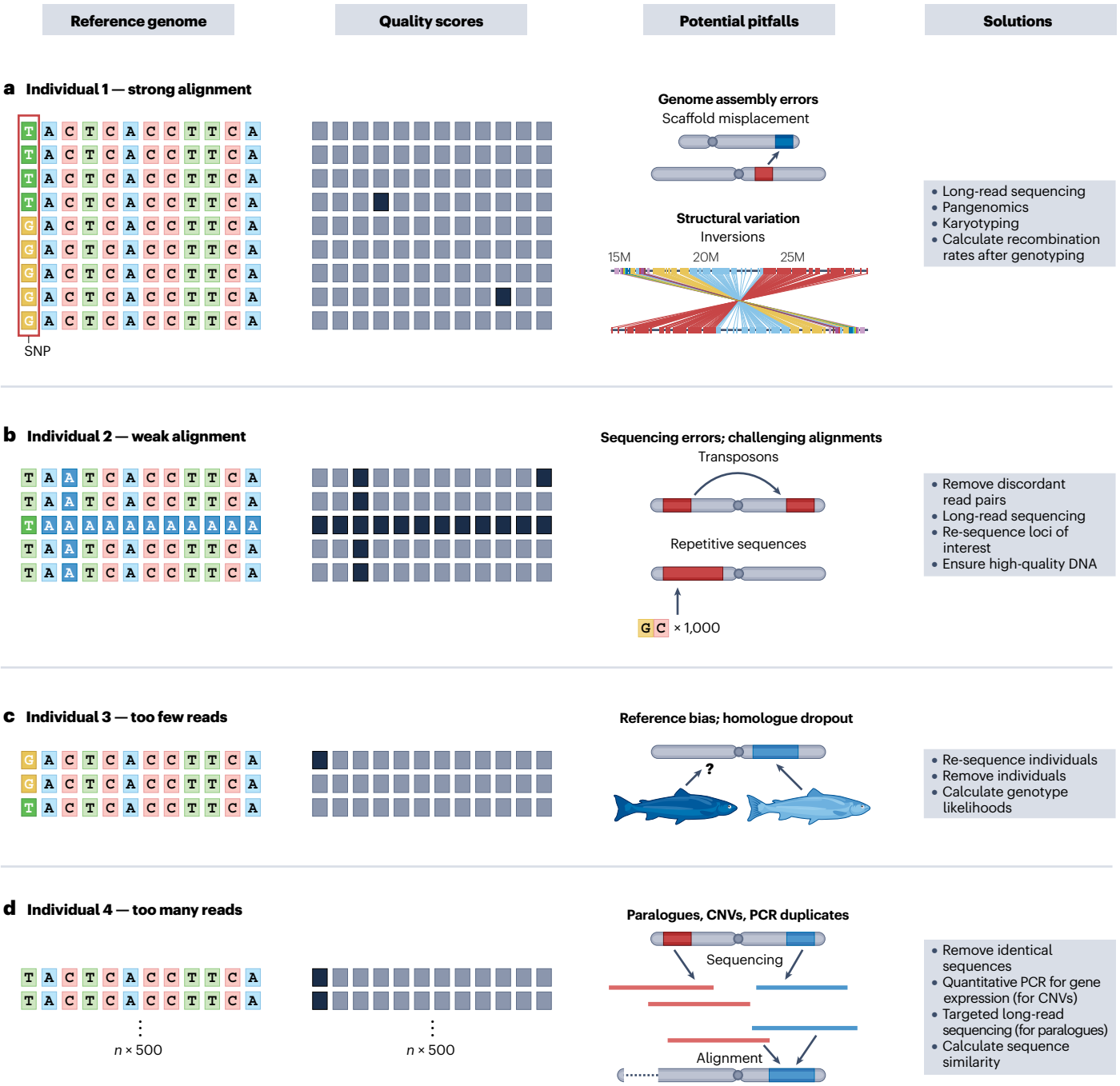


Fig. 1 | Pre-variant filtering — challenges and potential solutions related to filtering before variant calling. a, Individual 1 has both sufficient read depth (illustrated with nine reads, although a higher depth is preferred in practice) and read alignment quality ('quality scores'; grey indicates high quality, black indicates poor quality) to allow for successful variant calling (a single-nucleotide polymorphism (SNP) is highlighted with a red box). Despite this seemingly successful read alignment, challenges still exist: genome assembly errors (such as misplaced scaffolds on reference genomes), and structural variation (such as inversions) can cause issues for downstream analyses. Additional sequencing and filtering for linkage disequilibrium (LD) can resolve some of these concerns. **b**, 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. Weak alignments can also occur in repetitive regions of the genome or with short target sequences (for example, transposons); solutions include specialized filtering and long-read sequencing. **c**, 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, low-coverage approaches should be employed. Reference bias (for example, aligning to reference genomes from different species than the sequenced samples, including cryptic species) can also cause fewer reads to align than expected; solutions include removing samples or individuals and controlling for read depth across sample-groups. **d**, Individual 4 has too many reads, which can be caused by paralogous genes, highly repetitive regions, copy number variants (CNVs) or technical (for example, PCR) duplicates. These excess reads could be filtered out or carefully analysed to determine the underlying causes.

remove variable loci (including singletons) in small samples ($n < 10$ diploid individuals).

Hardy–Weinberg proportions. It is often desirable to filter out loci based on statistically significant (for a given α-value or P value) deviations from HWP. HWP are a common assumption of many downstream analytical tools (for example, STRUCTURE)[81], and removing loci that violate HWP can help to ensure unbiased results for downstream analyses in randomly mating populations[82]. Deviations from HWP often reflect sequencing, assembly or alignment errors (such as a heterozygote deficit caused by allelic dropout or a heterozygote excess caused by paralogous regions)[47,60,83,84]. However, loci out of HWP can also indicate real biological phenomena, such as cryptic population substructuring (Fig. 2c) or balancing selection[85]. As a result, it is crucial to filter HWP within sample-groups (for example, within populations) rather than study wide (for example, globally on all samples)[86] (discussed below) and to do so with a low stringency if the loci under selection or those that differ between populations are of interest. That said, some metrics, such as F_{ST}, can be biased upward by the careless removal of loci that are not in HWP within populations[86], which is potentially problematic if population delineations do not reflect biological realities (for example, in the case of demes). Tools such as HDplot[84] or ngsParalog[87] might be useful in such cases to identify and remove loci that are more likely to be out of HWP due to paralogy rather than relying on divergence from HWP alone. Given that the F_{IS} is fundamentally a directional measurement of HWP divergence, removing loci that are out of HWP with either positive or negative F_{IS} values (and thus either heterozygote deficits or excesses, respectively) may help to remove loci out of HWP by paralogy or other factors specifically.

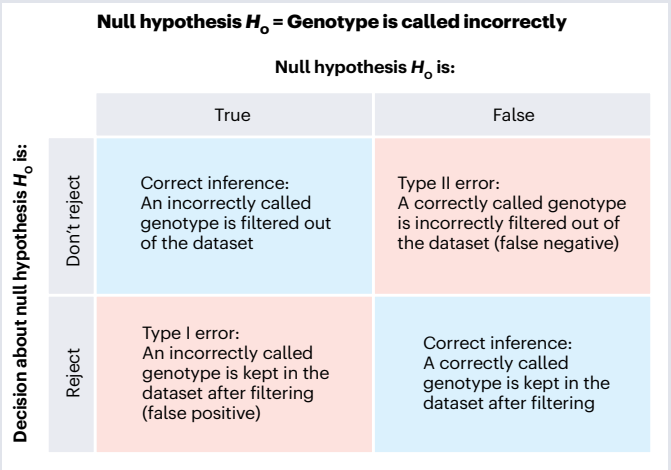
The thresholds used for testing HWP differ from the other filters because HWP testing is hypothesis-testing, and, as such, produces P values upon which filtering thresholds are set. Because thousands of loci may be tested simultaneously, corrections for multiple testing should be considered to adjust P values and thus avoid unnecessarily removing large numbers of quality (non-problematic) loci; however, such corrections are seldom performed[15]. That said, P-value correction for HWP differs from typical testing approaches in that, from a broader perspective, uncorrected P values are conservative in that they will result in the removal of more potentially problematic loci, not fewer. Researchers who do correct for multiple tests should explicitly report the reasoning behind the correction method they use, which

Box 1 | Filtering trade-offs

Different filtering choices result in trade-offs described here as false positives (type I or α-errors) and false negatives (type II or β-errors) (see the figure). During variant calling (that is, genotyping), an incorrectly called genotype is the null hypothesis. This simultaneously allows for a more conservative philosophical approach towards genotyping and allows for power ($1 - \beta$) to equal the proportion (or percentage) of correctly called genotypes retained in a dataset. Within this framework, a false positive occurs when an incorrectly called genotype (at a single locus) is retained and a false negative occurs when a correctly called genotype is incorrectly filtered out[142] (see the figure). False positives occur most frequently when filters are not stringent (for example, no minor allele frequency (MAF) filtering is performed) and/or when read depths at a locus are low. By contrast, false negatives are more likely when stringent filters are used (such as a high MAF filter), because more loci are assumed to be erroneous (and thus removed) even though many of those sites may represent real, correctly called genotypes.

Trade-offs between the two filtering error rates are inevitable for certain methods and questions. For example, when calculating Tajima's D value, many or most low-frequency variants (for example, singletons) should often be retained (Box 2); however, this will invariably allow more false positives (false variants) into the dataset and negatively affect accuracy or precision. Alternatively, low-frequency sites are often removed when performing genome-wide association studies (GWAS) or testing for outlier F_{ST} loci[103,143], which creates datasets with few false positives but may exclude real, causal variants that segregate at low frequency. Investigators should be cognizant of filtering trade-offs and consider solutions, such as creating two or more datasets with different filtering thresholds, sequencing loci of interest to higher depths and/or re-sequencing select samples.

Filtering trade-offs and challenges also arise when using low-coverage versus high-coverage data (or both). For example,



a range of relatively low-coverage restriction-site-associated DNA (RAD) sequencing data (average <8× read depth) was used to estimate the effective population size (N_e), heterozygosity and individual inbreeding in North American wolves between 1991 and 2020, during which heterozygosity and N_e seemed to decline in some populations[144]. However, samples were also sequenced to a lower depth over time, causing the proportion of missing genotypes to increase; reanalysis suggested that the reductions in sequencing depth likely caused the lower estimates of heterozygosity and N_e rather than actual inbreeding or reduced variation in recent decades[78]. Following the reanalysis, the authors recommended aggressive read-depth filtering of loci to retain more individuals given that larger numbers of individuals are more beneficial than additional loci when estimating contemporary N_e with genomic data.

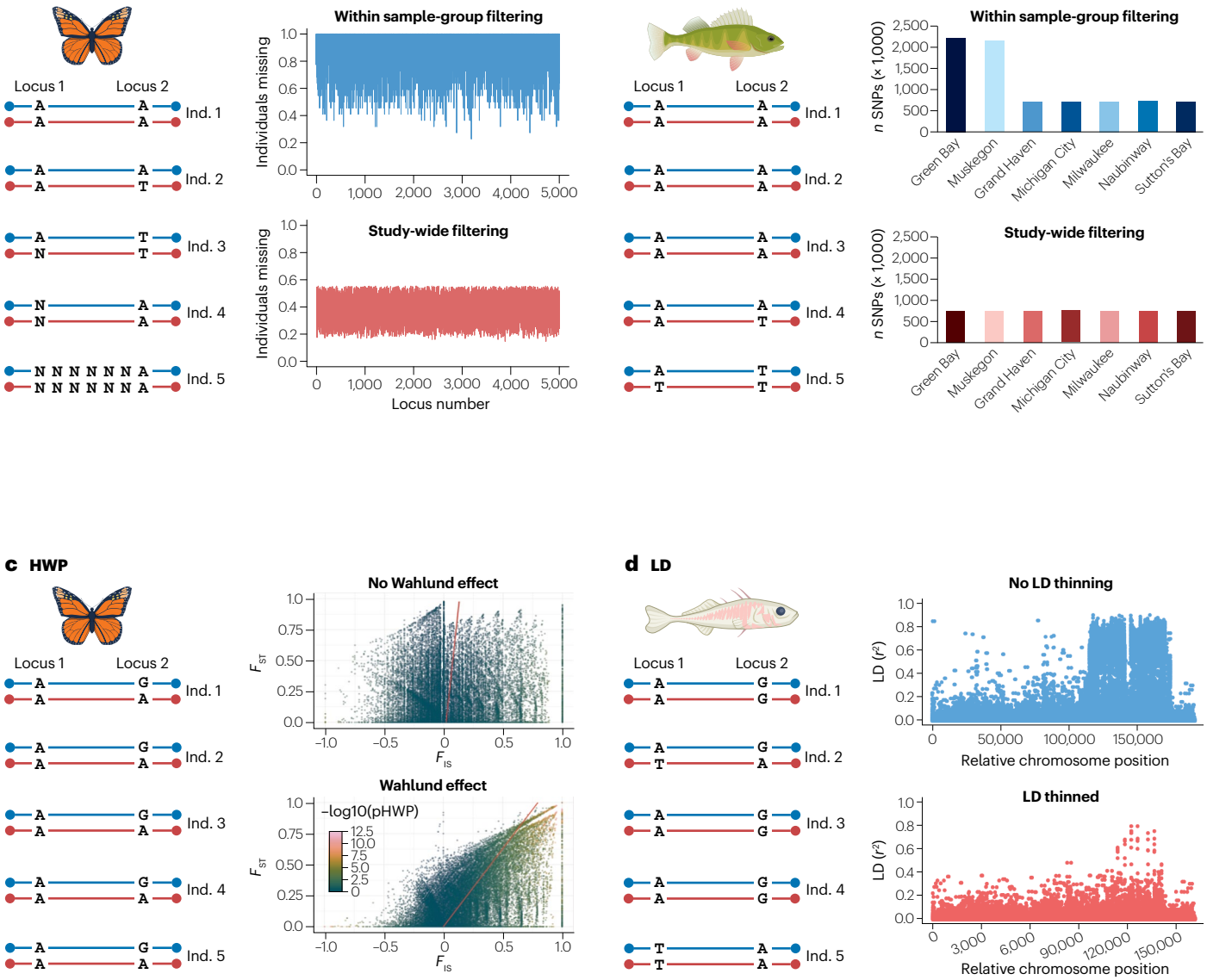


Fig. 2 | Post-variant filtering — challenges associated with four common filters after variant discovery. a, Missing data, which can occur across loci and individuals. Data from monarch butterflies[114] are used to show that the percentage of missing data can be high (21–100% per locus) when missing data filtering occurs within sample-groups, but is much lower if performed jointly across all samples (19–56%), the latter of which can obscure the differences in data quality among populations. **b**, Data from a study on yellow perch[102] showing that the number of single-nucleotide polymorphisms (SNPs) varies threefold among populations if a minor allele frequency (MAF) filter of 0.01 is applied within sample-groups, which would be missed if the same filter was applied study-wide. **c**, Loci can deviate from Hardy–Weinberg proportions (HWP) due to homozygote or heterozygote excesses owing to biological causes, which filtering may obscure. For example, unintentionally combining two divergent sample-groups will result in higher F_{IS} (fixation index, individuals versus subpopulation) and many more loci out of HWP (owing to Wahlund effects)[39]. Filtering study-wide would cause the erroneous removal of loci in such cases. **d**, Linkage disequilibrium (LD) thinning (that is, filtering out loci with high LD; red points) can obscure an inversion (blue points) in three-spined stickleback haplotypes[141]. F_{ST}, fixation index, subpopulation versus total; MAC, minor allele count.

can vary in stringency (from Benjamini–Hochberg[88] to sequential or simple/stringent Bonferroni[89]), and, ultimately, different approaches and α-thresholds should be applied depending on the questions being asked and the tolerance for including or excluding problematic loci[16] (Box 1). For an in-depth discussion on implementing and interpreting tests for HWP, see refs. 16,82,90.

Linkage disequilibrium. Pruning sets 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 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 such as the effective population size (N_e)[91]. 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 remove or overlook important structural variants (Fig. 2d).

Studies that lack a high-quality reference should also employ LD filters to ensure independence of loci or contigs[92], 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 (although this assumes distinct stacks or contigs are themselves unlinked, which may not be true and should not be the 'default assumption'). Corrections for multiple tests are also important for LD filtering if *P* values are used as a linkage measure, but such corrections are seldom conducted or reported[15].

Effects of filtering

The effects of filtering are often unknown in genomic studies. Although 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). For example, mapping quality filters can be applied both immediately after mapping and later during genotype calling. 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. An excellent example of this are GATK's 'hard filters' on called genotypes (for example, for strand bias or variant positions within reads)[67], which are routinely used but seldom discussed or varied from their recommended values. However, properly considering filtering choices and their effects is crucial because different filtering choices can lead to vastly different downstream results such that two researchers who make different decisions but analyse the same original dataset (for example, a set of FASTQ or VCF files) could reach entirely different biological conclusions.

To illustrate this principle, we systematically filtered ten published empirical and three simulated datasets by changing filtering thresholds for three key post-variant filters: MAF, missing data and HWP (Box 2). Although MAF filters are often applied to remove singletons or other rare variants, as described above, these variants are critical to several analyses including demographic history estimation and tests for selection. Most notably, Tajima's *D* value, a commonly used indicator of both demographic history and response to selection[93], is substantially biased by a MAF filter choice, leading to widely differing biological inferences depending on filtering stringency (Box 2). In this case, our recommendations are straightforward: because low-frequency alleles heavily influence Tajima's *D* value[94], researchers should apply both no MAF filter and a very minor one (such as a singleton filter) and compare the results when using the statistic. The effects of MAF and other filters can be substantial for diversity estimators[95], demographic inference[12,96], F_{ST} (ref. 47), gene flow[14], population structure estimates[97,98], estimating the distribution of locus effects on phenotypes[98] and allele frequency spectra[47,99–101]. Other filtering choices therefore require similar levels of care (Table 1).

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 (for example, all individuals) within a study jointly and simultaneously, we refer to this process as study-wide filtering (or 'global' filtering). When filtering occurs within each sample-group separately, we refer to this process as within-group filtering.

The effects of within-group versus study-wide filtering can be surprisingly large. For example, when applying a within-group MAF filter of 0.01 to a yellow perch (*Perca flavescens*) WGS dataset, the number of SNPs within each population varied by a factor of 3.3 (ranging from 670,578 to 2,275,935) (Fig. 2). However, when the same 0.01 MAF filter was applied globally, each sample-group was constrained to 714,000 SNPs[102]. In this case, some populations in the study had radically different SFS, likely caused by recent population expansions that resulted in an increase of rare variants[93,94]. In general, study-wide filtering can therefore lead to the removal of critically informative, globally rare but locally common alleles; thus, filtering MAF globally (a common practice) instead of within study groups is expected to have substantial 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 affect 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 work)[103]. The implications of these standardized pipelines are often not given much consideration, but the effects may be non-trivial. For example, when comparing populations with different SFS, 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, as populations with African ancestries tend to have more sites with low-frequency alleles than those with European ancestries[104]. 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. Although 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. For example, differences in downstream outcomes from filtering HWP[86] and LD[105] within-groups versus study-wide have been documented previously.

In light of these findings, it is crucial to consider why results differ when applying filters globally or within groups, particularly if sample-groups include individuals from different populations, locations or time points[39]. For example, tests for HWP should always be 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 (that is, a Wahlund effect), and their removal can mask the population structure[82,106] (Fig. 2). If a specific locus shows consistent deviation from HWP greater than the genome-wide trend in multiple different groups, this may indicate a genotyping error (such as allelic dropout) or alignment or genome assembly errors (that are not necessarily caused by biological processes)[90].

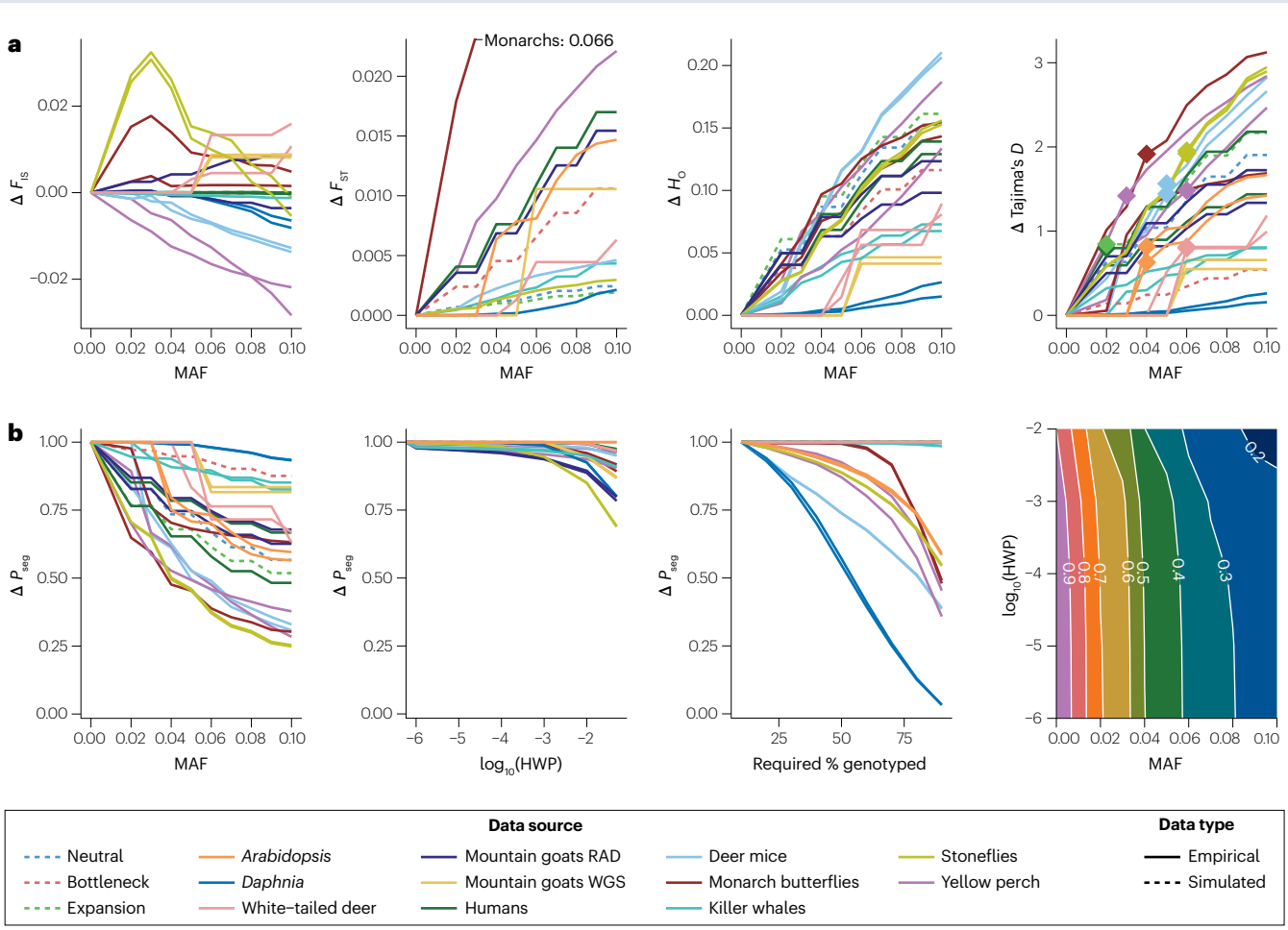
Of note, clearly defined putative populations are not always present. Forcing sample-groups on data with no clearly defined biological boundaries and then filtering on those sample-groups potentially risks creating biases, such as the artificial creation of population structure

Box 2 | Effects of post-variant 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*[145], *Daphnia*[146], white-tailed deer[147], mountain goats[148], humans[149], deer mice[150], killer whales[151], monarch butterflies[114], stoneflies[152] and yellow perch[102]) (Supplementary Methods and Supplementary Table 2) along with three simulated datasets (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, the minor allele frequency (MAF), on four commonly used parameters (see the figure, panel **a**). We also show how multiple filters (MAF, Hardy–Weinberg proportions (HWP) and missing data) and the higher-order interaction between MAF and HWP influence a single parameter (the proportion of retained segregating sites (P_{seg})) (see the figure, panel **b**). Parameter estimates were standardized to represent relative change across filter thresholds. A wide range of additional filtering effects and non-standardized values are presented in Supplementary Figs. 1–4.

Changing the filtering threshold for a single filter can result in large changes in F_{ST}, F_{IS}, H_O/H_e and Tajima's D estimates. Increasing MAF thresholds reliably increases the average F_{ST}, H_O (per segregating site) and Tajima's D value[47,94]; F_{IS}, however, is impacted variably among datasets (see the figure, panel **a**). Filtering with the most commonly used MAF threshold of 0.05 can often flip the genome-wide sign of Tajima's D value from negative to positive, changing its interpretation from a population expansion to a bottleneck (see the figure, panel **a**, right, diamonds indicating change in sign) (Supplementary Fig. 1). In addition to the parameters shown here, MAF filtering can also substantially change estimates of nucleotide diversity (π), private allele counts, Watterson's θ and effective population size (N_e) estimates derived from linkage disequilibrium (LD)-based approaches[120] (Supplementary Fig. 2).

Varying filtering thresholds across multiple filters also results in substantial changes to a single parameter, which we illustrate for the estimated number of segregating sites (see the figure). MAF filters, which are perhaps the most widely used, can have a particularly strong effect that, although constant in direction, can widely vary in



(continued from previous page)
magnitude depending, primarily, upon the shape of the underlying site-frequency spectra (SFS) across populations (Supplementary Fig. 6). HWP and missing data filtering (here "Required % genotyped" such that any loci or individuals with less than the noted percentage of called genotypes were removed) also impact different datasets in different ways — datasets where more loci are out of HWP lose more segregating sites with higher filters for HWP, as do those with more missing data. Filtering effects were generally not different across dataset types (reduced-representation versus whole-genome), although restriction site-associated DNA (RAD) datasets were generally more strongly impacted by higher HWP filters for most statistics (Supplementary Fig. 3). Higher-order interactions between filters may also be important — there are substantial changes in the average impact of filtering for HWP on the proportion of retained segregating sites depending on the MAF filter used (see the figure, panel **b**, right) — and are deserving of more study. In the examples provided here, researchers could come to different conclusions about demographic history (Tajima's *D* value), selection (Tajima's *D* value, HWP) and genetic diversity (H_O, proportion of segregating sites) based on the initial filtering thresholds selected and their higher-order interactions (which are commonly ignored).

The demographic context and genomic architecture of the study system (including model species) can also affect filtering. Populations that have undergone recent population expansions, for example, will lose far more rare alleles during MAF filtering than will those that have undergone population bottlenecks (Supplementary Fig. 6). This is the case with the monarch and yellow perch datasets, which correspondingly have the largest increases in F_{ST} with higher MAF filters (see the figure) (Supplementary Figs. 1 and 2). Genomic architecture also affects filtering impacts: for example, the removal of regions of elevated F_{ST} caused by selection occurring in areas with reduced recombination rates will have a larger effect on genome-wide principal component analysis (PCA) results than filtering elsewhere (Supplementary Fig. 7). Many parameters such as F_{ST}, F_{IS} and LD could also be influenced by function; genotypes adjacent to conserved exons, introns, centromeres, telomeres or sex-linked loci may all respond differently to filtering thresholds. WGS, whole-genome sequencing.

by forcing 'populations' to conform to HWP[86]. In cases where sample-group delineations do not accurately represent biological realities, a suitable approach might be to first identify population structure through PCA or another agnostic clustering approach, and then assess the impact of both within-group and study-wide filtering.

Solutions and best practices

Filtering is a powerful tool that should be applied thoughtfully, early and often throughout genomic dataset construction alongside tests for unintentional or unanticipated issues with a dataset (for example, experimental, sample collection, labelling library preparation

Table 1 | Recommended initial filtering thresholds for producing low-stringency and high-stringency filtered datasets

Objective	Individual missing data; <*X*% missing loci[a]	Loci missing data; <*X*% missing individuals	MAC or MAF		LD	HWP
Population structure	50%; 5%	50%; 5%	>1	>0.05	No; yes	1×10^{-6}; 0.01
Demography	25%; 5%	50%; 5%	1	No filter[b]	Yes	1×10^{-6}; 0.01
Selection	75%; 25%	75%; 25%	>1	>0.05	No	None; 1×10^{-4}
Genetic diversity	50%; 5%	50%; 15%	>1	>0.05	No, usually	None; 1×10^{-4}
Phylogenetic reconstruction	20%; 5%	50%; 10%	>1	>0.05	Yes	0.001; 0.01
GWAS	50%; 10%	50%; 10%	MAF only	>0.05, lower with large *n*	No, but correct *P* values	None; 1×10^{-6}
Mutation detection	Parents: 0%; offspring: 100%	Parents: 0%; offspring: 100%	No filter	No filter	None; 1×10^{-6}	–
Metagenomics or eDNA	–	–	>4	MAC only	Context-dependent	–
Relatedness or pedigree construction	<95%, provided sufficient remaining loci for power	20%; 5%	>2	>0.05	No, although CIs can be affected	0.001; 0.01

The thresholds proposed are suggestions and should not supplant existing knowledge of the study system or design (Box 3). These values represent relatively extreme values to examine effects, whereas moderate values may be preferable for final analysis. Justifications for these suggestions are provided in Supplementary Table 4. Columns represent the objective of the study; where individuals with missing data at more than *X*% loci are removed; where loci missing data in more than *X*% of individuals are removed; the MAC or MAF filtering threshold below which loci are removed (use MAC or MAF, not both together); whether loci should be removed to prevent excessive non-independent pairs; and where loci with a *P* value below the threshold are removed. CI, confidence interval; eDNA, environmental DNA; GWAS, genome-wide association studies; HWP, Hardy–Weinberg proportions; LD, linkage disequilibrium; MAC, minor allele count; MAF, minor allele frequency. [a]Some datasets may need a less strict filter if most individuals are removed. [b]For site-frequency-spectra-based estimators such as Tajima's *D* value, singletons could be retained for one filtering extreme.

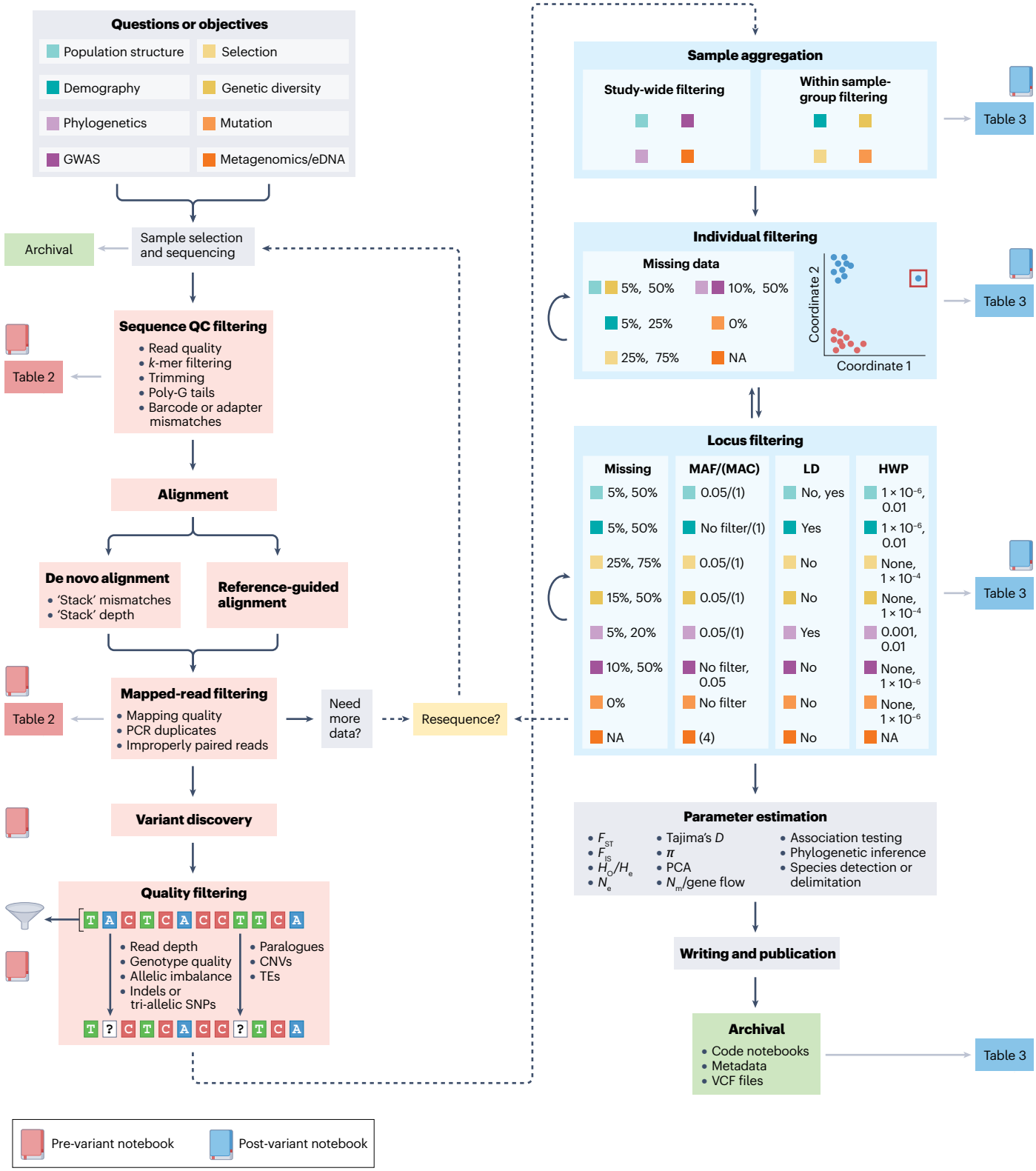

Questions or objectives
Population structure
Selection
Demography
Genetic diversity
Phylogenetics
Mutation
GWAS
Metagenomics/eDNA
Archival
Sample selection and sequencing
Sequence QC filtering
Read quality
k-mer filtering
Trimming
Poly-G tails
Barcode or adapter mismatches
Table 2
Alignment
De novo alignment
'Stack' mismatches
'Stack' depth
Reference-guided alignment
Mapped-read filtering
Mapping quality
PCR duplicates
Improperly paired reads
Need more data?
Resequence?
Variant discovery
Quality filtering
Read depth
Genotype quality
Allelic imbalance
Indels or tri-allelic SNPs
Paralogues
CNVs
TEs
Sample aggregation
Study-wide filtering
Within sample-group filtering
Table 3
Individual filtering
Missing data
5%, 50%
10%, 50%
5%, 25%
0%
25%, 75%
NA
Coordinate 2
Coordinate 1
Locus filtering
Missing
MAF/(MAC)
LD
HWP
5%, 50%
0.05/(1)
No, yes
1 × 10-6, 0.01
5%, 50%
No filter/(1)
Yes
1 × 10-6, 0.01
25%, 75%
0.05/(1)
No
None, 1 × 10-4
15%, 50%
0.05/(1)
No
None, 1 × 10-4
5%, 20%
0.05/(1)
Yes
0.001, 0.01
10%, 50%
No filter, 0.05
No
None, 1 × 10-6
0%
No filter
No
None, 1 × 10-6
NA
(4)
No
NA
Parameter estimation
FST
FIS
HO/HE
Ne
Tajima's D
π
PCA
Nm/gene flow
Association testing
Phylogenetic inference
Species detection or delimitation
Writing and publication
Archival
Code notebooks
Metadata
VCF files
Pre-variant notebook
Post-variant notebook

Fig. 3 | Flow chart to facilitate thoughtful, systematic and reproducible filtering for representative studies and questions using genomic DNA. Typical filtering workflows proceed through raw sequence quality control (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. Regardless of the study objectives, multiple datasets should be constructed to examine the effects of various filtering decisions. Suggested filtering thresholds per locus and for individual filtering are provided for each question and objective (Table 1 and Supplementary Table 4). Researchers should use a reproducible workflow to help them more easily repeat steps during analysis and the review process. Reproducible workflows can aid laboratories or research groups if more data will be produced in the future (for example, by students or postdoctoral researchers) as well as researchers in other laboratories. Data should be carefully archived before and after filtering, and all filtering methods and results carefully reported. See Supplementary Table 1 for a complete list of filters, Tables 2 and 3 for a simplified example of how to report filtering results, Box 4 for a filtering and reporting checklist, and the pre-variant and post-variant filtering R notebooks for examples of reproducible workflows (Supplementary Notebook 1 and 2). CNV, copy number variation; eDNA, environmental DNA; F_{IS}, fixation index, individuals versus subpopulation; F_{ST}, fixation index, subpopulation versus total; GWAS, genome-wide association studies; H_e, expected heterozygosity; H_o, observed heterozygosity; HWP, Hardy–Weinberg proportions; LD, linkage disequilibrium; MAC, minor allele count; MAF, minor allele frequency; NA, not applicable; N_e, effective population size; N_m, number of migrants; PCA, principal component analysis; SNP, single-nucleotide polymorphism; TE, transposable element.

errors, batch-effect sequencing and genome assembly errors) and to detect interesting biological phenomena (for example, natural selection, structural variants and Wahlund effects). Even in highly studied species, such as humans, a carefully considered and multi-faceted approach to filtering is important because novel structural and genetic variants can occur within every population[107], and failing to account for these variants may curtail power to identify causal associations or even lead to incorrect inferences[72]. To assist investigators in matching research questions with methods and filters, we have created a detailed flow chart describing the filtering process for a general genomics workflow that can be applied across disciplines and study systems (Fig. 3).

Quantification of filtering effects

Because different filtering choices can result in different downstream inferences, we recommend that distinctly 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 should be created — one with low filtering stringency (for example, allowing more missing data, a low, permissive MAF threshold and few loci removed due to deviations from HWP and LD); and one with high filtering stringency (for example, many loci or individuals removed due to missing data and a higher, restrictive MAF threshold). Creating two datasets using relatively broad filtering values (for example, low and high stringency) allows researchers to test whether distinct filtering thresholds affect analyses 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 (Box 3). For example, researchers should consider using low-stringency MAF filters for several demographic inferences (for example, Tajima's *D* value, SFS) but relatively stringent MAF filters for delineating populations[97], planning genomically informed breeding strategies[108] or estimating parentage or individual relatedness[109–111]. Studies interested in transposable elements may want to vary alignment thresholds (uniquely versus multiply mapped reads) but keep other filters stringent to strike a balance between sensitivity and accuracy[112,113].

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 filters[14,97,114,115], although they are in the minority. Note that we are not the first to suggest comparing outcomes from different filtering strategies[12,14,47], and we suspect that this recommendation will become more common, and more commonly followed, over time.

The concept of using multiple, distinctly filtered datasets requires a fundamental shift in the way genomics data are analysed:

Box 3 | The importance of study system knowledge

We recommend that researchers have a thorough understanding of both their study system and population-genetics theory before planning filtering strategies and interpreting results[16]. Critically, knowing a species' ecology, demography and pre-existing genetic results can provide important a priori expectations for analytical results and suggest sources of filtering (or other) errors. At every step, researchers should ask themselves whether, given their knowledge of their study species, theory and model assumptions, their results and filtering choices make sense.

For example, little geographical structuring might be expected in species with high gene flow, such as many migratory birds or organisms with highly dispersive early life-history stages (for example, many plants, arthropods and marine organisms), and thus the detection of relatively strong population subdivision (or high F_{ST} values) could be an artefact of data filtering[60,74]. Similarly, knowing a species' mating system, degree of population isolation and dispersal propensity can help to determine whether, for example, high inbreeding and/or low effective population size (Ne) estimates are biological or produced artificially by filtering choices (Box 1). Strong genetic signals of recent population bottlenecks in populations known to have undergone demographic expansions might also suggest filtering issues (Box 2). Discussions with local biologists, Indigenous peoples[153] and regional and federal agencies can also greatly assist with identifying spurious results. A solid understanding of a species' biology is also useful for model species, where known recombination rates, genomic organization[154] and different histories of captive breeding[155] can help to predict and interpret filtering results.

Glossary

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

Contigs
Contiguous sequences of DNA assembled from many overlapping sequence reads, representing a fragment of a chromosome.

De novo assembly
The reference-free alignment of sequencing reads into overlapping stacks or contigs for subsequent use in variant discovery and genotyping.

F_{IS}
A measure of inbreeding; the degree of subpopulation divergence from Hardy–Weinberg proportions — the correlation between alleles at specific loci within individuals relative to the subpopulation.

F_{ST}
A measure of population differentiation; the proportion of the total genetic variance due to differences in allele frequencies between subpopulations.

Genetic variants
Differences in DNA sequence compared with a reference sequence or other individuals within a population. The term includes short variants (single-nucleotide polymorphisms (SNPs) or insertions and deletions) and structural variants (chromosomal inversions and copy number variations (CNVs)). In the context of this Review, used interchangeably with 'locus'.

Genome-wide association studies
(GWAS). Tests for statistical relationships between a phenotype (including disease) and the allelic/genotypic state of an (ideally) large cohort of individuals across the entire set of sequenced loci.

Genotyping
Also referred to as genotype or variant calling. Calling allelic states at a locus (for example, A/A, A/C or C/C at a biallelic single-nucleotide polymorphism (SNP) in a diploid organism) or loci from sequence data. Genotyping algorithms often consist of multiple steps during which filtering can occur.

Haplotype phase
The complete sequence of variants that occur in a region along a single chromatid.

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

Imputation
The filling in of missing data for specific genotypes and/or loci by leveraging linkage disequilibrium (LD) between missing genotypes and genotypes called 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). The non-random association of alleles at different loci within a population or sample-group. This association can either be caused by physical linkage, when alleles are co-inherited due to non-independent assortment caused by close physical proximity, or occur across chromosomes when inbreeding, paralogy, genetic drift or other factors make certain alleles at different loci more likely to co-occur.

Low-coverage whole-genome sequencing
Whole-genome sequencing (WGS) with small numbers of reads covering most genomic loci (low coverage); the number of reads constituting low coverage varies widely depending on the discipline, methodology and research question. Low-coverage WGS often requires genotype likelihood-based methods.

Mapping quality
The score given to a read or other DNA sequence indicating the uniqueness of the alignment to a reference sequence; mapping quality score interpretations vary across alignment programs.

Minor allele count
(MAC). The number of gene copies or individuals carrying the minor (that is, 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; 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 factors, such as the absence of a sufficient number of reads covering a locus to call a genotype in an individual with any degree of confidence.

N50 or L50 scores
In a genome assembly after sorting contigs or scaffolds by length, either the length of the contig/scaffold that reaches 50% of the cumulative genome length (N50) or the number of contigs needed to reach 50% of the cumulative genome length (L50); used to evaluate the assembly quality.

Paralogues
Duplicated genomic regions that have arisen via either the duplication of that specific region or the duplication of the entire genome. A type of homologue (loci identical by descent) distinct from orthologues, which arise due to speciation events.

PCR duplicates
Technical duplicates resulting in spurious, usually identical read copies caused by repeatedly sequencing the same piece of template DNA multiple times.

Population structure
Also known as population subdivision. Non-independence among individuals in a study area/region caused by spatial, temporal, behavioural or other forms of reproductive isolation. Population structure is characterized by divergent allele frequencies across loci.

Read depth
The number of reads that cover a given or fixed genomic position. Also referred to as 'coverage'.

Reference bias
The propensity for reads containing the non-reference allele (the allele not in the reference genome) to have lower mapping quality scores or map to the wrong location compared with those containing the allele present in the reference genome.

Runs of homozygosity
Contiguous homozygous regions of the genome caused by the inheritance of identical haplotypes from both parents (for example, identical by descent). Useful for estimating inbreeding and population demographics.

Glossary (continued)

Sample-group
A group of samples that are not independent due to natural causes (such as geographic or temporal separation) and/or experimental treatments.

Single-nucleotide polymorphisms
(SNPs). Genetic variants where the allelic state of the population varies at a single base pair.

Singletons
Alleles that appear only once in a sample of individuals. Sometimes alternatively defined as an allele sequenced in only one individual (which may be homozygous for that allele).

Site-frequency spectra
(SFS). The distributions of allele frequencies across loci within a study or sample-group. Can be either an 'unfolded' or 'polarized' derived allele frequency spectrum which describes the frequency distribution of derived alleles or a 'folded' or 'unpolarized' minor allele frequency (MAF) spectrum which describes the frequency distribution of the minor alleles. Also known as the allele frequency distribution.

Structural variation
Genetic 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) rather than separately within each sample-group.

VCF file
A file in the variant call format, which contains genotype calls (or likelihoods, posteriors) alongside a flexible suite of metadata such as filtering and processing history and quality information.

Wahlund effect
A reduction in observed heterozygosity (H_O) relative to the expected heterozygosity (H_e) under Hardy–Weinberg proportions (HWP) (that is, $H_O < H_e$) at many/most loci caused by the 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}, the fixation index in individuals relative to a subpopulation).

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

investigators must realize 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 no substantial bias or error from filtering (Box 1). We provide recommendations for initial filtering thresholds for many types of questions and analyses in Table 1. These thresholds can and should be modified given the characteristics (such as sample sizes and general quality) of the data at-hand.

Best practices for pre-variant calling workflows

Most genomic workflows differ depending on the research question and data types (Fig. 3). The documentation of filtering decisions is therefore paramount for reproducibility. As a first step prior to any analysis, we recommend that raw data be immediately archived (privately or publicly) in independent, non-local repositories created for genomics data (for example, the NCBI Short-Read Archive, the European Variation Archive or the DNA DataBank of Japan Sequence Read Archive); other genomics data management best practices are reviewed elsewhere[116,117]. Given that filtering, by definition, requires manipulating data, the importance of archiving raw data cannot be understated. To this point, we refer the reader to ref. 118 for information on dataset and study organization.

After archiving, reads should be filtered for general QC (base quality, adapter removal, poly-G tails, sequencing artefacts) (Supplementary Table 1) and trimmed when appropriate and useful[119,120]. For most workflows, the alignment of reads to a reference or de novo assembly is the next step (Fig. 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[114,121]. This practice is particularly relevant to de novo reference assembly, as assembly decisions can result in very different references and, thus, very different filtering and analytical outcomes. For example, the *m* and *M* STACKS parameters and their impact on de novo reference construction have been well studied[41,42,122,123].

After alignment, the data should be filtered for technical (for example, PCR) duplicates. Although removing PCR duplicates has been suggested to be of little consequence[124,125], this is unlikely true for every study, such as those with low-coverage data[10,66,126]. 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 (Table 2). We have provided an R notebook (Supplementary Notebook 1) that uses a small example dataset to walk through an entire pre-variant filtering workflow — from raw reads to called genotypes — using various commonly implemented tools and provides an example of how to easily change, and importantly record, filtering parameters with minimal effort.

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 and paralogues) and, then remove poorly sequenced individuals (and/or samples with other quality or analytical concerns) (Fig. 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 individuals[8]. An iterative approach, where individuals and loci are first removed with low-stringency filters and then subjected to additional rounds of filtering stringencies may also improve data quality by removing individuals who reduce overall call rates in high-quality loci and vice versa[8]. Similarly, if a MAF filter is used to remove loci after variant discovery followed by the removal of individuals with too much missing data, 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 3 and Box 4).

Table 2 | Example of pre-variant filtering reporting standards

Filtering step	Results to report	Values
Sample selection and sequencing	Number of individuals or samples collected	*n*=250 individuals
	Number of samples initially sequenced	*n*=210 individuals
Sequence QC	Number of individuals who were successfully sequenced[a]	*n*=200 individuals
	Total number of reads prior to any filtering	1.5×10^8 reads
	Number of reads remaining after filtering for read quality[b]	0.8×10^8 reads
Mapped-read filtering	Number of reads that mapped	6.8×10^7 reads
	Number of reads remaining after filtering for mapping quality	5.1×10^7 reads
	Number of reads remaining after filtering for improperly paired reads	4.8×10^7 reads
	Number of reads remaining after filtering for PCR duplicates[c]	8×10^6 reads

All accompanying code and filtering steps should be reported in the pre-variant notebook; see Fig. 3 for a detailed flow chart and Supplementary Notebook 1 for the accompanying pre-variant notebook. The sequence of filtering events can affect downstream results, so rows should be arranged chronologically. Readers will be able to determine which filtering steps had the largest effects on dataset size from the table alone. Notice in this heuristic example that alignment (number of reads that mapped) and filtering for PCR duplicates had large effects. [a]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. [b]Any additional filters used during sequence quality control (QC), such as filtering for poly-G tails or adapter mismatches, should be given their own row. [c]Any additional filters used during filtering of mapped reads should be given their own row.

As data analysis proceeds, we suggest that re-filtering should be part of most genomics workflows. For example, PCA — or more sophisticated, related approaches[127] — can reveal individuals who were mislabelled, misclassified into an incorrect sample-group, contaminated during sample preparation or closely related (for example, full siblings)[128–130]. Such underlying causes should be carefully investigated, and problematic samples reported, and possibly removed, prior to the re-calculation of downstream statistics. Similarly, the decision to conduct new analyses (for example, Tajima's *D* value, transposable element annotation, parentage analysis) that were not initially considered may also require re-filtering of data. Lastly, many genomic datasets may contain batch effects — non-biological differences between samples that arise from independent sequencing runs — and these effects should be explored and explicitly accounted for when filtering samples that were sequenced in different batches[131].

After analyses are completed, the resulting data should again be archived and/or recorded, including all relevant metadata 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 limited access to old or out-of-date software), we strongly recommend that post-project archives include all filtered genotypic/variant data in the form of carefully annotated VCF files that include detailed filtering descriptions in the header[132].

We strongly suggest that authors and journals require supplementary tables that describe the final datasets, the specific filters and thresholds employed, the names of the final VCF files and the specific analyses for which each distinctly filtered dataset was used. We provide examples of these in Tables 2 and 3. Researchers should also explain whether they corrected for multiple testing along with a (brief) justification for the correction method used (for example, Bonferroni, false discovery rate (FDR))[15]. If reasonable, we also suggest

Table 3 | Representative table demonstrating post-variant reporting standards for different objectives

Filtering step	Population structure (dataset 1)	GWAS-stringent (dataset 2)	GWAS-relaxed (dataset 3)	Demography (dataset 4)
Minimum genotype quality	40	50	40	40
Minimum genotype coverage	10×	20×	15×	10×
Maximum genotype coverage	30×	30×	50×	30×
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[a]	15%	5%	10%	5%
MAF/MAC	MAF=0.05	MAF=0.05	MAF=0.01	MAC=2
LD	No filter used	No filter used	No filter used	$r^2>=0.25$
Hardy–Weinberg deviations	$P<0.05$; Bonferroni[b]	$P<0.05$; Bonferroni[b]	No filter used	$P<0.05$; no FDR correction
Results[c]	Fig. 1b,c	Figs. 2 and 3	Supplementary Fig. 1	Fig. 4
VCF md5sum	8d3d627940ee2a77 b4770db1fd710459	3dcccbf8d3fb869c3cf 5de291c0fe893	0b0681ad8b5bdab39e 7b76afc190d4c8	1f131fdc2ee6444e1b 94071195a1acd2

All accompanying code and filtering steps should be reported in the post-variant notebook; see Fig. 3 for a detailed flow chart and see 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 20× coverage per individual and 200 sequenced individuals (following from Table 1). FDR, false discovery rate; GWAS, genome-wide association studies; LD, linkage disequilibrium; MAC, minor allele count; MAF, minor allele frequency. [a]Missing data should be examined both study-wide and within sample-groups; different sample-groups may contain different amounts of missing data. [b]Hardy–Weinberg filters should only be applied within each sample-group, not study-wide; corrections for multiple comparisons should be reported. [c]Figure numbers indicate figures or supplementary material in the hypothetical paper for which this table is used to report filtering across datasets.

Box 4 | Filtering checklist

Throughout dataset assembly (from raw sequencing reads to genotypes), researchers should take care to explore the effects of alternative filtering strategies on downstream analyses and aid in reproducibility. To aid with this, the example checklist (see the table) should be consulted before and during a research project and checked-off prior to submitting a manuscript for peer review.

Analysis step	Reporting step
☐ Data archival	
☐ Decide on filtered datasets given a priori study questions, knowledge of the system and population-genetics theory (Fig. 3 and Box 3)	☐ Create filter recording and reporting tables (Tables 2 and 3)
☐ Filter on raw sequences, for example, read quality or poly-G tails (Table 2)	☐ 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, discordant read pairs (some variant callers will do so automatically if these are marked)
	☐ Report number of reads retained and filtered at each step
☐ Variant discovery	
☐ Begin or continue creation of multiple datasets	☐ 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, paralogues, coverage
	☐ Report the number of SNPs remaining after these steps to understand which steps remove the most loci
☐ Individual filtering	☐ Missing data; mislabelling or contamination
☐ Data analysis and parameter estimation	☐ Report effects of filters on parameters and questions of interest (Tables 2 and 3)
☐ Perform re-filtering and/or re-sequencing if necessary	
☐ Final filter recording	Report reads, loci, individuals lost at each step
☐ Archive all filtered datasets as VCF files	

HWP, Hardy–Weinberg proportions; MAF, minor allele frequency; SNP, single-nucleotide polymorphism.

that the downstream statistical effects of different filtering thresholds be reported (either in the main or supplementary text) to improve the scientific community's understanding of filtering impacts. 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 Supplementary Notebook 2.

The benefits of re-filtering and reproducible research

Thorough examination of filtering (and re-filtering) will necessitate extra time, computational resources and work from researchers. However, changing and testing workflows (that is, re-filtering) is 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 and their future students and laboratory members. A reproduction-friendly pipeline that runs a suite of analyses given a dataset and a set of filtering parameters is also easy to re-run a second time with a new (re-filtered) dataset[133] (Fig. 3). Indeed, reproduction-friendly pipelines show an additional benefit: they minimize the time needed to re-run filtering steps and, thus, ensure that testing several filter thresholds is a relatively painless process. For examples of studies with well-documented methods and easily accessible data that would be relatively straightforward to reproduce with new filters and thresholds, see refs. 97,134,135.

Journal reviewers should be reasonable when asking authors to reanalyse their data with different filtering parameters. 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 several filtering thresholds, the application of additional filters is likely not necessary.

Conclusions and future directions

Advancements in genomic sequencing technologies, improvements in reference quality[136–138] and the burgeoning field of pangenomics[139,140] 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 (for example, demography, life history) and the questions they intend to answer.

Filtering effects can be unpredictable and there is no single best strategy for filtering all genomic datasets. Critically, we highlight that different filtering thresholds can create different downstream results and conclusions for the same dataset. Most computational analyses should therefore be re-run on multiple datasets produced by re-filtering using different filters and thresholds to facilitate 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 the effects of filtering choices on downstream inferences, which will help to usher in the next generation of data filtering and improve genomics applications across disciplines from ecology and evolution to human health, agriculture and the conservation of biodiversity.

Data availability

Information on the empirical and simulated data used for the analyses shown in this review is available in the Supplementary Information.

Code availability

The simulation code is available on GitHub at: https://github.com/ChristieLab/filtering_simulation_paper.

Published online: 14 June 2024

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**, 30 (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. *Hum. 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).
   This review summarizes the sources of many common types of sequencing errors and provides some laboratory and bioinformatic ways to mitigate them.
6. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **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're looking for: principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* **27**, 3193–3206 (2018).
   This helpful review discusses the effects of missing data, MAC and other filters on genotyping error rates for RADseq data.
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).
    This study demonstrates the effects of different filtering and alignment choices on several downstream statistics and demographic reconstruction in RADseq data.
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**, 677009 (2021).
    This comprehensive analysis of empirical data demonstrates how missing data and MAF thresholds affect estimates of gene flow.
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 Univ. Press, 2022).
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. & Kimura, M. *An Introduction to Population Genetics Theory* (Scientific Publishers, 2017).
19. Van Etten, J., Stephens, T. G. & Bhattacharya, D. A *k*-mer-based approach for phylogenetic classification of taxa in environmental genomic data. *Syst. Biol.* **72**, 1101–1118 (2023).
20. 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).
21. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
22. 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).
23. Upton, R. N. et al. Design, execution, and interpretation of plant RNA-seq analyses. *Front. Plant Sci.* **14**, 1135455 (2023).
24. 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).
25. Boshuizen, H. C. & te Beest, D. E. Pitfalls in the statistical analysis of microbiome amplicon sequencing data. *Mol. Ecol. Resour.* **23**, 539–548 (2023).
26. Combrink, L. et al. Best practice for wildlife gut microbiome research: a comprehensive review of methodology for 16S rRNA gene investigations. *Front. Microbiol.* **14**, 1092216 (2023).
27. 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**, 2146 (2022).
28. Yang, L. & Chen, J. Benchmarking differential abundance analysis methods for correlated microbiome sequencing data. *Brief. Bioinformatics* **24**, bbac607 (2023).
29. Patin, N. V. & Goodwin, K. D. Capturing marine microbiomes and environmental DNA: a field sampling guide. *Front. Microbiol.* **13**, 1026596 (2023).
30. 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).
31. Deyneko, I. V. et al. Modeling and cleaning RNA-seq data significantly improve detection of differentially expressed genes. *BMC Bioinformatics* **23**, 488 (2022).
32. Giusti, A., Malloggi, C., Magagna, G., Filipello, V. & Armani, A. Is the metabarcoding ripe enough to be applied to the authentication of foodstuff of animal origin? A systematic review. *Compr. Rev. Food Sci. Food Saf.* **23**, 1–21 (2024).
33. da Fonseca, R. R. et al. Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Mar. Genomics* **30**, 3–13 (2016).
34. Zhao, M. et al. Exploring conflicts in whole genome phylogenetics: a case study within manakins (Aves: Pipridae). *Syst. Biol.* **72**, 161–178 (2023).
35. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med* **12**, 91 (2020).
36. 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).
37. Kumar, K. R., Cowley, M. J. & Davis, R. L. Next-generation sequencing and emerging technologies. *Semin. Thromb. Hemost.* **45**, 661–673 (2019).
38. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
39. 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).
    This reviews discusses the production and analysis of low-coverage WGS data.

40. Olson, N. D. et al. Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* **24**, 464–483 (2023).
41. Rochette, N. C. & Catchen, J. M. Deriving genotypes from RAD-seq short-read data using Stacks. *Nat. Protoc.* **12**, 2640–2659 (2017).
42. 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).
43. 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).
44. 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).
45. Bohling, J. Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecol. Evol.* **10**, 7585–7601 (2020).
46. 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).
47. Hendricks, S. et al. Recent advances in conservation and population genomics data analysis. *Evol. Appl.* **11**, 1197–1211 (2018).
48. Vaux, F., Dutoit, L., Fraser, C. I. & Waters, J. M. Genotyping-by-sequencing for biogeography. *J. Biogeogr.* **50**, 262–281 (2023).
49. 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).
50. 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).
51. 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).
52. 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).
53. Zeng, L. et al. Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* **20**, 79 (2019).
54. Zhernakova, D. V. et al. Genome-wide sequence analyses of ethnic populations across Russia. *Genomics* **112**, 442–458 (2020).
55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
57. Pfeifer, S. P. From next-generation resequencing reads to a high-quality variant data set. *Heredity* **118**, 111–124 (2017).
58. Lefouili, M. & Nam, K. The evaluation of BCFtools mpileup and GATK HaplotypeCaller for variant calling in non-human species. *Sci. Rep.* **12**, 11331 (2022).
59. 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).
60. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genet.* **15**, e1008302 (2019).
61. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
62. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
63. Singh, A. K. et al. Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Med. Genomics* **14**, 214 (2021).
64. 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).
65. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
66. Rochette, N. C. et al. On the causes, consequences, and avoidance of PCR duplicates: towards a theory of library complexity. *Mol. Ecol. Resour.* **23**, 1299–1318 (2023).
67. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (O'Reilly Media, 2020).
68. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
69. Eaton, D. A. R. & Overcast, I. ipyrad: interactive assembly and analysis of RADseq datasets. *Bioinformatics* **36**, 2592–2594 (2020).
70. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
71. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
72. Mona, S., Benazzo, A., Delrieu-Trottin, E. & Lesturgie, P. Population genetics using low coverage RADseq data in non-model organisms: biases and solutions. Preprint at *Authorea* https://doi.org/10.22541/au.168252801.19878064/v1 (2023).
73. 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).
74. 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).
75. 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).
76. Huang, H. & Knowles, L. L. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Syst. Biol.* **65**, 357–365 (2016).
77. Duntsch, L., Whibley, A., Brekke, P., Ewen, J. G. & Santure, A. W. Genomic data of different resolutions reveal consistent inbreeding estimates but contrasting homozygosity landscapes for the threatened Aotearoa New Zealand hihi. *Mol. Ecol.* **30**, 6006–6020 (2021).
78. Kardos, M. & Waples, R. S. Low-coverage sequencing and Wahlund effect severely bias estimates of inbreeding, heterozygosity, and effective population size in North American wolves. *Mol. Ecol.* https://doi.org/10.1111/mec.17415 (2024).
This study reports biases that could affect management decisions caused by next-generation sequencing filtering choices, low-coverage data and the sampling strategy.
79. Schmidt, T. L., Jasper, M.-E., Weeks, A. R. & Hoffmann, A. A. Unbiased population heterozygosity estimates from genome-wide sequence data. *Methods Ecol. Evol.* **12**, 1888–1898 (2021).
80. Sopniewski, J. & Catullo, R. A. Estimates of heterozygosity from single nucleotide polymorphism markers are context-dependent and often wrong. *Mol. Ecol. Resour.* **24**, e13947 (2024).
81. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
82. Waples, R. S. Testing for Hardy–Weinberg proportions: have we lost the plot? *J. Hered.* **106**, 1–19 (2015).
83. 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).
84. 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).
85. Bitarello, B. D., Brandt, D. Y. C., Meyer, D. & Andrés, A. M. Inferring balancing selection from genome-scale data. *Genome Biol. Evol.* **15**, evad032 (2023).
86. 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).
This study demonstrates the impact of pooling or splitting sample-groups when applying HWP filters to F_{ST} and other population structure inferences.
87. Linderoth, T. P. Identifying population histories, adaptive genes, and genetic duplication from population-scale next generation sequencing. *Genome Res.* **20**, 291–300 (2018).
88. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
89. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
90. 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).
91. 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).
92. 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).
93. 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).
94. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585 LP–585595 (1989).
95. 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.* https://doi.org/10.1111/1755-0998.13859 (2023).
96. Cubry, P., Vigouroux, Y. & François, O. The empirical distribution of singletons for geographic samples of DNA sequences. *Front. Genet.* **8**, 139 (2017).
97. 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).
This study demonstrates how MAF thresholds affect population structure inferences using both simulated and empirical data.
98. 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.* **23**, 1589–1603 (2023).
99. Díaz-Arce, N. & Rodríguez-Ezpeleta, N. Selecting RAD-seq data analysis parameters for population genetics: the more the better? *Front. Genet.* **10**, 533 (2019).
100. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nat. Rev. Genet.* **10**, 639–650 (2009).
101. Roesti, M., Salzburger, W. & Berner, D. Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* **12**, 94 (2012).
102. Yin, X. et al. Rapid, simultaneous increases in the effective sizes of adaptively divergent yellow perch (*Perca flavescens*) populations. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.21.590447 (2024).
103. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
104. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
105. 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).

106. De Meeûs, T. Revisiting F_{IS}, F_{ST}, Wahlund effects, and null alleles. *J. Hered.* **109**, 446–456 (2018).
107. Levy-Sakin, M. et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).
108. 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**, 189 (2019).
109. Anderson, E. C. & Garza, J. C. The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics* **172**, 2567–2582 (2006).
110. 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).
111. Thompson, E. The estimation of pairwise relationships. *Ann. Hum. Genet.* **39**, 173–188 (1975).
112. Goubert, C. et al. A beginner's guide to manual curation of transposable elements. *Mob. DNA* **13**, 7 (2022).
113. 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).
114. 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).
115. 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).
116. Brown, A. V. et al. Ten quick tips for sharing open genomic data. *PLOS Comput. Biol.* **14**, e1006472 (2018).
117. 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).
118. 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).
119. 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).
120. 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**, 737 (2019).
121. 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).
122. 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).
123. 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).
124. 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).
125. 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).
126. 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).
127. Martins, F. B. et al. A semi-automated SNP-based approach for contaminant identification in biparental polyploid populations of tropical forage grasses. *Front. Plant Sci.* **12**, 737919 (2021).
128. Deo, T. G. et al. High-resolution linkage map with allele dosage allows the identification of regions governing complex traits and apospory in guinea grass (*Megathyrsus maximus*). *Front. Plant Sci.* **11**, 15 (2020).
129. Zhang, F. et al. Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res.* **30**, 185–194 (2020).
130. Christie, M. R., Marine, M. L., Fox, S. E., French, R. A. & Blouin, M. S. A single generation of domestication heritably alters the expression of hundreds of genes. *Nat. Commun.* **7**, 10676 (2016).
131. Lou, R. N. & Therkildsen, N. O. Batch effects in population genomic studies with low-coverage whole genome sequencing data: causes, detection and mitigation. *Mol. Ecol. Resour.* **22**, 1678–1692 (2022).
132. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
133. Mirchandani, C. D. et al. A fast, reproducible, high-throughput variant calling workflow for population genomics. *Mol. Biol. Evol.* **41**, msad270 (2024).
134. 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).
135. Thompson, N. F. et al. A complex phenotype in salmon controlled by a simple change in migratory timing. *Science* **370**, 609–613 (2020).
136. Howe, K. et al. Significantly improving the quality of genome assemblies through curation. *Gigascience* **10**, giaa153 (2021).
137. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
138. Michael, T. P. & VanBuren, R. Building near-complete plant genomes. *Genome Stud. Mol. Genet.* **54**, 26–33 (2020).
139. Tettelin, H. & Medini, D. *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (Springer, 2020).
140. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
141. Hemstrom, W. *Thirty-Four Kilometers and Fifteen Years: Rapid Adaptation at a Novel Chromosomal Inversion in Recently Introduced Deschutes River Three-Spined Stickleback*. Thesis, Oregon State Univ. (2016).
142. 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).
143. 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).
144. vonHoldt, B. M. et al. Demographic history shapes North American gray wolf genomic diversity and informs species' conservation. *Mol. Ecol.* **33**, e17231 (2024).
145. Alonso-Blanco, C. et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
146. Maruki, T., Ye, Z. & Lynch, M. Evolutionary genomics of a subdivided species. *Mol. Biol. Evol.* **39**, msac152 (2022).
147. Kessler, C., Wootton, E. & Shafer, A. B. A. Speciation without gene-flow in hybridizing deer. *Mol. Ecol.* **32**, 1117–1132 (2023).
148. 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).
149. Lowy-Gallego, E. et al. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res.* **4**, 50 (2019).
150. 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).
151. Kardos, M. et al. Inbreeding depression explains killer whale population dynamics. *Nat. Ecol. Evol.* **7**, 675–686 (2023).
152. 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).
153. Robinson, J. M. et al. Traditional ecological knowledge in restoration ecology: a call to listen deeply, to engage with, and respect Indigenous voices. *Restor. Ecol.* **29**, e13381 (2021).
154. Lynch, M. *The Origins of Genome Architecture* (Sinauer Associates, 2007).
155. Lynch, M. & O'Hely, M. Captive breeding and the genetic fitness of natural populations. *Conserv. Genet.* **2**, 363–378 (2001).

Acknowledgements

The authors thank E. Anderson, A. Leaché, M. Kardos and the reviewers for their helpful comments that greatly improved this manuscript. The authors also thank M. Exposito-Alonso and the 1001 Genomes Consortium, the 1000 Genomes Project, B. Hand, M. Freedman, M. Kardos, C. Kessler, M. Lynch, R. Malison, D. Martchenko, M. Miller, R. Schweizer, A.B.A. Shafer and X. Yin for allowing their datasets to be reviewed and re-filtered. M.R.C. was funded, in part, by NSF DEB-1856710 and OCE-1924505. G.L. was funded, in part, by NSF-DOB-M66230.

Author contributions

All authors conceptualized, wrote and edited the manuscript. W.H. and J.A.G. conducted the simulations and analyses in Box 2.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41576-024-00738-6.

Peer review information *Nature Reviews Genetics* thanks Mark Ravinet and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Related links

DNA DataBank of Japan Sequence Read Archive: https://www.ddbj.nig.ac.jp/dra/index-e.html
European Variation Archive: https://www.ebi.ac.uk/eva/
GATK: https://gatk.broadinstitute.org/hc/en-us
NCBI Short-Read Archive: https://www.ncbi.nlm.nih.gov/sra