Read correction for non-uniform coverages

Camille Marchet[1], Yoann Dufresne[2], Antoine Limasset[1]

[1]CNRS, Université de Lille, CRIStAL UMR 9189, Lille, France
[2]Sequence Bioinformatics group and Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France

Next generation sequencing produces large volumes of short sequences with broad applications. The noise due to sequencing errors led to the development of several correction methods. The main correction paradigm expects a high (from 30-40X) uniform coverage to correctly infer a reference set of subsequences from the reads, that are used for correction. In practice, most accurate methods use k-mer spectrum [2, 4] techniques to obtain a set of reference k-mers. However, when correcting NGS datasets that present an uneven coverage, such as metagenomics, metatranscriptomics or RNA-seq data, this paradigm tends to mistake rare variants for errors [3]. It may therefore discard or alter them using highly covered sequences, which leads to an information loss and may introduce bias.

We present two new contributions in order to cope with this situation. First, we show that starting from non-uniform sequencing coverages, a De Bruijn graph can be cleaned from most errors while preserving biological variability. Second, we demonstrate that reads can be efficiently corrected via local alignment on the cleaned De Bruijn graph paths. We implemented the described method in a tool dubbed BCT and evaluated its results on RNA-seq and metagenomic data, reproducing data from the CAMI challenge [1]. We show that the graph cleaning strategy combined with the mapping strategy leads to save more rare k-mers, resulting in a more conservative correction than previous methods. BCT is also capable to better take advantage of the signal of high depth datasets. We suggest that BCT, being scalable to large metagenomic datasets as well as correcting shallow single cell RNA-seq data, can be a general corrector for non-uniform data. Finally, our results show that the correction step allows data reduction by removing a large quantity of k-mers from sequencing errors while keeping the majority of biological signal. This can be valuable for k-mer based indexing data structures, by reducing the number of objects to index.

Future applications include studying the impact of corrected short reads on hybrid correction of long reads (PacBio, Nanopore), for which little is know when it comes to non-genomic, non-uniform data.

Availability: BCT is open source and available at `github.com/Malfoy/BCT` under the Affero GPL License.

References

[1] Adrian Fritz, Peter Hofmann, Stephan Majda, Eik Dahms, Johannes Dröge, Jessika Fiedler, Till R Lesker, Peter Belmann, Matthew Z DeMaere, Aaron E Darling, et al. Camisim: Simulating metagenomes and microbial communities. *Microbiome*, 7(1):17, 2019.

[2] Yongchao Liu, Jan Schröder, and Bertil Schmidt. Musket: a multistage k-mer spectrum-based error corrector for illumina sequence data. *Bioinformatics*, 29(3):308–315, 2012.

[3] Li Song and Liliana Florea. Rcorrector: efficient and accurate error correction for illumina rna-seq reads. *GigaScience*, 4(1):48, 2015.

[4] Li Song, Liliana Florea, and Ben Langmead. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome biology*, 15(11):509, 2014.