Batch effects in population genomic studies with low-coverage whole genome sequencing data: causes, detection, and mitigation

R. Nicolas Lou[1]* and Nina O. Therkildsen[1]

1. Department of Natural Resources, Cornell University, Ithaca, New York, USA

* Corresponding author: RNL rl683@cornell.edu

Abstract

Over the past few decades, the rapid democratization of high-throughput sequencing and the growing emphasis on open science practices have resulted in an explosion in the amount of publicly available sequencing data. This opens new opportunities for combining datasets to achieve unprecedented sample sizes, spatial coverage, or temporal replication in population genomic studies. However, a common concern is that non-biological differences between datasets may generate batch effects that can confound real biological patterns. Despite general awareness about the risk of batch effects, few studies have examined empirically how they manifest in real datasets, and it remains unclear what factors cause batch effects and how to best detect and mitigate their impact bioinformatically. In this paper, we compare two batches of low-coverage whole genome sequencing (lcWGS) data generated from the same populations of Atlantic cod (*Gadus morhua*). First, we show that with a "batch-effect-naive" bioinformatic pipeline, batch effects severely biased our genetic diversity estimates, population structure inference, and selection scan. We then demonstrate that these batch effects resulted from multiple technical differences between our datasets, including the sequencing instrument model/chemistry, read type, read length, DNA degradation level, and sequencing depth, but their impact can be detected and substantially mitigated with simple

bioinformatic approaches. We conclude that combining datasets remains a powerful approach as long as batch effects are explicitly accounted for. We focus on lcWGS data in this paper, which may be particularly vulnerable to certain causes of batch effects, but many of our conclusions also apply to other sequencing strategies.

Introduction

The field of population genomics has been strongly influenced by two major advances over the past decades. First, high-throughput sequencing technology has rapidly evolved, steadily lowering the cost of DNA sequencing (Costello et al., 2018; Elango, Banaganapalli, & Shaik, 2019; Slatko, Gardner, & Ausubel, 2018). Second, the importance of reproducibility and reusability has increasingly been recognized by researchers, journals, and funding agencies alike, making data sharing an integral part of modern science (Gewin, 2016; Lowndes et al., 2017; Wilkinson et al., 2016). As a combined result of these shifts, a plethora of sequencing datasets from previous population genomic studies across the tree of life are now publicly available (Benson et al., 2013; Field et al., 2009; Kaye, Heeney, Hawkins, de Vries, & Boddington, 2009; Perez-Riverol et al., 2019).

This availability of pre-existing data brings new and exciting opportunities for combining datasets to achieve unprecedented resolution in empirical population genomic studies, e.g. with larger sample sizes, greater geographic coverage, or incorporation of temporal replication or time series analysis (see De-Kayne et al., 2021 for a recent review). As population genomics increasingly moves towards re-sequencing of entire genomes rather than specific sets of markers, the potential for combining datasets across studies should further grow. Yet, sequencing technology has evolved and diversified so quickly over the past decades that most datasets are likely to differ in various aspects, such as the library preparation method, sequencing platform, read type, and read length. In addition, DNA quality and depth of sequencing will often vary among different batches of data. These technical differences among datasets may create batch effects that can confound real biological patterns (Leek et al., 2010).

Typically, the best way to limit these technical artefacts is to keep data generation as consistent as possible across batches of samples (e.g., adhering to the same sequencing platform, read type, and read length) and/or to randomize samples from different groups (e.g., populations or time points) across different sequencing batches. However, neither of these options may be available when we combine pre-existing datasets (De-Kayne et al., 2021). When we supplement pre-existing datasets with new data, full randomization of samples is also not possible since we do not have control over which samples are included in pre-existing datasets. With sequencing platforms being gradually phased out (e.g., Illumina's recent discontinuation of the HiSeq platform), even generating new data with the same configurations as in pre-existing datasets may not be an option (De-Kayne et al., 2021; Leigh, Lischer, Grossen, & Keller, 2018). Furthermore, in addition to sequencing configuration, factors such as DNA degradation can also lead to batch effects that cannot be controlled by experimental design alone. We therefore need post-hoc bioinformatic approaches to detect potential batch effects in our data and mitigate their impact.

Although most researchers are aware of the potential risk of batch effects in sequencing data, only a few studies have explicitly discussed the causes and consequences of such issues, and even fewer have explored bioinformatic approaches to address them with real data. For example, Bálint et al. (2018) demonstrated that variation in DNA extraction and PCR protocols may lead to batch effects in eDNA studies, but they did not suggest any bioinformatic solutions to mitigate their impact. O'Leary et al. (2018) illustrated that differences in library preparation protocol and sequencing coverage may lead to batch effects with restriction site-associated sequencing (RAD-seq) data, but it is unclear whether their recommend mitigation methods would be applicable when samples are not randomly assigned to different batches. Leigh et al. (2018) provided one of the most thorough analyses of batch effects to date, demonstrating that differences in read lengths in time-series RAD-

seq data can lead to false signals of allele frequency shifts, but that stringent SNP filters, indel alignment, read trimming, as well as a species-specific reference genome can be effective remedies. Similarly, Kofler et al. (2016) examined how differences in read length and insert size could affect mapping performance with Pool-seq data, and they showed that intersecting results from two different mapping tools is an effective approach to reducing batch effects when working with datasets with differing read length and insert size. Most recently, De-Kayne et al. (2021) provided a broader conceptual overview of different causes of batch effects in sequencing data and best practices to address them, particularly highlighting important consequences of the shift from a four-channel to a two-channel sequencing chemistry on Illumina platforms. However, it is still unclear how such impact manifests in real data (but see Arora et al., 2019 for an example) and how effective their recommended mitigation methods are in practice.

In this paper, we present an empirical case study of batch effects in low-coverage whole genome sequencing (lcWGS) data. Whole genome sequencing is arguably the sequencing method that harbors the greatest potential for reusing and integrating datasets, since the ability to combine across studies does not hinge on selection of the same restriction enzymes (as in RAD-seq) or markers (as in SNP chips or microsatellites) (De-Kayne et al., 2021). In particular, as a powerful and cost-effective approach to obtain whole-genome data, lcWGS is becoming increasingly popular in the field of molecular ecology (Lou, Jacobs, Wilder, & Therkildsen, 2021), but its sensitivity to batch effects has not yet been examined. Here, we show how combining lcWGS datasets that differ in multiple ways can result in severe batch effects in downstream population genomic inference, and we highlight strategies for detecting and mitigating such impact using simple bioinformatic approaches. Although lcWGS data may be especially susceptible to certain causes of batch effects investigated in this paper because of the higher level of uncertainty in this data type, many of our

conclusions should also apply to other types of sequencing data, including high-coverage sequencing.

Materials and Methods

The data presented here originate from a lcWGS study of population structure and adaptive divergence in Atlantic cod (*Gadus morhua*) in Greenlandic waters (Lou et al. in prep). DNA was extracted from fin clips or gill tissue with the Qiagen DNeasy Blood & Tissue Kit and libraries were prepared with the protocol described in Therkildsen & Palumbi (2017). There was substantial variation in the preservation level of tissue samples from this difficult-to-sample locality. For a subset of our samples, the DNA was relatively well-preserved so we could prepare libraries with a sufficient insert size to make full use of cost-effective paired-end sequencing. For another subset of samples, however, the DNA was so degraded that the average insert size we could achieve in our libraries was only 100-150bp, and accordingly, paired-end sequencing would lead to substantial redundancy among the overlapping read-ends and adapter read-through (resulting in loss of >50% of the data). Naively unaware of how severely it would affect our analysis downstream, we decided to split our samples into two different batches for sequencing: one batch with single-end 125bp read (for short-insert libraries; we will refer to this as "HiSeq-125SE") and the other batch with paired-end 150bp reads (which was more cost-effective for libraries with longer inserts; we will refer to this as "NextSeq-150PE"). As the names imply, the two batches of data were sequenced on different Illumina platforms, so they differ in their sequencing chemistry (HiSeq 2500 with a four-color chemistry vs. NextSeq 500 with a two-color chemistry), and they were also sequenced to different average depth of coverage per sample (0.8x vs. 0.3x, see overview in Table 1).

To assign different samples to the two sequencing batches, we used gel electrophoresis to visually assess the level of degradation in all DNA extracts. We categorized samples with strong low-molecular-weight smears on a gel as “degraded” and sequenced these in the HiSeq-125SE batch. The majority of the remaining samples (well-preserved) were assigned to the NextSeq-150PE batch, but to fill up lane capacity, a subset of the well-preserved samples were sequenced in the HiSeq-125SE batch.

A total of 388 individuals were sequenced in these two batches, but we base our analysis here on a subset of 163 individuals from 9 populations for which individuals were split between the sequencing batches. For 3 of the 9 populations, the samples were split strictly based on their degradation level (i.e. the degraded samples were all sequenced in the HiSeq-125SE batch and the well-preserved samples were sequenced in NextSeq-150PE batch). For the other 6 populations, all samples were well-preserved and were randomly split between the two batches (sample sizes and depths of coverage in Figure S1). Since we do not expect there to be systematic biological differences between individuals from the same population, we have multiple independent sets of comparable samples split between batches, which allows us to assess the effectiveness of our bioinformatic mitigation strategies, both for well-preserved and degraded samples.

A detailed description of our entire data analysis pipeline is included in the supplementary materials, and all scripts used in this paper are available on GitHub (https://github.com/therkildsen-lab/batch-effect). Briefly, we first processed all samples with a standard bioinformatic pipeline for lcWGS data without explicitly taking the differences between the two sequencing batches into account (i.e., a “batch-effect-naive” pipeline). For data filtering and mapping, we used Trimmomatic-0.39 (Bolger, Lohse, & Usadel, 2014) to clip adapters, fastp-0.19.7 (Chen, Zhou, Chen, & Gu, 2018) to perform poly-G trimming, bowtie2-2.3.5.1 (Langmead & Salzberg, 2012) to align reads to the gadMor3 reference

154 genome (NCBI accession ID: GCF_902167405.1, Wellcome Sanger Institute, 2019),
155 samtools-1.11 (Li et al., 2009) to sort the resulting bam files, the MarkDuplicates module in
156 Picard Tools-2.9.0 (http://broadinstitute.github.io/picard/) to remove duplicated reads, the
157 clipOverlap module in BamUtil-1.0.14 (Jun, Wing, Abecasis, & Kang, 2015) to clip
158 overlapping read pairs, and GATK-3.7 (McKenna et al., 2010) to realign reads around indels.
159 We then used ANGSD-0.931 (Korneliussen, Albrechtsen, & Nielsen, 2014) to compute
160 genotype likelihoods and estimate individual heterozygosity across the entire genome in all
161 samples, taking both variable and invariable sites into account.
162 After processing all samples with this "batch-effect-naive" bioinformatic pipeline, we
163 noticed systematic differences in our estimates of individual heterozygosity between the two
164 batches of data (Figure 1A "before"). To identify what aspects of the data were driving these
165 differences and assess whether there were ways to mitigate them, we separately examined the
166 impact of each of the following potential sources of technical artefacts: poly-G tails, base
167 quality score miscalibration, reference bias in read alignment, DNA degradation level, and
168 sequencing depth. In the following sections, we describe in turn our approach to assessing
169 and mitigating the effects of each of these sources. Many of these efforts are simply based on
170 modifying part of our "batch-effect-naive" pipeline outlined above (e.g. using the sliding-
171 window trimming functionality in fastp to more effectively remove poly-G tails, or applying
172 more stringent filtering in ANGSD when estimating heterozygosity to alleviate the impact of
173 base quality score miscalibration). In other cases, we also used ANGSD for SNP calling,
174 principal component analysis (PCA), and F_{ST} estimation, and ngsLD-1.1.0 (Fox, Wright,
175 Fumagalli, & Vieira, 2019) for LD estimation and removal of strongly linked SNPs (i.e. LD
176 pruning). Finally, to examine the effect of varying depth of sequencing coverage, we used
177 simulated data in addition to the empirical cod datasets. We used SLiM-3.3 (Haller &
178 Messer, 2019) to simulate populations distributed in a two-dimensional space, and ART-

MountRainier (Huang, Li, Myers, & Marth, 2012) to simulate the lcWGS process to create comparable datasets with varying sequencing depth (more details in the supplementary material).

Results and Discussion

Across the two sequencing batches, we generated a total of 61.5 Gb raw sequencing data for the 163 samples. The systematic biases in population genomic inferences that we discovered pertain to estimates of genetic diversity, population structure, and selection scan. For example, the NextSeq-150PE samples consistently have substantially higher estimates of heterozygosity than HiSeq-125SE samples from the same population (Figure 1A "before"); samples from the same population but different batches cluster separately in a PCA (Figure 1B "before"); and when pooling all populations together, a large number of loci exhibit highly elevated levels of genetic differentiation (compared to the genome-wide mean) between the HiSeq-125SE and NextSeq-150PE batches, which is not expected because they are composed of samples from the same populations (Figure 1C "before").

Based on our bioinformatic analysis, we found that all the potential sources of technical artefact that we investigated (poly-G tails, base quality score miscalibration, reference bias, DNA degradation level, and varying sequencing depth) contributed to the batch effects observed in our data. We summarize these different causes of batch effects in Table 2, and in the following sections discuss each issue separately in detail.

Presence/absence of poly-G tails

A key factor that can cause batch effects when compiling data generated on different sequencing platforms is variation in their sequencing chemistry. Across Illumina platforms, an important change is the shift from a four-channel system (used e.g. on HiSeq instruments) where each DNA base is detected with a different fluorescent dye, to a two-channel chemistry, that uses the combinations of two different dyes. With the two-channel system (implemented on newer platforms like NextSeq and NovaSeq), G is called when there is little to no fluorescence signal. Accordingly, the absence of a signal can result from a true G base in the DNA template, but any low-intensity fluorescence signal (regardless of the true base) may also lead to a G call, which becomes problematic. Since the intensity of the fluorescence signal tends to decrease with sequencing cycles, false calls of G tend to be enriched at the end of reads, forming poly-G tails (De-Kayne et al. 2021). Although one might expect that reads with poly-G tails would fail to map to the reference genome and therefore would not cause problems downstream (especially with global alignment settings), we found that many of these reads can in fact map to the reference genome with high confidence (i.e. with mapping quality scores higher than 20, see Figure S2, also see Arora et al., 2019). Making the problem worse, these erroneous G calls are sometimes associated with high base quality scores (Figure 2A), so they can survive per-base quality trimming, and can also pass base quality filters in data analysis tools downstream. In our case, we found that poly-G tails were the main culprit behind the inflated heterozygosity estimates of the samples in the NextSeq-150PE batch (see comparison between Figure 1A "before" and Figure S3).

Poly-G trimming, as implemented in the program fastp (Chen et al., 2018), has been proposed as a possible solution to this problem (De-Kayne et al. 2021). However, we found that calls of other bases are often interspersed within poly-G tails, and fastp only allows a maximum of five non-G bases in a poly-G tail. As a result, longer poly-G tails cannot be completely removed by this functionality (Figure 2). In fact, although we included a poly-G

trimming step in our "batch-effect-naive" pipeline, the enrichment of G bases at the end of reads in our NextSeq-150PE data remained strong (Figure 2).

Instead, we found a sliding-window quality trimming approach more effective for removing poly-G tails. This is based on the observation that going from the start to the end of a read, there tends to be a region in which base quality starts to decrease significantly before a poly-G section appears (Figure 2A, https://sequencing.qcfail.com/articles/illumina-2-colour-chemistry-can-overcall-high-confidence-g-bases/). Therefore, we can move a sliding window from the start to the end of a read; once the average base quality drops below a threshold, we cut the window along with the remaining sequence after it. This approach is implemented in fastp as the `cut_right` option, and in Trimmomatic as the `SLIDINGWINDOW` option. Because a drop in base quality is often not immediately followed by a poly-G tail, sliding-window base quality trimming may result in greater data loss than necessary, but we found it to be much more effective at removing poly-G tails than targeted poly-G trimming with existing tools (Figure 2A). Indeed, after applying this method (with window size of 4 and average base quality threshold of 20), G bases are no longer enriched at the end of reads in our samples sequenced in the NextSeq-150PE batch (Figure 2B), and the initial disparity in heterozygosity estimates between the two batches is significantly reduced (Figure S3). We therefore use the sliding-window-trimmed NextSeq-150PE data in all subsequent analyses so that poly-G tails will not be a confounding factor.

Difference in levels of base quality score miscalibration

In an ideal scenario, a base quality score should accurately reflect the probability of the base call being correct. In practice, however, these scores are often incorrectly calibrated (Callahan et al., 2016; Ni & Stoneking, 2016), which can lead to batch effects if the levels of

such biases differ across sequencing runs. For example, overestimated base qualities in one batch of data may result in inflated estimates of genetic diversity because sequencing errors are more likely to be interpreted as true variants. Such inflated estimates can lead to erroneous conclusions about relative levels of diversity when compared to estimates generated from other sequencing batches with more accurate quality scores. Base quality score miscalibration can be particularly problematic for low-coverage data, because the estimated probability of a base call being correct is central to the underlying probabilistic analysis framework based on genotype likelihoods rather than called genotypes (Korneliussen et al., 2014; Lou et al., 2021; Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012).

A simple way to diagnose base quality score miscalibration is to compare diversity estimates (e.g., individual heterozygosity) obtained with a relaxed and a stringent base quality filter. If base quality scores are accurate, there should not be systematic differences between these estimates. However, if base quality scores are miscalibrated, sequencing errors cannot be accurately accounted for and can cause greater biases when they are more prevalent (i.e. when a relaxed filter is used). Therefore, if systematic differences are observed between diversity estimates obtained with different filters in one batch of data but not in others, base score quality miscalibration could be causing batch effects. In this case, using a more stringent base quality threshold for diversity estimates in all batches can provide more comparable results. In our data, we found that heterozygosity estimates consistently decreased in NextSeq-150PE samples after a more stringent base quality filter of 33 is applied (as opposed to 20), suggesting that the base quality scores in the NextSeq-150PE batch are overestimated (Figure 3). In contrast, heterozygosity estimates from HiSeq-125SE samples slightly increased after the filter, suggesting that their base quality scores are somewhat underestimated (Figure 3). As a result, within the same population, individuals in

278 the NextSeq-150PE batch tend to have higher heterozygosity estimates than their
279 counterparts in the HiSeq-125PE batch when a relaxed base quality filter is applied (Figure
280 S3), but this difference is greatly reduced with a more stringent base quality filter (Figure 1A
281 "after), suggesting that this bioinformatic filtering is an effective mitigation strategy.
282 Using more stringent base quality thresholds is a logistically simple approach, but it has
283 the downside of potentially wasting large amounts of data. In comparison, base quality score
284 recalibration is a more robust, yet computationally involved, method to counteract base
285 quality score miscalibration. However, some of the most widely used recalibration methods
286 (e.g. as implemented in GATK and ANGSD) require a database of known variable sites,
287 which is not readily available for most non-model species. Methods that do not rely on such
288 databases are also available (e.g. Chung & Chen, 2017; Kousathanas et al., 2017; Ni &
289 Stoneking, 2016; Orr, 2020; Zook, Samarov, McDaniel, Sen, & Salit, 2012), but their
290 effectiveness has not been extensively tested especially with low-coverage data.
291
292 Difference in levels of reference bias / alignment error
293
294 When the read type and/or read length differ between batches, batch effects can arise
295 from systematic differences in reference bias and alignment error. Specifically, compared to
296 longer paired-end reads, shorter single-end reads carrying bases that are different from the
297 reference are less likely to be aligned to the reference genome (either correctly or incorrectly)
298 with high confidence, and therefore tend to receive low mapping quality scores. Also, shorter
299 single-end reads are more prone to alignment errors caused by insertions and deletions
300 (indels), leading to erroneous identification of SNPs in genomic regions adjacent to indels
301 (Leigh et al., 2018).

We did not find indel-related alignment errors to be a cause of batch effects in our data, presumably because we had a species-specific genome and performed indel realignment. However, this issue has been discussed in detail in Leigh et al. (2018), so here we just summarize their recommendations in Table 1 and focus our analyses on reference bias. After removing poly-G tails from the NextSeq-150PE samples, we estimated F_{ST} between the two batches of data and found that although the background level of F_{ST} is very low, allele frequencies at a large number of SNPs are strongly differentiated between the two batches (Figure 1C "before"). This is not expected since they are composed of samples from the same populations. Therefore, we closely examined the read alignment at several of these outlier SNPs with the Integrative Genomics Viewer (Robinson et al., 2011). We show a typical example in Figure 4A, where the F_{ST} outlier SNP appears to have similar allele frequencies in the two batches before a mapping quality filter is imposed. With a minimum mapping quality filter of 20, the allele frequency at this SNP remains unchanged in the NextSeq-150PE batch. However, in the HiSeq-125SE batch, reads with the non-reference allele (A) are entirely removed by the mapping quality filter. As a result, the filtered data (erroneously) suggest a strong differentiation between the batches.

The strong reference bias in the HiSeq-125SE batch as exemplified in Figure 4A is not a singular case. We calculated the proportion of mapped reads surviving a mapping quality filter of 20 in the HiSeq-125SE batch at all SNPs, and found that this proportion is significantly lower in F_{ST} outlier SNPs (those with $F_{ST} > 0.3$ between the two batches, Figure 1C "before") compared to all other SNPs (t-test, p=2e-322, Figure 4B). In other words, F_{ST} outliers are enriched at sites that have large numbers of reads filtered out due to the mapping quality filter in the HiSeq-125SE batch of data.

Based on this pattern, a simple mitigation strategy is to locate the sites that have a high proportion of low-mapping-score reads mapping to them (e.g. >10%) in a batch of data with

single-end reads and/or shorter reads, and exclude them from further analyses. With this method, we were able to eliminate the majority of the most conspicuous F_{ST} outliers between the two batches (Figure 1C "after"). When different batches are composed of samples from the same populations (as in our case), another effective approach could be to remove the private alleles (those that are absent in one batch of data and are at intermediate frequency in the other batch) from certain analyses (e.g. genome-wide PCA, Figure 1B "after"). Similarly, calling SNPs with only one batch of data has been proposed as a potential strategy (De-Kayne et al., 2021), but in our data, this approach resulted in strong ascertainment bias (Figure S4).

Alternatively, Günther & Nettelblad, (2019) recommended a second round of read alignment with a modified reference genome, where a randomly chosen third base replaces the original base at each variable site identified in the first round of alignment. As suggested by Kofler et al. (2016), using different alignment tools and intersecting their results may be yet another promising mitigation method, since every tool has its own unique biases, which can be minimized by considering results from another tool. However, both of these approaches are computationally intensive and are not tested in this study.

Difference in levels of DNA degradation

Elevated levels of DNA degradation in one batch of data can also contribute substantially to batch effects. This is particularly relevant for temporal studies as older samples are likely to be more degraded, although other factors such as DNA preservation methods can also introduce variation in DNA degradation levels between batches of samples from the same time point (which is the case in our datasets where 34 samples from 3 populations were poorly-preserved and were sequenced in the HiSeq-125PE batch).

A major consequence of DNA degradation is deamination of cytosines (i.e., transition of C bases into U bases), causing enrichment of C-to-T and G-to-A substitutions in more degraded batches of data. Similar to base quality score miscalibration, these errors will also inflate diversity estimates, as degradation patterns will be regarded as true variants. Indeed, in our data, the degraded samples tend to have higher heterozygosity estimates than well-preserved samples from the same population, after batch effects caused by poly-G tails and base quality miscalibration are accounted for (Figure 5A "before"). In addition, we found that samples with different degradation levels also cluster separately on a PCA (Figure 5B "before"), although in this case, the effect of DNA degradation is potentially confounded with that of reference bias, and both are likely to play a role.

DNA degradation levels can often be assessed by visualizing the fragment length distribution of the extracted DNA on an agarose gel, but a simple bioinformatic method to detect degradation directly from sequencing data is to calculate the frequencies of different base substitutions among the private alleles in each batch of data. Degraded samples should show enrichment of C-to-T and G-to-A substitutions among its private alleles, which is indeed the case for our HiSeq-125SE batch (which has 34 degraded samples) (Figure 5C). An alternative method is to compare the change in diversity estimates after excluding all C-to-T and G-to-A transitions (e.g., the `-noTrans 1` option in ANGSD). Ignoring a subset of variant types certainly results in decreases in diversity indices in all samples, but if some samples are more strongly impacted, it means that DNA degradation levels are uneven among samples (Figure 5D). In this case, the diversity estimates excluding transitions will be more comparable between batches and less biased in a relative sense (Figure 5A "after"). Similar to the case of reference bias, when different batches are comprised of individuals from the same populations, it could also be effective to exclude all private alleles in both batches of data from certain analyses (e.g., genome-wide PCA, Figure 5B "after").

More robust, yet more computationally involved, methods to correct for batch effects caused by DNA degradation include base quality score recalibration for degraded DNA (e.g., mapDamage) (Jónsson, Ginolhac, Schubert, Johnson, & Orlando, 2013), or using genotype likelihood models that explicitly incorporate DNA damage (e.g., ATLAS) (Link et al., 2017).

Difference in sequencing depth

When datasets with different levels of sequencing depth are combined, the dataset with lower depth is likely to generate less accurate population genetic parameter estimates (Lou et al., 2021). For certain types of analysis, difference in sequencing depth between batches may also lead to systematic biases. For example, some PCA methods are unsuitable with extremely low-coverage data (Lou et al., 2021), and when extremely low-coverage and higher-coverage data are combined, clustering patterns can become driven by read depth. Here, to better illustrate the effect of sequencing depth without other factors interfering, we first used simulated data instead of our empirical data. In Figure 6, we simulated nine populations on a three-by-three grid, each connected to its neighbors by gene flow (this is the same model used in Section 4.2 in Lou et al., 2021). We then simulated two batches of sequencing data generation from variable numbers of samples in each population. The only difference between the two batches of simulated data is their sequencing depth (either 0.125 or 4x, see supplementary material for details about the simulations). At low sample size (5 or 10 per population), PCAs generated from PCAngsd-0.98 (Meisner & Albrechtsen, 2018) and the `-doCov 1` option in ANGSD tend to group samples with the same read depth together along one of the top PC axes, creating false patterns of clustering (Figure 6). In comparison, the PCoA generated from the `-doIBS 2` option in ANGSD is less prone to such biases (Figure 6). We observed a similar pattern in our empirical data, where the PCA generated

from the `-doCov 1` option in ANGSD does not show obvious signs of batch effects when other causes of batch effects are controlled despite the difference in sequencing depth between the two batches (Figure 1B "after", 4B "after"). In contrast, PCA generated from PCAngsd still has individuals from different batches clustering separately (Figure S5).

Therefore, when dealing with batches of data with different sequencing depths, we recommend using methods that are known to be less sensitive to read depth when possible. Downsampling the batch of data with higher coverage and comparing the results generated from before and after downsampling is another effective strategy to detect and mitigate such batch effects.

Practical Considerations

In this paper, we provide an example of how batch effects can manifest when sequencing datasets generated on different platforms are combined, and we showcase several simple bioinformatic approaches to identify the potential causes of batch effects and mitigate their impact. Researchers may wonder whether these mitigation measures should always be implemented when different datasets are combined. We argue that this will depend on the experimental design of each project. Specifically, if samples (or a subset of samples) are randomly assigned to batches as in the case of our project, researchers can follow their standard pipeline, but check for evidence of batch effects on all their results. For example, they could color PCA plots by batches to examine if individuals from the different batches tend to cluster separately, and they could verify whether heterozygosity estimated from one batch of data is consistently higher/lower than other batches when biological factors are controlled for (e.g. Figure 1A "before", 1B "before"). If batch effects are observed in such results, they can go through our list of potential causes and use the relevant filters to evaluate

and mitigate the impact (Table 1). We also emphasize that a complete randomization is not always necessary. Particularly, when new data is generated to supplement existing datasets, it would be very helpful to sequence just a few individuals that are comparable to individuals included in the existing datasets (e.g. these can be exactly the same individuals, or individuals from the same populations at the same time points). In downstream analyses, comparisons of these individuals among different batches could be used to detect potential artefacts. This also highlights the importance of tissue banks in ensuring reusability of sequencing data (De-Kayne et al., 2021).

However, if samples are not randomly assigned and if true biological signals may be confounded with batch effects, it may no longer be possible to determine the presence / absence of batch effects using standard analyses such as PCA or heterozygosity estimation. In such cases, we would recommend researchers to take a subset of data from each batch, and perform some of the tests that we have mentioned in this paper (e.g., comparing heterozygosity estimates before and after applying a stringent base quality filter, calculating the frequencies of different base substitutions in private alleles in each batch of data, etc.) as a means to determine the presence/absence of batch effects.

We focused our investigation on lcWGS data in this paper. Compared with other sequencing strategies, lcWGS has its unique challenges due to low data redundancy, reliance on accurate base quality scores, and the difficulty in dealing with low-frequency SNPs (Lou et al., 2021). Therefore, batch effects caused by poly-G tails, base quality score miscalibration, and DNA degradation are likely to be more problematic for low-coverage data. However, it is not difficult to imagine that all these issues can sometimes affect high-coverage data as well, especially when the analysis in question depends on accurate genotype calling at low-frequency SNPs (e.g., estimations of individual heterozygosity, site frequency spectrum, Watterson's theta, etc.). The reference bias / alignment error issue can be just as

problematic for high-coverage data as it is for low-coverage data or pooled data (Kofler et al., 2016; Leigh et al., 2018). Disparities in sequencing depth is unlikely to become an issue if depth is higher than 20x in all batches. Otherwise, genotype calling in the batch with lower coverage (even at medium coverage, e.g., 5x-20x) is likely to be more inaccurate, and may therefore cause batch effects (Warmuth & Ellegren, 2019). In these cases, genotype-likelihood-based inference may be preferable to genotype calling.

Conclusion

As we have illustrated in this paper, batch effects can be a pervasive source of bias in various types of population genomic inference from combined lcWGS datasets. This is further complicated by the fact that multiple factors can introduce batch effects and their signals can be confounded. Accordingly, when possible (e.g., if new datasets are generated), we should try to limit the extent of batch effects through experimental design. However, we have also shown that, when treated meticulously, different causes of batch effects can be disentangled, and their impact can be mitigated with simple bioinformatic filtering. Therefore, we conclude that combining datasets remains to be a promising approach, as long as batch effects are explicitly accounted for.

Data Availability Statement

Sequencing data that support the findings of this study will be openly available in Dryad at [URL], reference number [reference number]. The entire bioinformatic pipeline will be available in a GitHub repository release deposited in Zenodo (DOI: XXX).

Acknowledgements

We would like to thank Harmony Borchardt-Wier for assistance in the laboratory, Einar Eg Nielsen, Anja Retzel, and Rasmus Hedeholm for supplying the samples used for this work, and the Therkildsen Lab at Cornell University for valuable comments on earlier versions on this manuscript. This study was funded through a National Science Foundation grant to NOT (OCE-1756316).

References

Arora, K., Shah, M., Johnson, M., Sanghvi, R., Shelton, J., Nagulapalli, K., … Robine, N. (2019). Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. *Scientific Reports*, *9*(1), 19123. doi: 10.1038/s41598-019-55636-3

Bálint, M., Márton, O., Schatz, M., Düring, R.-A., & Grossart, H.-P. (2018). Proper experimental design requires randomization/balancing of molecular ecology experiments. *Ecology and Evolution*, *8*(3), 1786–1793. doi: https://doi.org/10.1002/ece3.3687

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42. doi: 10.1093/nar/gks1195

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. doi: 10.1038/nmeth.3869

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. doi: 10.1093/bioinformatics/bty560

Chung, J. C. S., & Chen, S. L. (2017). Lacer: Accurate base quality score recalibration for improving variant calling from next-generation sequencing data in any organism. *BioRxiv*, 130732. doi: 10.1101/130732

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., … Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, *19*(1), 332. doi: 10.1186/s12864-018-4703-0

De-Kayne, R., Frei, D., Greenway, R., Mendes, S. L., Retel, C., & Feulner, P. G. D. (2021). Sequencing platform shifts provide opportunities but pose challenges for combining genomic data sets. *Molecular Ecology Resources*, *21*(3). doi: https://doi.org/10.1111/1755-0998.13309

Elango, R., Banaganapalli, B., & Shaik, N. A. (2019). Driving Forces of Bioinformatics. In N. A. Shaik, K. R. Hakeem, B. Banaganapalli, & R. Elango (Eds.), *Essentials of Bioinformatics, Volume II: In Silico Life Sciences: Medicine* (pp. 1–8). Cham: Springer International Publishing. doi: 10.1007/978-3-030-18375-2_1

Field, D., Sansone, S.-A., Collis, A., Booth, T., Dukes, P., Gregurick, S. K., … Wilbanks, J. (2009). 'Omics Data Sharing. *Science*, *326*(5950), 234–236. doi: 10.1126/science.1180598

Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: Evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, *35*(19), 3855–3856. doi: 10.1093/bioinformatics/btz200

Gewin, V. (2016). Data sharing: An open mind on open data. *Nature*, *529*(7584), 117–119. doi: 10.1038/nj7584-117a

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, *15*(7), e1008302. doi: 10.1371/journal.pgen.1008302

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. doi: 10.1093/molbev/msy228

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594. doi: 10.1093/bioinformatics/btr708

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, *29*(13), 1682–1684. doi: 10.1093/bioinformatics/btt193

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. *Genome Research*, gr.176552.114. doi: 10.1101/gr.176552.114

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics—Re-shaping scientific practice. *Nature Reviews Genetics*, *10*(5), 331–335. doi: 10.1038/nrg2573

Kofler, R., Langmüller, A. M., Nouhaud, P., Otte, K. A., & Schlötterer, C. (2016). Suitability of Different Mapping Algorithms for Genome-Wide Polymorphism Scans with Pool-Seq Data. *G3: Genes|Genomes|Genetics*, *6*(11), 3507–3515. doi: 10.1534/g3.116.034488

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, *15*, 356. doi: 10.1186/s12859-014-0356-4

Kousathanas, A., Leuenberger, C., Link, V., Sell, C., Burger, J., & Wegmann, D. (2017). Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics*, *205*(1), 317–332. doi: 10.1534/genetics.116.189985

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. doi: 10.1038/nmeth.1923

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., … Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews. Genetics*, *11*(10), 10.1038/nrg2825. doi: 10.1038/nrg2825

Leigh, D. M., Lischer, H. E. L., Grossen, C., & Keller, L. F. (2018). Batch effects in a multiyear sequencing study: False biological trends due to changes in read lengths. *Molecular Ecology Resources*, *18*(4), 778–788. doi: https://doi.org/10.1111/1755-0998.12779

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., & Wegmann, D. (2017). ATLAS: Analysis Tools for Low-depth and Ancient Samples. *BioRxiv*, 105346. doi: 10.1101/105346

Lou, R. N., Jacobs, A., Wilder, A., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *n/a*(n/a). doi: 10.1111/mec.16077

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., … Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, *1*(6), 0160. doi: 10.1038/s41559-017-0160

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. doi: 10.1101/gr.107524.110

Meisner, J., & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics*, *210*(2), 719–731. doi: 10.1534/genetics.118.301336

Ni, S., & Stoneking, M. (2016). Improvement in detection of minor alleles in next generation sequencing by base quality recalibration. *BMC Genomics*, *17*(1), 139. doi: 10.1186/s12864-016-2463-2

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLOS ONE*, *7*(7), e37558. doi: 10.1371/journal.pone.0037558

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you'e looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, *27*(16), 3193–3206. doi: https://doi.org/10.1111/mec.14792

Orr, A. J. (2020). *Methods for Detecting Mutations in Non-Model Organisms* (Ph.D., Arizona State University). Arizona State University, United States -- Arizona. Retrieved from https://www.proquest.com/docview/2476130546/abstract/12747DA614FF4224PQ/1

Perez-Riverol, Y., Zorin, A., Dass, G., Vu, M.-T., Xu, P., Glont, M., … Hermjakob, H. (2019). Quantifying the impact of public omics data. *Nature Communications*, *10*(1), 3512. doi: 10.1038/s41467-019-11461-w

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. doi: 10.1038/nbt.1754

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. doi: https://doi.org/10.1002/cpmb.59

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194–208. doi: 10.1111/1755-0998.12593

Warmuth, V. M., & Ellegren, H. (2019). Genotype-free estimation of allele frequencies reduces bias and improves demographic inference from RADSeq data. *Molecular Ecology Resources*, *19*(3), 586–596. doi: https://doi.org/10.1111/1755-0998.12990

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. doi: 10.1038/sdata.2016.18

Zook, J. M., Samarov, D., McDaniel, J., Sen, S. K., & Salit, M. (2012). Synthetic Spike-in Standards Improve Run-Specific Systematic Error Analysis for DNA and RNA Sequencing. *PLOS ONE*, *7*(7), e41356. doi: 10.1371/journal.pone.0041356

Tables and Figures

Table 1. Summary of the key differences between our two sequencing batches.

Sequencing batch	Sequencing platform	Sequencing chemistry	Read type	Read length (in bp)	DNA degradation level	Average depth of coverage per sample*	Sample size
HiSeq-125SE	HiSeq 2500	four-color	single end	125	34 samples in 3 populations are degraded; all others are well-preserved	0.8x	88
NextSeq-150PE	NextSeq 500	two-color	paired end	150	well-preserved	0.3x	75

* After deduplication, overlap clipping, removal of reads with low mapping scores and poly-G trimming. See Figure S1 for sample size in each population and the depth of coverage in each sample.

Table 2. A summary of possible causes of batch effects in population genomic studies with low-coverage whole genome sequencing data, and methods to detect and mitigate their impact.

Cause	Identification	Mitigation
Presence/absence of poly-G tails	• Examine the base composition at each read position in raw fastq files (e.g., with FastQC) (Figure 2B)	• Trim off ends of reads with low base quality within sliding windows (e.g., the `cut_right` option in fastp, or the `SLIDINGWINDOW` option in Trimmomatic). (Figure 1A "after", 2)
Difference in levels of miscalibration in base quality scores	• Compare diversity estimates (e.g., individual heterozygosity) using a relaxed vs stringent base quality threshold within each batch (Figure 3)	• Use a more stringent base quality threshold in all batches (Figure 1A "after", 3) • Use base quality score recalibration (e.g., the SOAPsnp genotype likelihood model) (Korneliussen et al., 2014)
Difference in levels of reference bias / alignment error	• Spot check read alignments at outlier loci (Figure 4A) • Check for enrichment of outlier loci in genomic regions that have a high number of low-mapping-score reads mapped to them (Figure 4B) • Compare results using different alignment tools	• Perform indel realignment (e.g., the IndelRealigner tool in GATK3) (Leigh et al., 2018; McKenna et al., 2010) or use a haplotype-based variant discovery software (e.g. FreeBayes or HaplotypeCaller of GATK) • Use a species-specific reference genome (Leigh et al., 2018) • Trim all reads to the same length (Leigh et al., 2018) • Exclude genomic regions that have a high proportion of low-mapping-score reads mapped to them (Figure 1C "after") • Exclude private alleles of each batch from certain analyses (Figure 1B "after", 5B "after") • Change variable sites in the reference genome to a randomly chosen third base and redo read alignment (Günther & Nettelblad, 2019) • Use different alignment tools and intersect their results (Kofler et al., 2016)

Difference in levels of DNA degradation	• Examine the fragment size distribution in DNA extracts with gel electrophoresis • Compare the frequencies of different types of base substitutions among the private alleles in each batch (Figure 5C) • Compare the drop in diversity estimates (e.g., individual heterozygosity) after excluding all transitions between different batches of data (Figure 5D)	• Exclude transitions from certain analyses (Figure 5A "after") • Exclude private alleles of each batch from certain analyses (Figure 5B "after") • Recalibrate base quality scores for degradated DNA (e.g., mapDamage) (Jónsson et al., 2013) • Use genotype likelihood models that take post-mortem damage into account (e.g., ATLAS) (Link et al., 2017)
Difference in sequencing depths	• Color individual by batch or sequencing depth in a PCA to spot non-random clustering (Figure 6) • Examine whether down-sampling of high-coverage individuals systematically changes the results	• Use methods that are known to be less sensitive to differences in sequencing depth (Figure 6) • Down-sample data to achieve similar coverage across all individuals

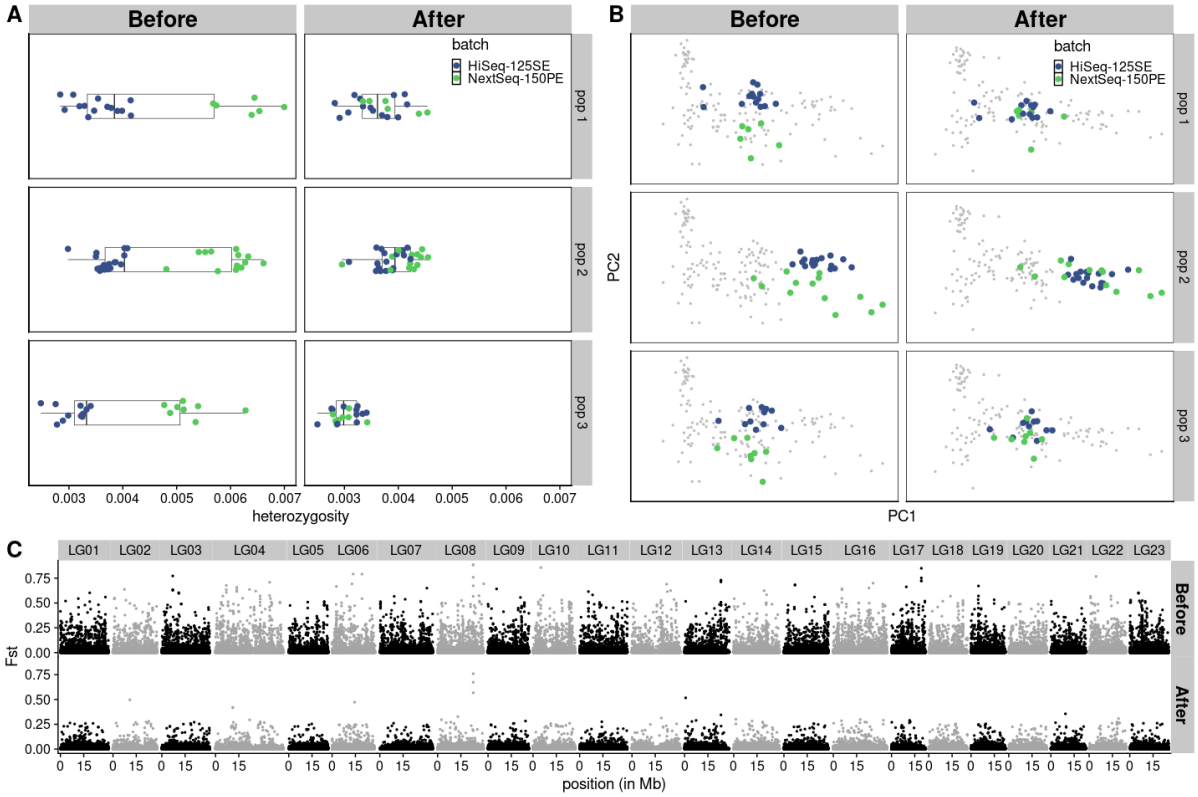

Figure 1. Examples of how batch effects are manifested in our data before correction, and how they were mitigated after our correction. **(A)** Individual heterozygosity estimated from each sample, grouped by populations on the y-axis and colored by batches, before and after batch effect correction (i.e., using sliding-window quality trimming in addition to poly-G trimming, and applying a more stringent base quality filter), in three representative populations where no samples suffered from DNA degradation. **(B)** Genome-wide PCA with all samples using an LD-pruned SNP list, colored by batches, before and after batch effect correction (i.e., excluding SNPs that are invariable in one batch of samples but are at intermediate frequencies in the other batch), in the same three populations as in **(A)**. Grey points represent samples from other populations. Two outlier points are removed from these plots to better illustrate the broader pattern in the data. Sliding-window quality trimming is performed in both "before" and "after". **(C)** F_{ST} between two batches of samples, before and after batch effect correction (i.e., excluding SNPs that have a high number of low-mapping-score reads mapped to them). Sliding window quality trimming is performed in both "before" and "after".

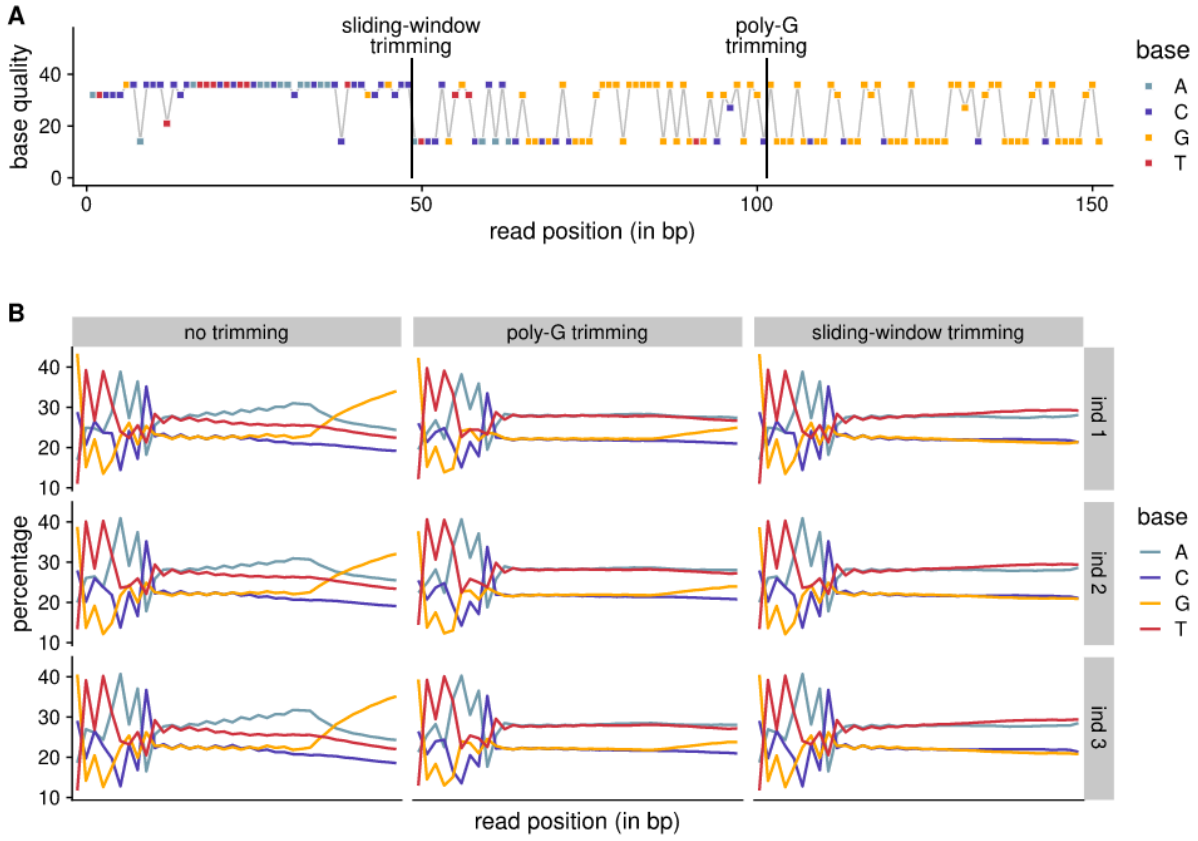

Figure 2. Sliding-window quality trimming (`cut_right` option in fastp) is more effective at removing poly-G tails in data generated by two-color-chemistry sequencing platforms than poly-G trimming (`trim_poly_g option` in fastp). **(A)** An example of how poly-G trimming and sliding-window trimming affect a typical read with a poly-G tail. Base quality score is shown on the y axis and fastp cut sites are indicated by vertical lines. **(B)** Base composition at each read position in three randomly chosen samples, before trimming, after poly-G trimming, and after sliding-window trimming. In both **(A)** and **(B)**, poly-G trimming is shown to remove part of the G enrichment towards the ends of reads, whereas sliding-window trimming removes the poly-G tails entirely.

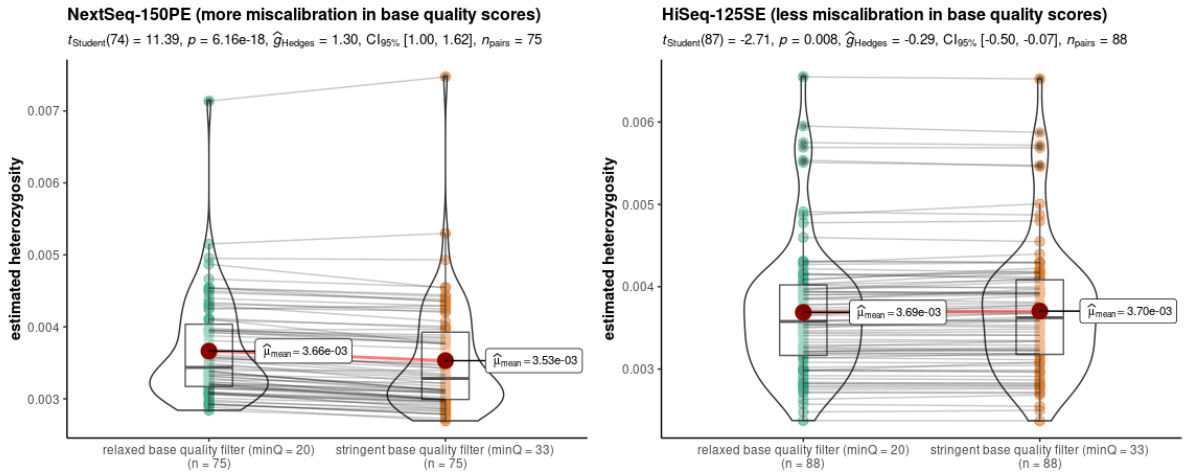

Figure 3. Comparing individual heterozygosity estimates obtained with relaxed vs. stringent base quality filter is a simple way to detect batch effects caused by base quality miscalibration. Individual heterozygosity estimates in two batches of data before and after applying a more stringent base quality filter (from 20 to 33) are shown on the y axis. Samples in the NextSeq-150PE batch (left) have significantly lower heterozygosity estimates after a more stringent filter is applied (paired samples t-test, p=6e-8) and therefore are likely to have overestimated base quality scores. In contrast, samples in the HiSeq-125SE batch tend to have slightly higher heterozygosity estimates after this filter is applied (paired samples t-test, p=0.008), suggesting that their base quality scores are somewhat underestimated.

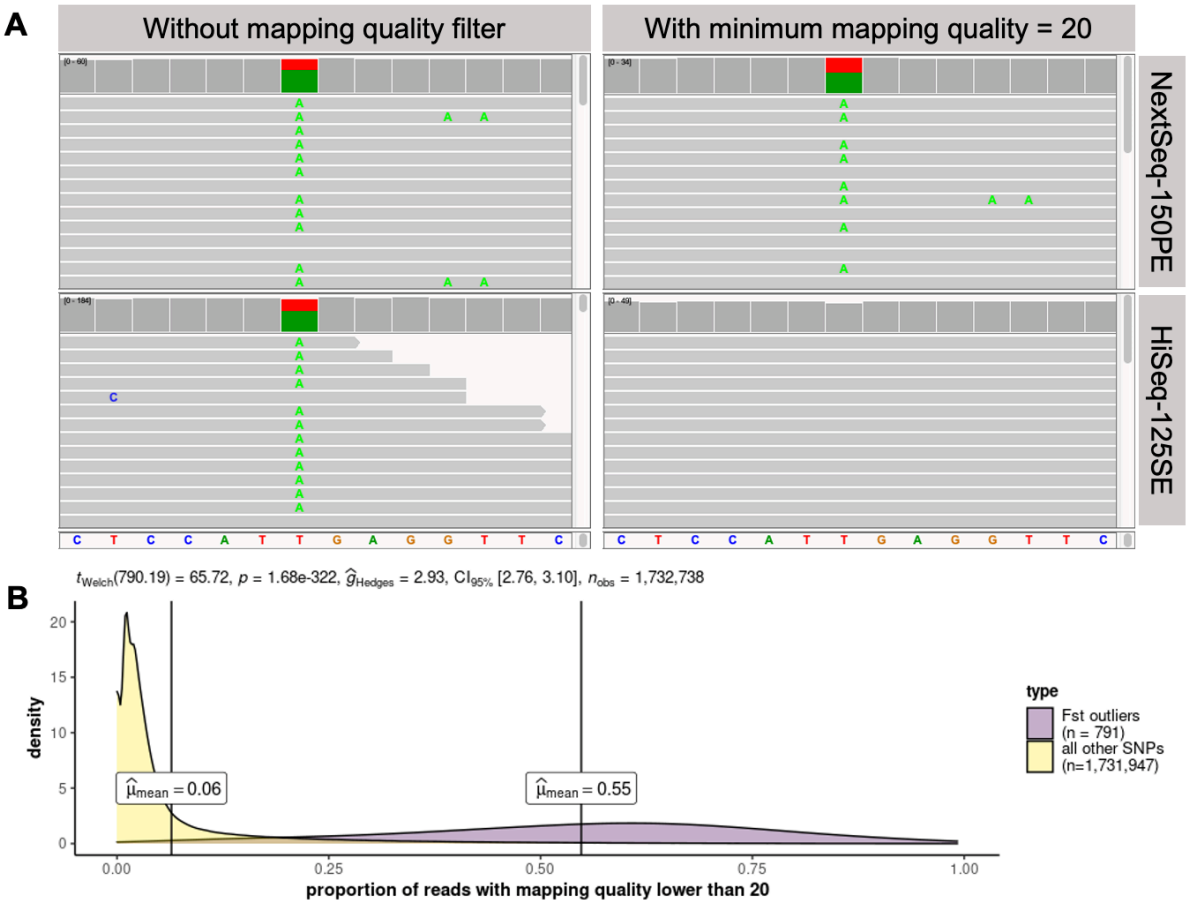

Figure 4. Batch effects caused by different levels of reference bias can be detected and mitigated by adjusting the mapping quality filter. **(A)** Screenshot from the Integrative Genomics Viewer showing read alignment from NextSeq-150PE batch (top) vs. HiSeq-125SE batch (bottom), and with (right) vs. without (left) a minimum mapping quality of 20. Reads with the non-reference allele are all removed after imposing the mapping quality filter in the HiSeq-125SE batch, leading to a false signal of allele frequency divergence between the batches. **(B)** Distribution of the proportion of HiSeq-125SE reads failing a minimum mapping quality filter of 20 in F_{ST} outliers and all other SNPs. F_{ST} outliers are enriched in genomic regions where higher proportions of reads are filtered out. This proportion can thus be used as a filter to remove the regions most affected by reference bias and mitigate the false signals of allele frequency divergence between batches.

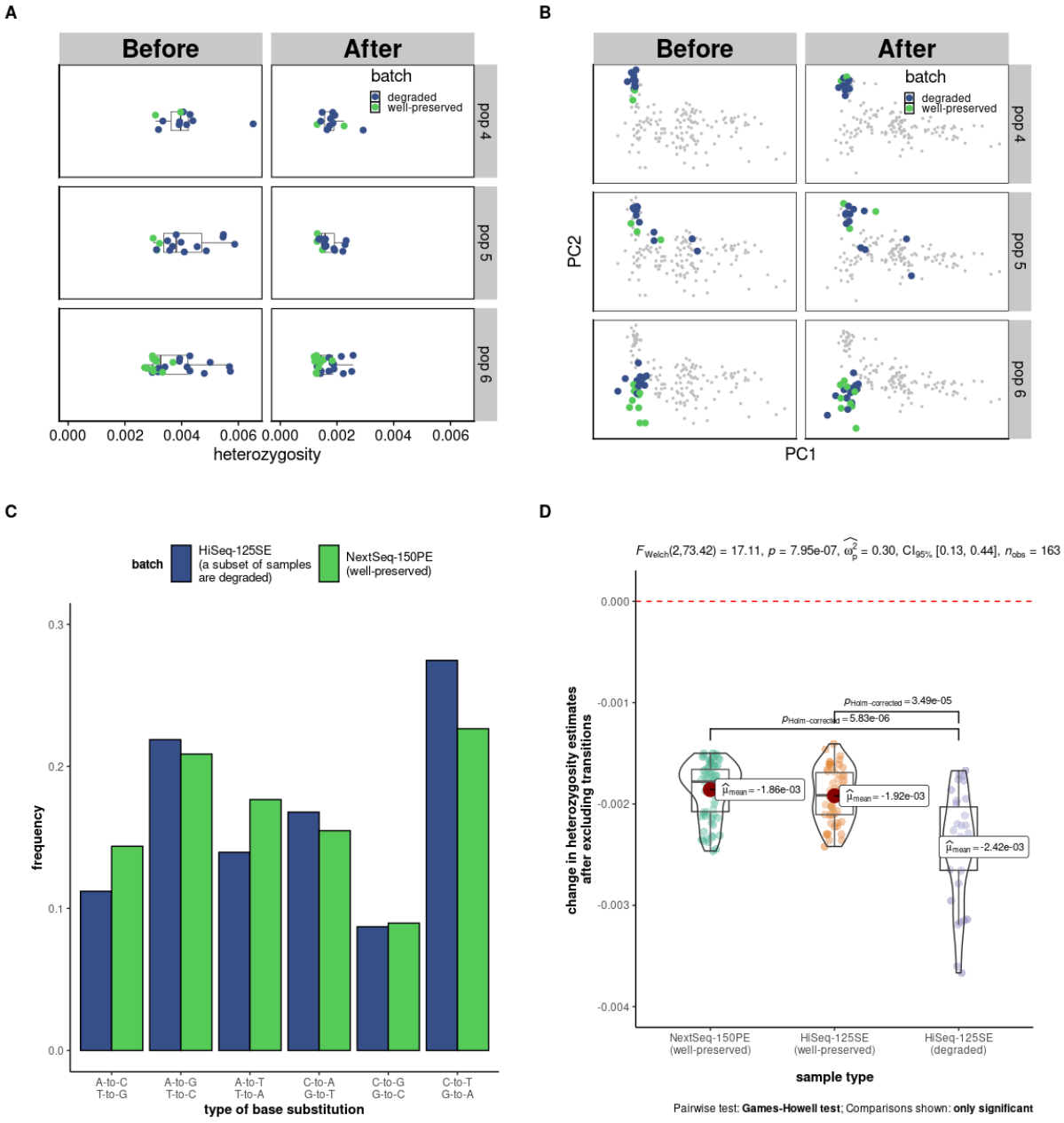

Figure 5. Batch effects caused by difference in DNA degradation level and strategies to detect and mitigate them. **(A)** Individual heterozygosity estimated from each sample, colored by batches, before and after batch effect correction (i.e., excluding all transitions), in three populations for which samples were split into batches based on their degradation level. Sliding-window quality trimming and a more stringent base quality filter are applied in both "before" and "after". **(B)** Genome-wide PCA with all samples using an LD-pruned SNP list, colored by batches, before and after batch effect correction (i.e., excluding SNPs that are invariable in one batch but are at intermediate frequencies in the other batch), in the same three populations as in **(A)**. Grey points represent the rest of samples. Two outlier points are removed from these plots to better illustrate the broader pattern in the data. Sliding-window quality trimming is performed in both "before" and "after". **(C)** Using the frequencies of different base substitutions in private alleles to detect DNA degradation. There is an

enrichment of C-to-T and G-to-A substitutions in the HiSeq-125SE batch, suggesting higher levels of DNA degradation in this batch. Reference alleles are assumed to be the wild-type alleles in this figure. **(D)** Using the change in individual heterozygosity estimates after filtering out transitions to detect DNA degradation. As expected, heterozygosity estimates in all samples are negatively affected, but the degraded samples in the HiSeq-125SE batch (right) are shown here to be more negatively affected than well-preserved samples in either batch, suggesting that their heterozygosity estimates are inflated when transitions are included (one-way ANOVA, p=8e-7).

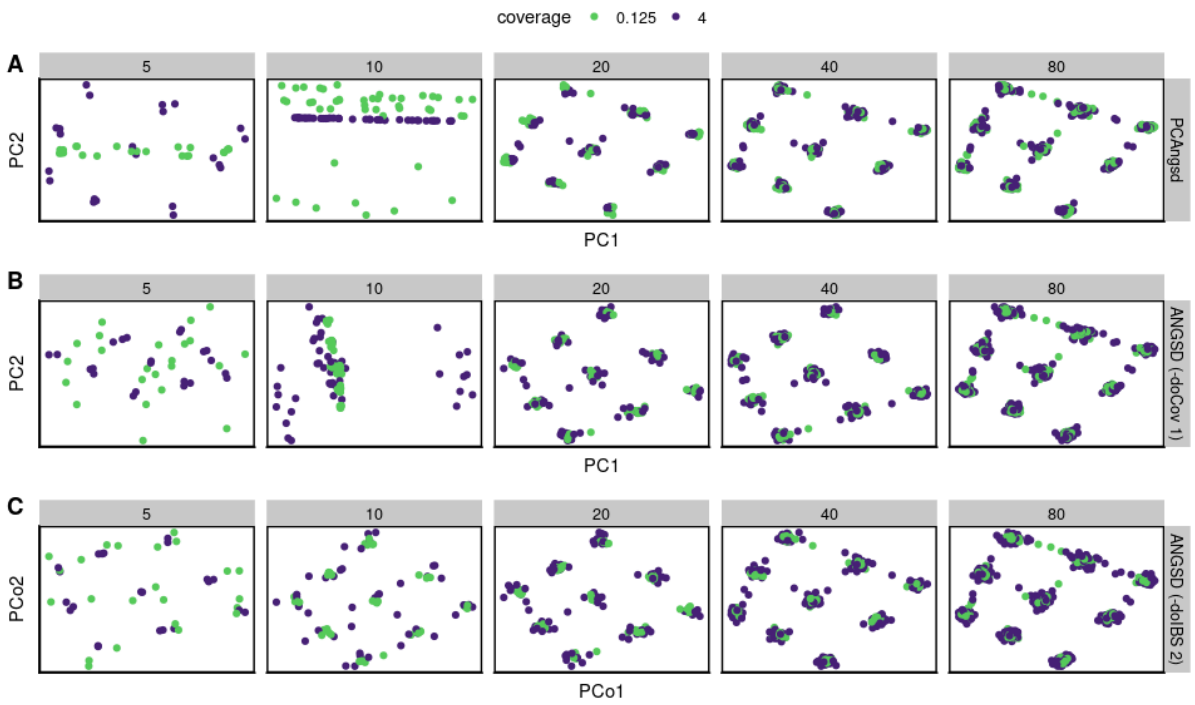

Figure 6. Some software programs are less sensitive to batch effects caused by different sequencing depths than others. **(A)** PCA generated from PCAngsd. **(B)** PCA generated from the `-doCov 1` option in ANGSD. **(C)** PCoA generated from the `-doIBD 2` option in ANGSD. A total of nine populations are simulated, and the sample size per population increases from left to right (as noted in panel headers). Green points mark individuals that are sequenced at 0.125x, whereas dark blue points mark individuals that are sequenced at 4x.