February 8, 2024

Dear Dr. Koch,

Please consider the attached revision of our manuscript (manuscript # NRG-23-092V3), entitled “Next-generation data filtering in the genomics era”, as a Review in Nature Reviews Genetics. Recent and rapid advances in both short (*e.g.*, Illumina) and long (*e.g.*, PacBio) read sequencing have resulted in the proliferation of genome-wide data. These sequencing methods all result in large numbers of reads that, importantly, are not entirely free of sequencing errors. Furthermore, these sequences are typically aligned back to a reference genome, a process that further increases the error rate especially when genetic variants are called (*e.g.*, SNPs). Most researchers, therefore, perform multiple types of filtering, including filtering on minor allele frequency, minor allele counts, missing data, and deviations from Hardy-Weinberg proportions. This filtering may seem trivial at first glance, but in reality, filtering is an issue of paramount importance because: 1.) every genomic data set must be filtered, often repeatedly, 2.) filtering choices can be confusing, are often subjective, and there are currently no standard or agreed-upon guidelines, and 3.) the same data set filtered in different ways often yields entirely different results; if poor filtering choices are made, all downstream analyses can be affected. Thus, a comprehensive review of different filtering strategies and their downstream effects will be valuable to researchers across diverse fields.

In this revision, we have thoughtfully addressed all reviewer comments, including comments from 3 additional ad hoc reviewers (named in Acknowledgements). Our revision now includes an illustration of filtering effects on ten published empirical datasets spanning mammals, fish, arthropods, and plants and three simulated expanding, declining, and equilibrium (stable) populations (see Box 1). We also present filtering parameter guidelines that researchers can reference, given their specific research questions/objectives. Furthermore, we now provide two RStudio Notebooks as SI files - one for pre-variant and a second for post-variant filtering - to serve as examples for how researchers can easily automate, test, and report the impact of filtering on their own data. Our checklist and guiding principles for filtering will serve authors, reviewers, and even journal editors across disciplines.

Our point-by-point responses are appended below. The attached revision is formatted as a Review Article and contains 5,817 words, 3 figures, 3 tables, and 4 boxes. We also include substantial Supporting Information, with 7 supplementary figures and 4 tables.

Thank you for your consideration.

Sincerely,

A signature on a white background

Description automatically generated

William Hemstrom, PhD

Department of Biological Sciences

Purdue University

West Lafayette, IN 47906

<whemstro@purdue.edu>

Below are the responses to the editor and reviewers’ comments. All of our responses below are in blue font. In the manuscript and supporting materials, all additions are also in blue font; minor wording improvements made to the manuscript during editing have not been illustrated to expedite review. We thank the editor and reviewers for their helpful and thoughtful comments, which we believe have greatly improved the manuscript.

Editor Comments to Author:

While all referees acknowledged the importance of the topic, they suggested that you delve more specifically into ways to illustrate the impact of filtering methods and settings on outcomes. Rather than an overview of filtering criteria and their potential impact, all three referees commented (in comments to authors as well as separately to the editor) that they thought the community would benefit from the inclusion of simulations that illustrate how different filters/thresholds affect various datasets, e.g. stemming from various sequencing technologies, reference genome qualities, or population demographics. One referee specifically suggested providing links to Jupyter notebooks or similar resources that demonstrate the application of some of these filters.

We have done this before, for example, in the article 'Navigating the pitfalls of applying machine learning in genomics' (https://www.nature.com/articles/s41576-021-00434-9), which included interactive notebooks with data and code that can be downloaded and run locally or in a web browser without installing any software. Please note that Nature Reviews Genetics cannot publish what could be considered primary data/results, so it would be important to use pre-existing datasets or run simulations in a purely illustrative way to demonstrate the effects of filters on outcomes/interpretations – as long as the actual data/interpretations themselves are not important, and no new biologically relevant conclusions are reached, it would be fine to include such ‘analyses’.

If you decide to revise your manuscript with this more in-depth approach, we would be willing to consider a revised manuscript version for publication. Please note that the article would likely undergo another round of peer review before a final decision is made.

In response to this comment, and the specific comments from reviewers below, we have substantially revised the manuscript to include illustrations for how different filters and filtering thresholds can affect various data sets. We have now increased the number of filters and filtering thresholds illustrated in Box 1 by increasing including more datasets (10 empirical and 3 simulated), filters (MAF/HWP/missing data, etc.), and outcomes (11 descriptive statistics). These additional contributions allowed us to now visualize the general trends associated with specific filters (Box 1) as well as provide general filtering threshold recommendations (e.g., Figure 3, Table 1). We do not discuss biological interpretations of specific studies but instead discuss the trends caused by changing filtering thresholds across datasets.

We have also now added two R markdown notebooks (Supplementary notebooks 1 and 2), which are set-up to automatically download the relevant data from the NCBI, which provide an example workflow, from reads to genotypes, that allows for the easy modification of filtering thresholds. These workbooks should be simple to modify to fit new data, and we hope will aid researchers more easily compare filtering thresholds in their own datasets. Please see below for additional details.

Reviewer comments to author:

Referee #1:

This review report was conducted by a PI (who hasn’t filtered genomic data for over 5 years) and two senior members of the laboratory who are heavily involved in filtering lcWGS datasets. The amount of next generation sequencing data being produced has increased exponentially over the last decade. However, guidelines for analyzing these datasets have not kept up with the rate of data production. Therefore, this paper is timely and is an important contribution to the field of population genomics. The authors outline different steps and stages of the filtering process with specific methods for filtering during pre- and post-variant calling. Their outline of these methods will be extremely useful for students, postdocs, and other early career scientists that are trying to identify appropriate ways to filter their datasets. In addition, the authors provide a nice comparison of the filtering effects of three datasets from different species. That all being said, the paper gives a good overview, but by the end of the paper we felt that the “guidelines” we were expecting from the introduction were lacking. The main takeaway from the paper was to filter your data multiple ways and get to know how filtering affects your dataset. We do agree with this statement and furthermore, we feel understanding your system and species is of the utmost importance when determining filtering criteria. However, the way the paper is set up, the reader gets the sense that there will be extensive testing of datasets or simulations to help guide which parameters will be useful under different scenarios stemming from either species/system specifics, sampling specifics, or sequencing specifics (e.g., different population demographies, population sizes, sample sizes, different sequencing technologies, quality of the reference genome, etc.). After reading the paper, it’s clear that this was not the goal of the paper, but this should be reflected in the way it is set up. One other thing that was severely lacking was encouragement to understand one's study system. The context of life history and previous genetic studies is absolutely vital to the process of analyzing genomic datasets and this is getting less emphasis than it should. This paper represents a great opportunity to emphasize the importance of knowing one's system (more on this below). Finally, it’s commendable that the paper is advocating for reproducibility and data documentation. The field has been moving more and more towards requiring better documentation and more papers stressing the importance of data archiving in reproducible ways are needed. We recommend that the paper be rejected in its current form with encouragement to resubmit after addressing the issues raised.

We greatly appreciate the positive comments suggesting this paper is timely and will be “extremely useful for students, postdocs and others”. We especially appreciate the helpful advice to improve the paper; we have now incorporated all suggestions below. Specifically, we have now systematically filtered both previously published empirical datasets and simulated datasets that we generated, to generate general guidelines for filtering parameter values (Box 1, Table 1, Supplemental Table S4). We also agree that understanding of one’s study system and species is of utmost importance when determining filtering criteria and we have now added an entirely new Box to address this point (Box 3 in resubmission) – see also below. We additionally discuss the benefits of reproducible research in the realm of filtering on lines 446 – 459.

Major comments

The paper is more of an overview of the different methods of filtering rather than in an in-depth review of how filtering parameters should be applied. Specifically, we thought the comparison of the three datasets would be the crux of how the paper tackled their proposal of key best practices and standardization of filtering approaches. However, those data sets were used to only explore how MAF filters and the amount of missing data affect certain analyses (i.e., Tajima’s D, proportion of segregating sites, observed heterozygosity, and FST). A better way to assess and answer this question would be using a sensitivity analysis of different parameters to see which parameters most affect different datasets.

We agree and have now substantially modified the manuscript in the following ways: 1.) We now have analyzed three simulated datasets alongside 10 published empirical datasets (spanning mammals, fish, arthropods, and plants) to explore how a single filter can affect multiple downstream parameter estimates, how multiple filters can affect the same parameter, and how higher-order interactions among filters can affect downstream analyses (Box 1; lines 1000 – 1042). Specifically, we illustrate that MAF filtering choices affect *F*ST, *F*IS, Tajima’s D, and observed heterozygosity in similar ways (Box 1, panel A). We also illustrate that filtering by Hardy-Weinberg proportions, missing data, and MAF affect the proportion of segregating sites (Box 1, panel B). These results illustrate the effects of single and multiple filters, respectively. In the SI, we systematically illustrate the effects of MAF, HWP, and missing data filters against a wide range of summary statistics.

We feel like this opening statement was not relieved after completing the paper: “Researchers are confronted with a multitude of choices when filtering genomic data; they must choose which of the many available filters to apply and select appropriate thresholds.” Specifically, the appropriate thresholds are going to be study-specific and some guidelines on thresholds across a range of datasets would be beneficial. In addition, although we think Table 1 is useful for outlining all the different filters that can be applied, a large proportion of these filters are not mentioned in the manuscript apart from this table. We think as a new researcher looking at this table we would still feel overwhelmed and unsure of what filters/thresholds would be appropriate for analyses of different data types (RAD, lcWGS, etc.).

We agree that a new researcher may have come away from this first draft feeling overwhelmed. In response, we have taken several explicit steps: 1) we have shortened/condensed our explanation on the potential effects of filtering, 2) we now test and mention a larger proportion of filters and have moved the former Table 1 to the supplementary material (now Supplementary Table 1), and 3) we now provide much more explicit guidelines for filtering across a range of datasets. We have now added a table detailing conservative filtering thresholds for a range of possible filters across a suite of research questions (Table 1) designed for researchers to detect an effect, if present, within their datasets. We have also updated the flowchart (Fig. 3) to start with the questions first and then provide a series of specific guidelines. Justifications for our filtering recommendations are contained in Supplementary Table S4. While we do not address lcWGS data directly in this manuscript, we do refer the reader to lcWGS resources (lines 150 – 153).

One thing we feel was a major omission was discussing the importance of researchers understanding their study system. We would like to see at least a paragraph describing how important it is to understand your study system and have a good feel for expected results. Do you expect to see low population structure because it’s a marine organism so observing high population structure would be atypical and potentially an artifact of data processing? Do you expect to see high inbreeding because of a certain mating system? Do low or high Ne estimates make sense in the context of the demography of the animal? What have previous genetic studies in the system with last generation markers suggested? Understanding this context is extremely vital for ensuring that results are biologically reasonable and accurate. Seasoned PIs often know when something is wrong even when they aren’t in the details of the filtering. Why is that? Fred Allendorf has bemoaned the fact that genomic researchers are now almost ubiquitously excellent programmers and bioinformaticians but are losing touch with population genetics theory. And we could see a similar trend occurring when it comes to knowing one’s study system. Every time we start a project we urge the analyst working on it to start by reading up thoroughly on the life history of the species and reading last generation genetics papers on the species. We think this is absolutely vital and we have seen numerous examples of where this has not occurred and has led to major issues in data interpretation. To this point, when we’re reviewing work in our lab or manuscripts we are constantly asking ourselves “do these results make sense in the context of the study system?” And “if not why not?” Diving into the filtering methods is obviously important but we still believe that the another crucial way to identify erroneous results is to understand your study system. We want to make sure this doesn’t get lost as the focus on bioinformatics increases.

We wholeheartedly agree. In response to this helpful comment and ideas mentioned directly above, we have now added several paragraphs on the importance of study system knowledge in a new box, Box 3 (lines 1065 – 1081), where we discuss the importance of understanding one’s study system, both in terms of life history and ecology, but also in terms of prior studies, genetics, and demographic history. We also include new simulated data in Box 1 illustrating the effects of past demographic events (Box 1, panels A and B). For example, we simulated both growing and declining populations and now show that stringent MAF filters (removing rare alleles) can completely change demographic inferences – e.g. causing a growing population to appear to be declining (e.g., Tajimas D in Box 1 panel A).

Figure 2 is helpful for highlighting the challenges of four common filters, but the associated text really lacks an assessment of how to appropriately apply these filters. Again I think this relates to the scope of what the paper is trying to convey. Is it guidelines for these filters or is it an overview of the challenges of different filters applied to different datasets?

We would like for this manuscript to be both an overview of the challenges of different filters and guidelines for filtering. The first two figures are to illustrate the former – the challenges of different filters. In response to this comment and the comments above, we now have shortened the overview of the challenges of different filters and expanded and clarified the second half of the manuscript focusing on specific guidelines. We point the reader to these guidelines that are now presented in multiple places: Tables 1-3, Figure 3, and Supplementary Table 4.

It should be made much clearer in the introduction that the authors advocate for a thoughtful filtering strategy that matches the trade-off in type I and II error when evaluating their hypotheses. It is unclear in the introduction if the authors are advocating for ‘comprehensive sequence data filtering standards’ that should be implemented within the context of their flowchart or if a population genetics theory driven method that weighs the pros and cons of different filtering methods is preferable. The discussion gives the impression that the authors favor the latter. The references are appropriate, but the overall conclusion that researchers need to think critically about their study objectives and tailor their data filtering to balance type I and II error for the hypotheses being tested needs to be clarified in the abstract/intro.

We agree and now provide specific filtering recommendations that are designed to lead researchers down the path to first seeing the effects of filtering (if any) – by considering both population genetics theory and knowledge of their study system to prioritize which filtering effects to test for, and then thinking about the causes of those downstream effects. We suggest researchers should report effects of key filtering choices (on downstream parameter estimates) and justifying their choices of filter made in the final set of analyses (lines 121 – 126; 349 – 355).

In terms of population genetics theory, it should be noted that the authors suggestion of removing loci that deviate from Hardy-Weinberg proportions prior to analyzing their data with STRUCTURE is misguided. By removing loci / individuals that cause violations of HWE the STRUCTURE algorithm would be essentially nudged to find the sampled collections. And in general we feel that advising removing loci based on HWP may not be the best approach. Looking at FIS distributions to identify a cutoff based on multiple modes in the distribution could help with paralogs, but Hdplot could do the same. We think it might be better to advise researchers to examine HWP and FIS in interesting regions they identify after they identify them rather than advocating an HWP filter.

We agree and have reworded this section to think about why deviations occur, rather than removing HWP/FIS-violating loci without further thought (lines 300 – 308).

We found Box 1 to be the most interesting of the visual aids, but its discussion in text is severely lacking. It appears that there are 4 groups of Monarch butterflies (structure plot – 3 mostly pure and 1 admixed) but only three lines plotted in A and B? In general the discussion of this box felt rather cursory. We think quite a bit more detail is necessary. Alternatively, if the authors are simply trying to show that filtering impacts downstream analyses they could focus on one aspect like MAF. Also it should be noted somewhere that these are all RAD datasets. It took looking up the references to figure that out.

We have now substantially revised Box 1 in light of these comments. In particular, between Box 1 and the Supplementary Material, we now include 10 empirical datasets, four different filters, and three additional simulated datasets. This should provide a much more detailed portrait of how filtering systematically affects SNP genotype datasets.

We really liked the idea that different analysis methods should use differently filtered datasets as input (this is super important) but we’d like to see more specifics on this topic. The authors went into this a little bit, but more on what type of filtering approaches and parameters matter for what analyses would be very useful and we think would add some good “meat” to the paper.

In response to this comment, we have now provided a table showing our filter threshold recommendations for different types of analyses (Table 1) and discuss many of the specifics more explicitly throughout the paper. Our more extensive simulations and the discussion on the results (Box 1) should also aid with this. These recommendations are also now explicitly included in the flow chart (Figure 3). However, we also caution readers that this and any review cannot fully quantify effect of all filter types for all study systems and thus knowledge of the study system, population genetics theory, and assumptions of the downstream estimators and software are crucial for choosing re-filtering thresholds to test (Box 3).

We are all for making sure filtering methods are understood, scrutinized and well documented but we have also seen data filtering become a quagmire for many scientists, especially early career scientists. We think that at the end of the day, investigators need to provide good documentation of the filters they chose and have good reasons for choosing them. We also think it is important to recognize that asking authors to reanalyze their data with multiple filters during the review process should not be done casually. Often first authors are students who may have moved to new jobs and asking for something like this could be the difference between something getting published or not. We think that this paper would be a nice place to make a plea to reviewers to be thorough but reasonable when it comes to asking authors to reanalyze their data with different parameters. For example, if reviewers see a result that doesn’t match up to expectations the authors should provide reasons why that is the case. But if results generally match expectations based on previous work, asking authors to filter the data many many ways to most likely get the same answer seems like a waste of time. To be clear, we agree that confirming conclusions with multiple datasets is a good idea, but we want to make sure that reasonable bounds for data exploration exist so researchers aren’t forced into unnecessary data filtering exercises to get things published. Hopefully this paper moves the needle forward in terms of coming up with best practices that both authors and reviewers can use to ensure that reasonable data filtering approaches are followed but analyses do not become overly onerous. But we still think pointing out the potential quagmire of data filtering and the need for reviewers and journal editors to be fair and reasonable would be a good addition to this paper.

We agree and have given this comment a lot of thought. In light of our revised analyses provided in Box 1, we now provide some clear guidelines while simultaneously emphasizing that any filtering regime must be applied in a thoughtful way while considering the research question and the intricacies of the study system (Box 3). Specifically, we have remade Figure 3 to start with the Questions and Objectives, such that new students can follow through the flow chart with their specific objective in mind. It is unlikely that all studies will have all questions/objectives and so this approach will hopefully make the idea of filtering less overwhelming. We also suggest with this figure that researchers start with the understanding that building a reproducible workflow from the start will help them repeat analyses both before and after reviews (if required). Importantly, we highlight that such workflows will also help future researchers in the local PI’s lab or other labs to use the same analytical and filtering workflows when new data are produced. We now also mention that reviewers should be reasonable when asking authors to reanalyze their data with different parameters, and authors can sometimes simply discuss and justify their choice of filters and thresholds (e.g., if their question is unlikely affected by a certain filter; lines 456 – 460).

Minor comments

lcWGS is gaining popularity quickly. I think a few sentences on specific considerations for lcWGS and how they might differ from higher coverage data would be warranted.

We agree; given that some aspects of lcWGS filtering can be quite different we now refer the reader to an existing paper which discusses filtering in lcWGS data (lines 150 – 153).

Line 25 – standardized approaches are great… until they aren’t. This should be caveated that the analyst should use their head when making decisions because they’re data may not be standard and a standardized approach may not apply.

We mention that authors need to think critically about filtering choices, and that the effects of filtering do not bias downstream results (lines 345 – 346). We also highlight the importance of thinking critically about filtering choices in the new Box 3 (lines 1061 – 1077).

Ln 29 – ‘do not all have’ = ‘have variable error rates’

Updated (line 36).

Line 29 – ‘In addition to the inherent error rate of the genotyping chemistry, errors can arise when…’

Updated (lines 36 – 37).

Line 38 – ‘Genomic filtering’ you aren’t filtering the genome but the data that may be aligned to a genome.

This sentence has now been deleted.

Line 44 – ‘imputation, modify the data’ imputation adds data to individuals where it may be lacking based on observed patterns in other individuals. Modify the ‘data set’.

We changed to “dataset” (line 51) and provide the reader with a further in-depth description of “imputation” in the glossary (line 904).

Line 47 – 3. could add that the explicit filtering decisions have unknown consequences on downstream applications

Agreed. Added at line 55.

Line 54 – ‘limited standardization of filtering efforts’, one shoe fits all is not always the best. Emphasize importance of looking at data

We agree this is important and now point out the importance of thinking critically about filtering choices in Box 3 and lines 345-346. We also changed our wording here slightly away from “standardization” (line 60).

Line 60 – I would change methodologies to approach

Updated (line 67).

Line 83 – ‘minimize’ instead of avoid?

Updated (line 89).

Line 91 – Change exhaustive to something like “as comprehensive as possible” as no table will be exhaustive

Changed to “comprehensive” (line 99).

Line 109 – ‘calling confidence’ = expected base calling accuracy?

31% should be 68.4% accuracy. Phred scores are error probabilities so you have to subtract from 1 to get accuracy

P(Incorrect Base) = 10^(-Q/10)

P(Correct Base|Q5) = (1-(10^((-5)/10)))\*100=68.37%

P(Correct Base|Q40) =(1-(10^((-40)/10)))\*100=99.99%

You are correct. We double checked the math and changed the notation to “maximum allowable error rates” (line 113-114).

Line 169 – ‘samples’ refers to individuals within collections for a study? MAF would be across the full dataset?

We changed “samples” to “sample-groups” (line 179).

Line 182 – Any guidance on which type of p value correction might be best for certain situations?

We feel that this is outside the scope of this paper, but now refer the reader to a study that goes into further depth on Hardy-Weinberg testing (lines 216-217).

Line 189 – comma at end not needed

Updated.

Line 210: add “which is often likely to be untrue” after “are themselves unlinked).” With some RAD techniques like bestRAD sequencing happens in both directions from the cut site so adjacent sites are likely linked. Keeping one SNP from each tag does not address this linkage.

Updated (lines 232 – 233).

Line 219 – We’re a bit confused by this sentence. Can you elaborate? We think you may mean something like filtering based on missing data multiple times to iteratively remove low quality individuals and loci?

We added a short example to clarify what we mean here—some filters, like mapping quality in our example, can be applied at different steps in the analytical pipeline and may actually be applied multiple times (lines 240-241). We also later discuss individual vs. locus removal order during post-genotype filtering, which is another example of the same filter being applied more than once (lines 405-413). We have chosen to keep our discussion of such filtering workflows short, as they quickly become intractable (but we think it is useful for readers to know that they exist).

Line 254 ‘SNPS’ should be SNPs

Fixed.

Line 257 - ‘likely caused by recent population expansions’. We’d like to see more detail on this. Maybe the authors could apply multiple filters and show if this is a reasonable explanation?

We clarified this statement (line 279) but want to avoid delving into this example too much since we feel like it to be beyond the scope of this review.

Line 304 – We think ‘follows from this’ is not a great phrase to put in a topic sentence

We agree and have removed this sentence.

Line 312 – I think what the authors are trying to get at is that there is a tradeoff in type I and II error and we have to think critically about the balance between these when reporting novel findings. This should be emphasized more.

We now mention ideas related to this theme in lines 354 – 356 (“No filtering method will remove all errors, but re-filtering with different thresholds can provide higher certainty that there is not substantial bias or error from filtering”), and 214 – 217 (“and, ultimately, different approaches and alpha thresholds should be applied depending on the questions being asked and the tolerance for including or excluding problematic loci”).

Line 317 – We totally agree that there is no “best strategy” and recommend emphasizing this early and often in the paper including in the abstract

We now state in the abstract “Filtering effects can be unpredictable and there is no single best strategy for filtering all genomic datasets” (lines 28 – 29 and 352 – 359).

Line 332 - I would specify archived privately. Data are expensive to generate and the lab that generated them should have the first shot at publishing them.

Good point. We have now added “(privately or publicly)” on line 375.

Line 385: what about recommendations for tests for relatedness? That would manifest in a PCA with outlier clusters. We think it is important to add here that the researcher should figure out why individuals look strange rather than just removing them.

Good point. We now address this on lines 418 – 421.

Line 427 and throughout the manuscript: We’re not sure if standardize is a good term to use in the paper because what the paper is advocating for and we would advocate for is to explore your data and use filters that make the most sense for your data. Therefore, we’d say that the filtering process is not meant to be standardized, but rather the documentation of filtering steps (and some shared approaches) should be standardized.

We very much agree—the term “standardization” is dropped from most places in the text (it is only now in reference to our recommendations for filter threshold reporting).

Line 429: take out “for the same dataset”

Updated.

Ln 810 – ‘improvements’ assumes knowledge of the correct answer. Maybe “potentially increased resolution”

This is a good point, but Box 1 has been substantially changed and this text no longer exists.

Line 966: homozygosity should be heterozygosity

Thank you for the detailed check of the glossary—we have fixed this typo (line 984).

Box 2 may be better represented in the text instead of as a box.

We think that it fits slightly better as a box, since attempts to place it in the main text felt jarring, but are happy move this box to either the main text or the supplementary material after reading our revision.

Referee #2:

Overall, I think is a good, timely and important reference paper. It is well written and clearly presented. The figures in particular very nicely illustrate the issues and workflow in a succinct way that a reader can easily grasp.

For much of the manuscript, I found myself nodding vigorously while reading. I completely agree that attempts to standardise filtering are necessary. This is especially true if we are to begin to leverage the large number of genomic datasets for comparative analyses to answer really big questions. This MS provides an excellent summary and will act as a good jumping off point for those wishing to address these issues in their data. I don’t really have any major comments beyond a couple of relatively simple points that I think could be expanded on or revised.

I think one point to recognise that is not addressed is that a lot of the recommendations are based on being able to clearly define groups a priori. But it is often not that straightforward to define groups clearly or easily. The freshwater fish species focused on in the main text, the Yellow Perch, is restricted to lakes and so there is an obvious discrete population boundary. But this is not clear for many other species where there is no clear break in distribution or structure. Indeed, many of the researchers this MS will target might be sequencing in order to investigate population structure that could be informative for group identification. I think it would be very helpful for the MS to address this point and perhaps give a suggestion for researchers in this situation.

We agree and have now added a short paragraph that focuses on populations without discrete boundaries and how that would affect filtering when one cannot clearly define groups *a priori* (lines 308 – 314).

Another point I think that is not clearly addressed is the possibility of interaction among filters. For example, setting a GQ or depth threshold might mean that some genotypes are masked as null and then this makes a variant drop out due to missing data thresholds. I recognise that investigating this is difficult but I think it warrants a mention, especially in the best practices section.

We agree and have now clearly illustrated interactions between two filters (Box 1; panel B). We also discuss the possibility of higher-order interactions among filters and that the sequence of filtering events can also add to the challenges associated with filtering (lines 1027 – 1030 and 405 – 413, respectively).

Finally, I think there needs to be more space given to the possibility of reproducible workflows and ways of reporting. There are now many tools available for this – i.e. RMarkdown, Nextflow, Snakemake. There is already a hint towards reproducible workflows at the end of the MS (LN407-413) but I think there is space to really integrate this approach to filtering right from the start. Clearly a move to more reproducible workflows will address many of the issues this MS focuses on and I think it would be better to mention this earlier and throughout.

We agree and now mention the idea of reproducible workflows much earlier and more often in the manuscript (lines 31, 70, 374, 442-460; see also Tables 2 and 3). We have also added two R markdown (Supplementary Notebooks 1 and 2) notebooks which implement an example of read alignment, SNP calling, and VCF filtering that allow for easy re-running with new filter thresholds. These notebooks are presented as an example workflow—they require only minor path tweaks to run with new data and are clearly commented to allow for more extensive tweaks depending on research needs.

Anyway, the rest of my comments are minor, hoping to improve and clarify where needed:

LN29: “do not all have low error rates” - quite a clumsy way to put this – surely a number of reads with non-negligible error rates or something similar would be preferable?

We have rephrased this sentence to be clearer (line 36).

LN137: I think it would also be useful to note here that higher-than-average depth can also be informative – i.e. highly repetitive regions. Authors should not only filter low-coverage sites, but also be wary of high coverage. In certain pipelines, like STACKS, this is explicitly integrated.

Agreed. We now mention this explicitly at lines lines 141 – 144.

LN174: Why HWP and not HWE – the more recognisable and widespread acronym?

HWE (Hardy-Weinberg Equilibrium) is a bit of a misnomer and perhaps this manuscript could be a good place to change how the field thinks about this nomenclature. Please see Waples (2015)—located at <https://doi.org/10.1093/jhered/esu062>—for a more in-depth discussion. We are happy to change to HWE, if the reviewer thinks it more appropriate.

LN220: Might be helpful to highlight a specific example in the text here too – MAF is a good candidate here and this is consistent with the main focus of the later text.

We agree and have added an example using GATK’s hard filters, which we essentially never see explored or varied from their defaults. These filters are a bit more abstract than MAF, and so seem to fit here a bit better (lines 246 – 248).

LN225: Delete nonetheless – redundant with however at the start of the sentence.

Removed.

LN252-254: The way this example is initially presented is a little confusing – when I first read it I initially expected the conclusion to be that within-group filtering introduces biased. Perhaps it would help to flip the example – i.e. state a range of SNPs per population and then show that whole group filtering is more restrictive with 714K SNPs per pop.

We’ve reworded this section to be clearer according to this suggestion.

LN316-317: It is potentially also worth mentioning here that there is no filtering method that will remove all errors but rather instead produces a dataset where we can be more certain that there is not a substantial bias from error.

We have now added quite a bit more detail here in line with this comment (lines 354 – 359).

LN400-402; LN783-785: Is it really necessary to archive all VCF files? Why not a single, unfiltered VCF and a script that can filter for each of the different analyses? This would still allow reproducibility and would be more efficient in terms of storage. It is not clear to me what is meant by limited tool availability and why this would mean it is necessary to store all the filtered outputs.

Storing just the core VCF file and a set of filtering instructions is ideal in the long run, but tool versions can change filtering syntax or approaches and user error during future reproduction efforts are not unlikely. Storing the actual compressed, filtered files is still probably ideal. Storage concerns are absolutely valid, but we nonetheless feel like the “best practice” is nonetheless to store everything. We would be happy to include a brief note of this in the manuscript if helpful.

Box 1 and figure: Is there a good reason why the cutthroat trout data is so variable with the differences in filters? It seems Fst and Tajimas D are especially vulnerable to missing data filters in this dataset. It seems a shame not to explore this or at least mention it in the box (although I recognise there is limited space) – it might provide insight to readers.

We no longer use cutthroat data in the current revision—as we already had too many fish examples and this dataset was fairly low-quality and messy.

LN802-804: Perhaps I am misreading the figure but I do not see this transition at MAF 0.01 – do you mean MAF 0.1? Also at least one of the cutthroat datasets appears to approach zero but not increase above it.

Please see comment above and the new/improved Box 1.

Referee #3:

Overall, I think the paper is well written, well structured, and will be useful to many researchers. Besides one specific issue detailed below, I did not find any technical issues. My main high-level issue is that I would like to see a larger emphasis on how filtering can unintentionally bias the data, especially when considering under-represented groups. The simple example would be to remove rare variants using an AF reference that includes mostly European individuals on a set of samples that includes people of African origin. These issues are discussed, but they are not held out as top-level issues.

We have now substantially revised Box 1 in response to this comment and similar comments from reviewers. In particular, we now include human data as one of the 10 empirical datasets that we analyze (Box 1). We also now mention that using reference alleles from one population (e.g. Europeans) to remove rare variants in data from another populations, e.g., non-Europeans or minority ethnic group, can unintentionally bias the data (lines 291 – 296).

I liked the examples and I think they are helpful to people. It would be even more helpful if the authors include links to Jupyter notebooks (or something equivalent) demonstrating the process of applying some of these filters.

We agree and have now added two R markdown notebooks (Supplementary notebooks 1 and 2) which provide repeatable filtering workflows for both pre- and post-variant calling filters. Hopefully these will be useful to researchers looking to make repeatable, easily altered filtering workflows for their labs.

…since high depth of coverage allows for more confidence in genotyping (and subsequently downstream inferences)...

This is not true, “good” or “expected” depth of coverage allows for more confidence in genotyping. We often see some sites with extremely high coverage where we do not trust the given genotype b/c the site likely violates the genotyper’s ploidy assumption.

Agreed! We have now added more detail to our discussion of depth filtering to address this point (lines 141 – 144)

Loci (typically SNPs) for which the less frequent allele (that is, the minor allele) occurs below a certain frequency are also often filtered out. This is clearly experiment specific and also depends on the number of samples and the diversity of samples in the reference. If the target sample is at all divergent from the reference population (even something quite small like differing ancestry), the AF could be misleading. This issue holds true for the rest of the population filters suggested here. I think some discussion must be included before these population filters that warns users of the issues of bias in our population datasets. The authors discuss some of these issues in a subsequent section, but in a paper like this a reader may not get to the details that are not relevant to their needs.

We mention reference bias and note issues with aligning to more distantly related genomes on lines 131 and 84-89 respectively. We also provide an example of the risk of generating misleading site frequency spectra when filtering on MAF using divergent populations more extensively on lines 192-198.