Reliable NGS genotyping of MHC class I and II genes requires template-specific optimization of pipeline settings

Artemis Efstratiou[1], Arnaud Gaigher[1], Sven Künzel[2], Ana Teles[1], and Tobias Lenz[1]

[1]University of Hamburg Faculty of Mathematics Computer Science and Natural Sciences
[2]Max Planck Institute for Evolutionary Biology

June 5, 2023

Abstract

Using high-throughput sequencing for precise genotyping of multi-locus gene families, such as the Major Histocompatibility Complex (MHC), remains challenging, due to the complexity of the data and difficulties in distinguishing genuine from erroneous variants. Several dedicated genotyping pipelines for data from high-throughput sequencing, such as next-generation sequencing (NGS), have been developed to tackle the ensuing risk of artificially inflated diversity. Here, we thoroughly assess three such multi-locus genotyping pipelines for NGS data, using MHC class IIβ datasets of three-spined stickleback gDNA, cDNA, and "artificial" plasmid samples with known allelic diversity. We show that genotyping of gDNA and plasmid samples at optimal pipeline parameters was highly accurate and reproducible across methods. However, for cDNA data, the same configuration yielded decreased overall genotyping precision and consistency between pipelines. Further adjustments of key clustering parameters were required to account for higher error rates and larger variation in sequencing depth per allele, highlighting the importance of template-specific pipeline optimization for reliable genotyping of multi-locus gene families. Through accurate paired gDNA-cDNA genotyping and MHC-II haplotype inference, we show that MHC-II allele-specific expression levels correlate negatively with allele number across haplotypes. Lastly, sibship-assisted cDNA genotyping of MHC-I revealed novel variants and haplotype-based allelic segregation with a higher-than-previously-reported individual allelic diversity for MHC-I in sticklebacks. In conclusion, we here provide novel genotyping protocols for MHC-I and -II genes of the three-spined stickleback, but also evaluate the performance of popular NGS-genotyping pipelines and highlight the need for template-specific optimization for reliable multi-locus genotyping.

Hosted file

`Manuscript.docx` available at https://authorea.com/users/625506/articles/647480-reliable-ngs-genotyping-of-mhc-class-i-and-ii-genes-requires-template-specific-optimization-of-pipeline-settings