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

Runyang Nicolas Lou[1] and Nina Overgaard Therkildsen[1]

[1]Cornell University

August 3, 2021

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.

Hosted file

`batch_effect_072221_no_field_code.pdf` available at https://authorea.com/users/380682/articles/532568-batch-effects-in-population-genomic-studies-with-low-coverage-whole-genome-sequencing-data-causes-detection-and-mitigation

Hosted file

`supplement_batch_effect_072221.pdf` available at https://authorea.com/users/380682/articles/532568-batch-effects-in-population-genomic-studies-with-low-coverage-whole-genome-sequencing-data-causes-detection-and-mitigation