The future of next generation sequencing datasets: technological shifts provide opportunities but pose challenges for reproducibility and reusability

Rishi De-Kayne[1], David Frei[1], Ryan Greenway[1], Sofia Mendes[2], Cas Retel[1], and Philine Feulner[1]

[1]Eawag, Swiss Federal Institute of Aquatic Science and Technology
[2]Universidade de Lisboa

September 11, 2020

Abstract

Technological advances in DNA sequencing over the last decade now permit the production and curation of large genomic datasets in an increasing number of non-model species. Additionally, this new data provides the opportunity for combining datasets, resulting in larger studies with a broader taxonomic range. Whilst the benefits of new sequencing platforms are obvious, shifts in sequencing technology can also pose challenges for those wishing to combine new sequencing data with data sequenced on older platforms. Here, we outline the types of studies where the use of curated data might be beneficial, and highlight potential biases that might be introduced by combining data from different sequencing platforms. As an example of the challenges associated with combining data across sequencing platforms, we focus on the impact of the shift in Illumina's base calling technology from a four-channel to a two-channel system. We caution that when data is combined from these two systems, erroneous guanine base calls that result from the two-channel chemistry can make their way through a bioinformatic pipeline, eventually leading to inaccurate and potentially misleading conclusions. We also suggest solutions for dealing with such potential artifacts, which make samples sequenced on different sequencing platforms appear more differentiated from one another than they really are. Finally, we stress the importance of archiving tissue samples and the associated sequences for the continued reproducibility and reusability of sequencing data in the face of ever-changing sequencing platform technology.

Hosted file

`Main_Text_Final.pdf` available at https://authorea.com/users/358315/articles/480505-the-future-of-next-generation-sequencing-datasets-technological-shifts-provide-opportunities-but-pose-challenges-for-reproducibility-and-reusability