SeqBIM

Abstract

UMI-Gen: a UMI-based read simulator for variant calling evaluation

Vincent Sater[1,2,3]*, Pierre-Julien Viailly[2,3], Thierry Lecroq[1], Philippe Ruminy[2,3], Élise Prieur-Gaston[1], Caroline Bérard[2,3] and Fabrice Jardin[2,3]

[1]*Normandie Univ, UNIROUEN, LITIS EA 4108, 76000 Rouen, France*

[2]*Centre Henri Becquerel, 76000 Rouen, France*

[2]*Normandie Univ, UNIROUEN, INSERM U1245, Team "Genomics and Biomarkers of Lymphoma and Solid Tumors", 76000 Rouen, France*

***Corresponding author**: vincent.sater@gmail.com

Abstract

With Next Generation Sequencing becoming more affordable every year, NGS technologies asserted themselves as the fastest and most reliable way to detect Single Nucleotide Variants (SNV) and Copy Number Variations (CNV) in cancer patients. These technologies can be used to sequence DNA at very high depths thus allowing to detect abnormalities in tumor cells with very low frequencies. Multiple variant callers are publicly available and are usually efficient at calling out variants. However, when frequencies begin to drop under 1%, the specificity of these tools suffers greatly as true variants at very low frequencies can be easily confused with sequencing or PCR artifacts. The recent use of Unique Molecular Identifiers (UMI) [1] in NGS experiments has offered a way to accurately separate true variants from artifacts. UMI-based variant callers are slowly replacing raw-read based variant callers as the standard method for an accurate detection of variants at very low frequencies. However, benchmarking done in the tools publication are usually realized on real biological data in which real variants are not known, making it difficult to assess their accuracy. We present UMI-Gen, a UMI-based read simulator for targeted sequencing paired-end data. UMI-Gen generates reference reads covering the targeted regions at a user customizable depth. After that, using a number of control files, it estimates the background error rate at each position and then modifies the generated reads to mimic real biological data. Finally, it will insert real variants in the reads from a list provided by the user.

References

[1] Y. Kukita, R. Matoba, J. Uchida, T. Hamakawa, Y. Doki, F. Imamura, K. Kato, High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients, DNA Res 22 (2015) 269–277. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4535617/. doi:10.1093/dnares/dsv010.