DOI 10.18699/SBB-2022-439

Swaveform: a genome-wide survey of structural variation profiles

Bezdvornykh I.*, Kanapin A., Samsonova A.

Institute for Translational Biomedicine, Saint Petersburg State University, St. Petersburg, Russia
** latur@me.com*

Key words: structural variation, waveforms, motif discovery, whole genome sequencing, signal processing, machine learning

Motivation and Aim: Genome structural variation (SV) is a potent source of genetic diversity and may have a profound effect upon human health [1, 2]. Typically, discovery algorithms rely on uniformity and evenness of sequencing coverage profile, as well as read depth information for accurate detection of structural variants [3]. However, as sequencing coverage signal is discontinuous, heterogeneous, and irregular, often even erratic existing SV detection tools still generate highly discordant results [4–6]. A comprehensive survey of signal coverage profiles in the vicinity of SV breakpoints may significantly improve robustness and reliability of SV discovery. Here, for the first time we present a detailed catalogue of various waveforms and patterns observed in the sequence coverage signal associated with different types of SVs, as well as a toolkit for coverage data management and analytics.
Methods and Algorithms: We used HGDP sequencing project data [7] which includes 929 whole-genome sequenced human samples with an average depth of coverage of about 30x. Structural variation data generated by the consortium contains annotated breakpoints for the following types of the events: CNV gain and loss, insertions, inversions, deletions and duplications [8]. We have developed a software suite to extract depth of coverage (DOC) profiles in a ± 256 bp vicinity of the breakpoints. Furthermore, to speed up coverage data processing and optimize storage we introduced a simple binary format (.bcov) for recording of coverage values. The extracted profiles and sequencing samples metadata are stored in a relational database (SQLite) to facilitate data search, retrieval and visualization.
A profound variability of waveforms associated with different classes of SVs has long impeded the reliability and reproducibility of the discovery algorithms. We, therefore, sought to identify repeated patterns found within DOC profiles (i.e., motifs) and characterize conserved structures in a signal. Coverage profiles attributed to specific classes of SVs were compressed, normalised and clustered using dynamic time warping distance. To reveal predominant waveforms and to discover the most significant motifs from the profiles, the signal was transformed into a SAX (Symbolic Aggregate approximation) representation [9]. The SAX-based motifs, stored in the database may have an important utility for development of novel improved approaches for breakpoint detection.
Results: In this study we present Swaveform – a recourse providing easy access to 15705161 DOC signal profiles extracted from 911 human sequencing samples generated by the HGDP. A portable database architecture and provided API facilitate easy and seamless on-premises deployment encompassing data processing routines on all levels i.e., from raw aligned data to visual representation of coverage profiles. We also propose

a new binary format to manage sequence coverage data. Finally, as motif discovery has been successfully applied throughout a large range of domains such as medicine, finance, robotics and DNA analyses we designed an algorithm for motif extraction from the coverage signal.

Conclusion: Robust and reproducible structural variation discovery still poses significant computational and algorithmic challenges and as of today we are nowhere near to resolving structural variation in personal genomes with accuracy required for translational research. The developed catalogue and accompanying toolkit form an indispensable resource that will facilitate development and honing of computational tools for discovery of specific genomic rearrangement events. It is expected that Swaveform will be of immense value to the machine learning and biomedical communities.

Acknowledgements: This work is supported by the Russian Scientific Foundation grant 20-14-00072.

References

1. Collins R.L. et al. A structural variation reference for medical and population genetics. *Nature*. 2020;581(7809):444-451. doi: 10.1038/s41586-020-2287-8.
2. Liu Z., Roberts R., Mercer T.R., Xu J., Sedlazeck F.J., Tong W. Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biol*. 2022;23(1):68. doi: 10.1186/s13059-022-02636-8.
3. Cameron D.L., Stefano L., Papenfuss A.T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun*. 2019;10(1):1-11. doi: 10.1038/s41467-019-11146-4.
4. Parikh H. et al. Svclassify: a method to establish benchmark structural variant calls. *BMC Genomics*. 2016;17(1):64. doi: 10.1186/s12864-016-2366-2.
5. Chaisson M.J.P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun*. 2019;10(1):1-16. doi: 10.1038/s41467-018-08148-z.
6. Chapman L.M. et al. A crowdsourced set of curated structural variants for the human genome. *PLoS Comput Biol.* 2020;16(6):e1007933. doi: 10.1371/journal.pcbi.1007933.
7. Bergström A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020;367(6484):eaay5012. doi: 10.1126/science.aay5012.
8. Almarri M.A. et al. Population structure, stratification, and introgression of human structural variation. *Cell*. 2020;182(1):189-199.e15. doi: 10.1016/j.cell.2020.05.024.
9. Malinowski S., Guyet T., Quiniou R., Tavenard R. Advances in Intelligent Data Analysis XII. *Lect Notes Comput Sci.* 2013;8207:273-284. doi: 10.1007/978-3-642-41398-824.