A setback into a success: What can batch effects tell us about best practices in genomics?

Xavier Dallaire, Claire Merot

MR. XAVIER DALLAIRE (Orcid ID : 0000-0003-2375-561X)

DR. CLAIRE MÉROT (Orcid ID : 0000-0003-2607-7818)

Article type : Perspective

A setback into a success: what can batch effects tell us about best practices in genomics?

Xavier Dallaire[1], Claire Mérot[1,2]*

[1] Institut de Biologie Intégrative des Systèmes, Département de Biologie, Université Laval, Québec, Canada

[2] UMR 6553 Ecobio, Université de Rennes, OSUR, CNRS, Rennes, France

* corresponding author: claire.merot@gmail.com

The increasing access to high-throughput sequencing is certainly one of the major changes that molecular ecology has gone through over the last decade. With the positive trend towards more open science, most sequencing datasets are now available on public databases, which holds amazing potential, but also risks of introducing batch effects in studies combining datasets. In this issue of Molecular Ecology Resources, Lou & Therkildsen (2022) offer a timely discussion on the matter by analyzing an imperfect low-coverage Whole Genome Sequencing dataset, in which they test the effects of differences in sequencing choices, DNA degradation, and read depth on routine population genomics analyses. Through a series of diagnostic tools, they uncover multiple factors producing technical artefacts that can bias estimates of genetic diversity, inference of population structure, and selection scans. For each confounding factor, they demonstrate the effectiveness of mitigation approaches and suggest other avenues to deal with the issue. In this perspective, we highlight

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.13615

considerations regarding (1) effects that arise from differences between batches of sequencing, (2) unavoidable heterogeneity within datasets, and (3) more general concerns around the use of next-generation sequencing in population genomics. Altogether, by exploring what may have appeared at first glimpse as a "failed" sequencing project, Lou & Therkildsen (2022) end up setting a standard of best practices to make the most of heterogeneous whole-genome sequences, opening a promising avenue towards efficient reuse of published datasets.

The reduced costs of next-generation sequencing have translated into a rapidly growing number of genomic studies, which are providing new insights into fundamental and applied questions in evolutionary biology and ecology. For comparative genomics and population genomics, increasing data availability brings the opportunity to increase sample size and to broaden the spatial or temporal scale of studies by reusing and merging datasets. However, uncorrected bias, missing data, filtering parameters, sample heterogeneity, and some analytical choices are known to generate artefactual signals (Goh et al., 2017; Tom et al., 2017). Moreover, the need to analyze large genomic datasets turned most of us into apprentices at bioinformatics. Along such a tortuous road, who hasn't wondered what the best practices are? How can we be sure that we are using the best tools or the best parameters to analyze our specific dataset? Is there uncorrected bias that may explain the observed patterns more than biological reasons? The matter can complicate further when reusing and/or combining previously generated datasets: how to deal with uncontrolled study design, variable qualities and coverage, differences in sequencing technologies, etc? The study by Lou and Therkildsen (2022) provides helpful answers to such questions as they explore the consequences of different technical artefacts that emerged in their dataset.

The first issue faced by Lou & Therkildsen (2022) is batch effect, which is a pattern of variation due to the merging of datasets produced separately that can confound or mask biological patterns. While usually mitigated by randomization or consistency in methods (e.g., library preparation and sequencing procedure), batch effect represents a major concern for studies carried over several years or across multiple labs, as well as for re-analysis of public datasets. In particular, with the rapid evolution of sequencing technology, whole-genome sequences may have been produced by different methods and may differ in read type and length. For instance, the switch from a 4-colour technology (Illumina HiSeq) to a 2-colour technology (NextSeq) as well as a different assessment of quality score between technologies led to spurious SNPs (single-nucleotide polymorphisms) and an overestimate of heterozygosity differences (Lou & Therkildsen, 2022). Moreover, the switch from single-end to paired-end, and from 125 to 150 base pairs, affected read mapping and lead to artifactual high differentiation between the two batches. Fortunately, while such artefacts are worrying, they can be detected by simple bioinformatic procedures and mitigated by filtering choices accounting for technology differences.

Besides batch effects, the paper also draws attention to usually overlooked biases that may arise from heterogeneity in DNA quality and sequencing depth, even within the best-controlled single batch design. For example, the authors observed higher heterozygosity in samples with degraded DNA, a pattern they link to the identification of false polymorphism through deamination of cytosines. This effect might be particularly strong in long-term studies if samples vary significantly in age and were preserved in various conditions. DNA quality, as well as fragment size of the libraries, imperfectly balanced pooling, and sequencing choices, may also lead to heterogeneity in sequencing depth between samples and between batches. Lou & Therkildsen explored this issue with simulated data and showed that the higher proportion of missing data in shallow-covered samples might dampen population structure signals, although this can be controlled with appropriate tools even for low (<4X) to very low coverage (<1X).

More generally, some emerging artefacts reported by Lou & Therkildsen (2022) are worth considering in most sequencing datasets. For instance, this study confirms the warning from De-Kayne et al. (2021) about the enrichment of guanine (G) at the end of NextSeq reads that might cause artifactual SNPs and an overestimation of genetic diversity, suggesting that the trimming procedure may need adjustment. Moreover, the current trend is reducing sequencing depth to allow the analysis of large numbers of biological samples in a cost-effective way (Lou et al., 2021). While this shift has merits based on the fact that many population genomics analyses produce more robust results with more samples at low coverage than with fewer samples at high coverage (Alex Buerkle & Gompert, 2013), Lou and Therkildsen (2021) remind us that low-coverage whole-genome sequencing is more sensitive to artefacts due to DNA degradation, depth heterogeneity or DNA quality. That being said, some artefacts such as reference bias and alignment errors are equally problematic with high-coverage data (Gage et al., 2019; Lloret-Villas et al., 2021), and more importantly, Lou & Therkildsen show that appropriate bioinformatic procedures are key to control and correct for the impact of multiple factors. Most of those mitigation methods include more stringent filtering that may reduce the fraction of genome actually analyzed or the number of polymorphic markers. This strategy remains nevertheless viable considering that whole genome sequencing produce amounts of data far superior to what is needed for many population genomics applications.

Beyond warning and practical recommendations to deal with heterogeneous sequencing datasets, Lou & Therkildsen (2022) also foster good practices for both designing and analyzing next-generation sequencing studies. The step-by-step approach leads to a collection of tests and tools that molecular ecologists can adapt to the peculiarities of their own datasets in a modular way. Moreover, even though the list of studied technical artefacts is inevitably incomplete, the approach undertaken by Lou & Therkildsen provides an inspiring example of how to explore and deal with future problems. For example, testing the effects of different levels of filtering on key

statistics and using simulations to validate some methods appears as a sensible framework applicable to other datasets. Altogether, this paves the way towards a more accurate use of different kind of genomic datasets and draws a promising picture of future research taking advantage of the incredible amount of sequence data already available. This will encourage ambitious research in molecular ecology based on larger datasets, more temporal or geographical replicates, and more non-model species.

Last but not least, it should be noted that Lou & Therkildsen (2022) clearly reported a problem in their data, describing not only what did work but also what didn't, and which steps of analyses and filtering succeeded or failed. In an academic world heavily focused on high-impact positive results, there is a tendency to shorten methods section in peer-reviewed papers. We believe that method-orientated work is equally important and highlights the relevance of methodological robustness in genomics.

References

Alex Buerkle, C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go?. *Molecular ecology*, *22*(11), 3028-3035.

De-Kayne, R., Frei, D., Greenway, R., Mendes, S. L., Retel, C., & Feulner, P. G. (2021). Sequencing platform shifts provide opportunities but pose challenges for combining genomic data sets. *Molecular Ecology*, *21*, 653-660.

Gage, J. L., Vaillancourt, B., Hamilton, J. P., Manrique-Carpintero, N. C., Gustafson, T. J., Barry, K., ... & de Leon, N. (2019). Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. The plant genome, 12(2).

Goh, W. W. B., Wang, W., & Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends in biotechnology*, *35*(6), 498-507.

Lloret-Villas, A., Bhati, M., Kadri, N. K., Fries, R., & Pausch, H. (2021). Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. BMC genomics, 22(1), 1-17.

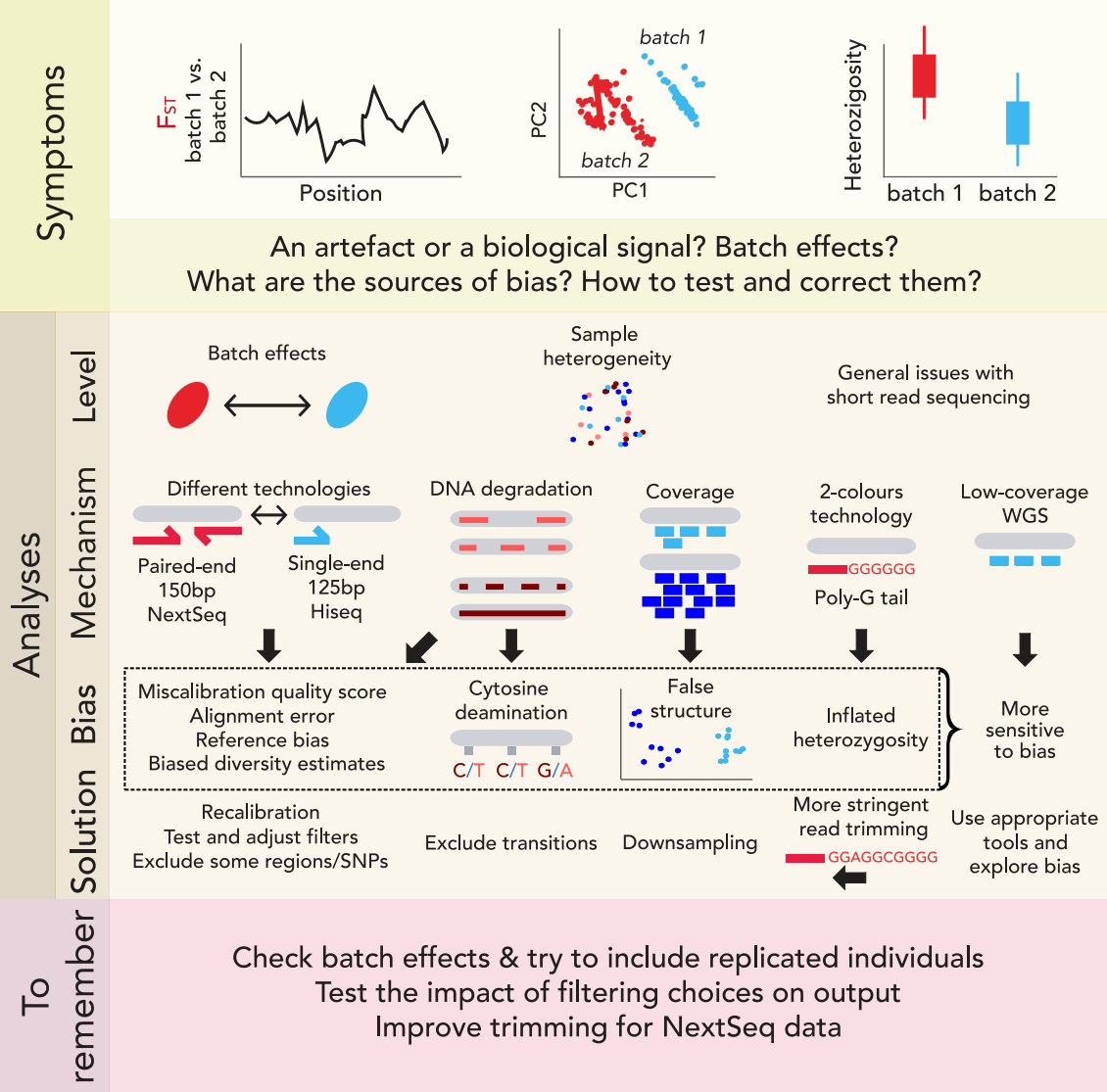
Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *30*(23), 5966-5993.

Lou, R. N., & Therkildsen, N. O. (2022). Batch effects in population genomic studies with low-coverage whole genome sequencing data : Causes, detection, and mitigation. *Molecular Ecology Ressources*,

Tom, J. A., Reeder, J., Forrest, W. F., Graham, R. R., Hunkapiller, J., Behrens, T. W., & Bhangale, T. R. (2017). Identifying and mitigating batch effects in whole genome sequencing data. *BMC bioinformatics*, *18*(1), 1-12.

Figure 1: Schematic overview of the technical artefacts explored and discussed in Lou & Therkildsen (2022).

The top section (*Symptoms*) displays some of the artefactual patterns observed by Lou & Therkildsen in the data that led to further exploration. The middle section (*Analyses*) splits the technical effects identified by Lou & Therkildsen into three levels that may re-appear in other research projects with different designs (Batch effects, which appear when combining two datasets; Sample heterogeneity, which may emerge even in a single dataset if samples are heterogeneous; General issues, which are of concern in any short-read sequencing project). Some of the mechanisms leading to technical bias are schematized and some solutions are proposed. It is worth noting that those factors are interrelated and interact with each other. For example, low-coverage WGS may exacerbate the bias due to other factors, and the 2-colours specificities is also a problem when combining batches sequenced with different technologies. The bottom section (*To remember*) summarizes the main takeaways.


Symptoms
F_{ST} batch 1 vs. batch 2
Position
batch 1
batch 2
PC2
PC1
Heterozigosity
batch 1
batch 2
An artefact or a biological signal? Batch effects?
What are the sources of bias? How to test and correct them?
Analyses
Level
Mechanism
Bias
Solution
Batch effects
Sample heterogeneity
General issues with short read sequencing
Different technologies
Paired-end 150bp NextSeq
Single-end 125bp Hiseq
DNA degradation
Coverage
2-colours technology
GGGGGG
Poly-G tail
Low-coverage WGS
Miscalibration quality score
Alignment error
Reference bias
Biased diversity estimates
Cytosine deamination
C/T C/T G/A
False structure
Inflated heterozygosity
More sensitive to bias
Recalibration
Test and adjust filters
Exclude some regions/SNPs
Exclude transitions
Downsampling
More stringent read trimming
GGAGGCGGGG
Use appropriate tools and explore bias
To remember
Check batch effects & try to include replicated individuals
Test the impact of filtering choices on output
Improve trimming for NextSeq data