DOI 10.18699/SBB-2022-215

Seeking an optimal approach to variant calling in medical genetics

Barbitoff Y.A.[1, 2]*, Abasov R.[2], Predeus A.V.[2], Glotov A.S.[1]

[1] *Dpt. of Genomic Medicine, D.O. Ott Research Institute of Obstetrics, Gynaecology, and Reproductology, St. Petersburg, Russia*

[2] *Bioinformatics Institute, St. Petersburg, Russia*

* *barbitoff@bioinf.me*

Key words: exome sequencing, coverage bias, variant calling, benchmark

Motivation and Aim: Accurate variant detection in the coding regions of the human genome is a key requirement for molecular diagnostics of Mendelian disorders. Efficiency of variant discovery from next-generation sequencing (NGS) data depends on multiple factors, including reproducible coverage biases of NGS methods and the performance of read alignment and variant calling software.

Methods and Algorithms: In this work, we systematically evaluated the performance of 5 kits for whole-exome (WES) and whole-genome (WGS) sequencing, as well as 4 popular short read aligners and 9 novel and well-established variant calling and filtering methods. We used a large dataset of both "gold standard" (provided by the Genome In A Bottle (GIAB) consortium) and in-house WES and WGS samples.

Results: We showed that, contrary to a common view, most of the observed bias in modern WES stems from mappability limitations of short reads and exome probe design rather than sequence composition. When non-repetitive targeted exome regions are considered, WES performs similarly to WGS in terms of coverage and variant calling accuracy. Among software tools tested, Bowtie2 performed significantly worse than other aligners, suggesting it should not be used for medical variant calling. When other aligners were considered, the accuracy of variant discovery mostly depended on the variant caller and not the read aligner. Among the tested variant callers, DeepVariant consistently showed the best performance and the highest robustness. We have also compared the consistency of variant calls in GIAB and non-GIAB samples. With few important caveats, best-performing tools have shown little evidence of overfitting.

Conclusions: Our analysis showed that modern strategies of exome sequencing and NGS data analysis allow for high accuracy of genetic variant discovery within coding regions of the human genome. However, there is still a need for development of new library preparation and variant calling methods to enhance variant discovery in the repetitive regions of the coding genome.

Acknowledgements: This work was supported by he Ministry of Science and Higher Education of Russian Federation (project "Multicenter research bioresource collection – Human Reproductive Health" contract No. 075–15–2021-1058 from September 28, 2021). We thank the Systems Biology Fellowship for providing financial assistance to Y.A.B.