***In review, please do not distribute!***

**Abstract:**

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

**Figure 3:**

**A diagram of a process

Description automatically generated with medium confidence**

**Box 3: Filtering checklist**

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

|  |  |
| --- | --- |
|  | Data archival |
|  | Decide on filtered data sets given *a priori* study questions |
|  | Create filter recording/reporting table (see Table 2) |
|  | Filter on raw sequences (*e.g.*, read quality, poly-G tails; see Table 1) |
|  | Report exact filters used for filtering on raw sequences |
|  | Report total number of reads in study |
|  | Report total number of reads filtered out |
|  | Perform sequence alignment; report alignment parameters |
|  | Report total number of reads that aligned successfully |
|  | Report total number of reads that mapped uniquely |
|  | Report total number of reads that were filtered out |
|  | Perform filtering on successfully mapped reads |
|  | Filter on mapping quality, PCR duplicates, read pairs |
|  | Report number of reads retained/filtered at each step |
|  | Variant discovery |
|  | Begin or continue creation of multiple data sets |
|  | Decide on study-wide versus within-sample group filters |
|  | Decide on filter values to employ and order of filters |
|  | Locus filtering (see text for when individual filtering should go first) |
|  | Filter for MAF, HWP, paralogs, coverage etc. (see Table 1) |
|  | Individual filtering |
|  | Missing data; Mislabeling/contamination |
|  | Dataset construction; continue reporting of all filters and reads filtered |
|  | Data analysis and parameter estimation |
|  | Report effects of filters on parameters/questions of interest |
|  | Perform re-filtering and/or re-sequencing if necessary |
|  | Final filter recording (report reads, loci, individuals lost at each step) |
|  | Archive all filtered data sets as VCF files |