Structural Variants Identification From Long And Short-Read Sequencing Technologies for daughters from families of the 1000 Genomes Project

Laboratory of Functional and Structural Genomics

CeNT

Sachin Gadakh[1], Mateusz Chiliński[1,2], Karolina Jodkowska[1], Jan Gawor[3] and Dariusz Plewczynski*[1,2]

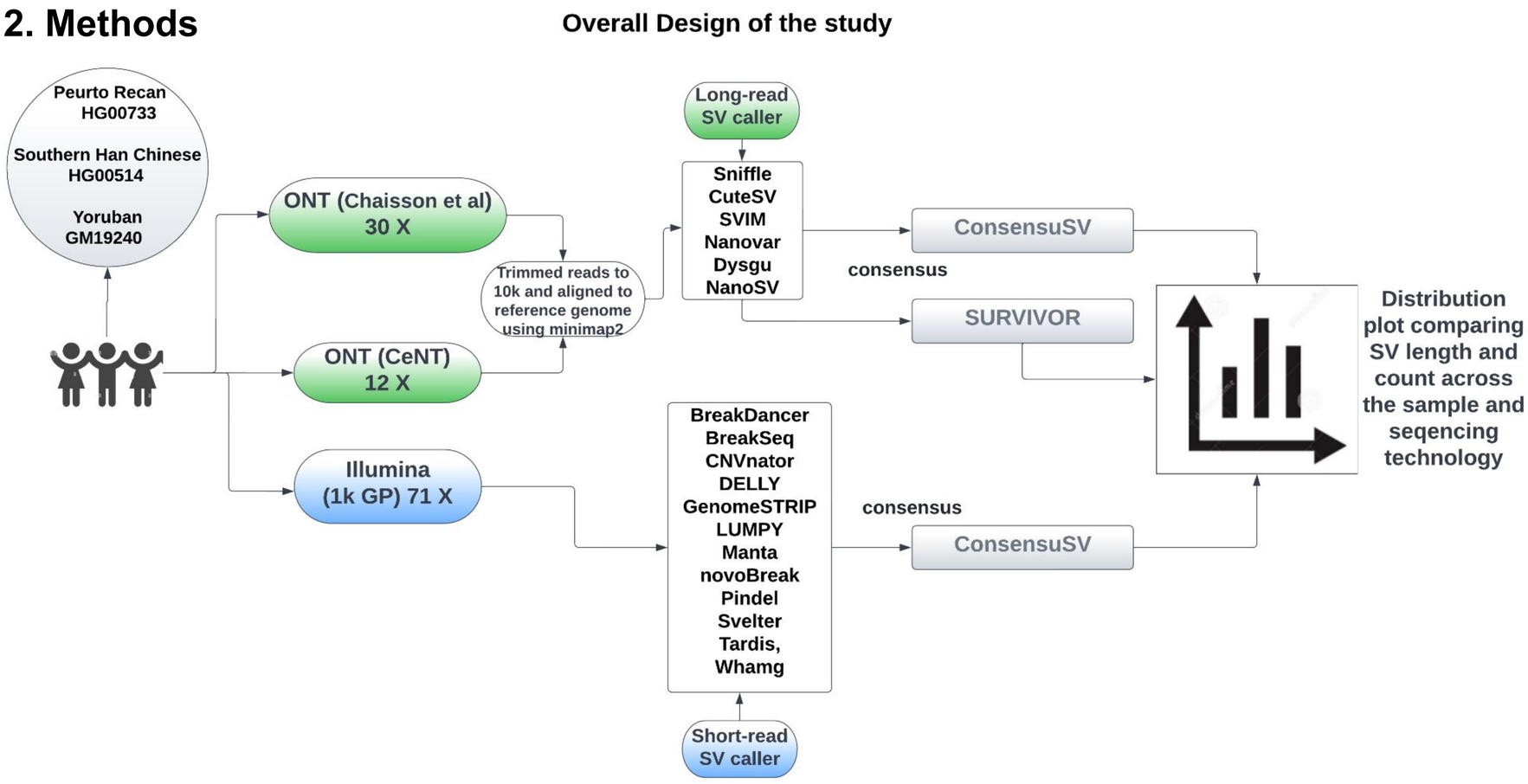
[1]Centre of New Technologies, University of Warsaw, Warsaw, Poland
[2]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
[3]Institute of Biochemistry and Biophysics, Warsaw, Poland
*Corresponding author d.plewczynski@cent.uw.edu.pl

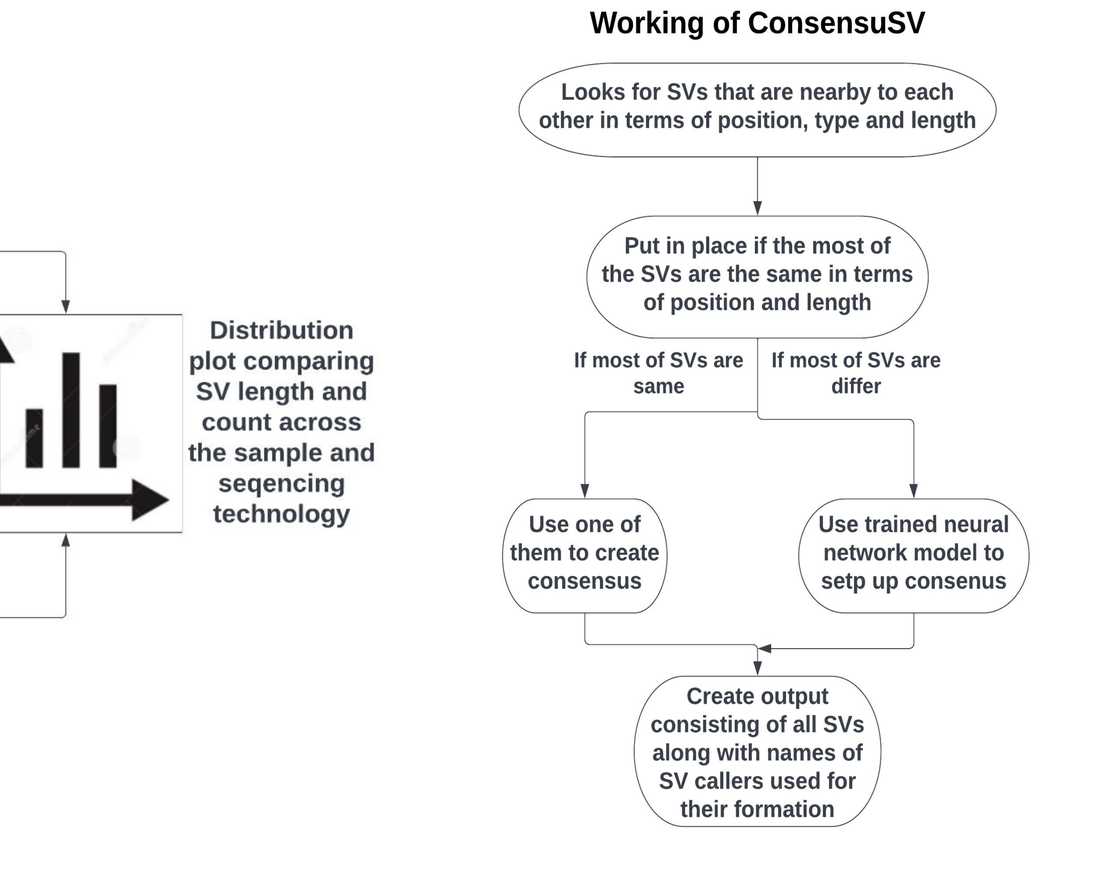
WYDZIAŁ BIOLOGII · UNIWERSYTET WARSZAWSKI ·

1. Introduction

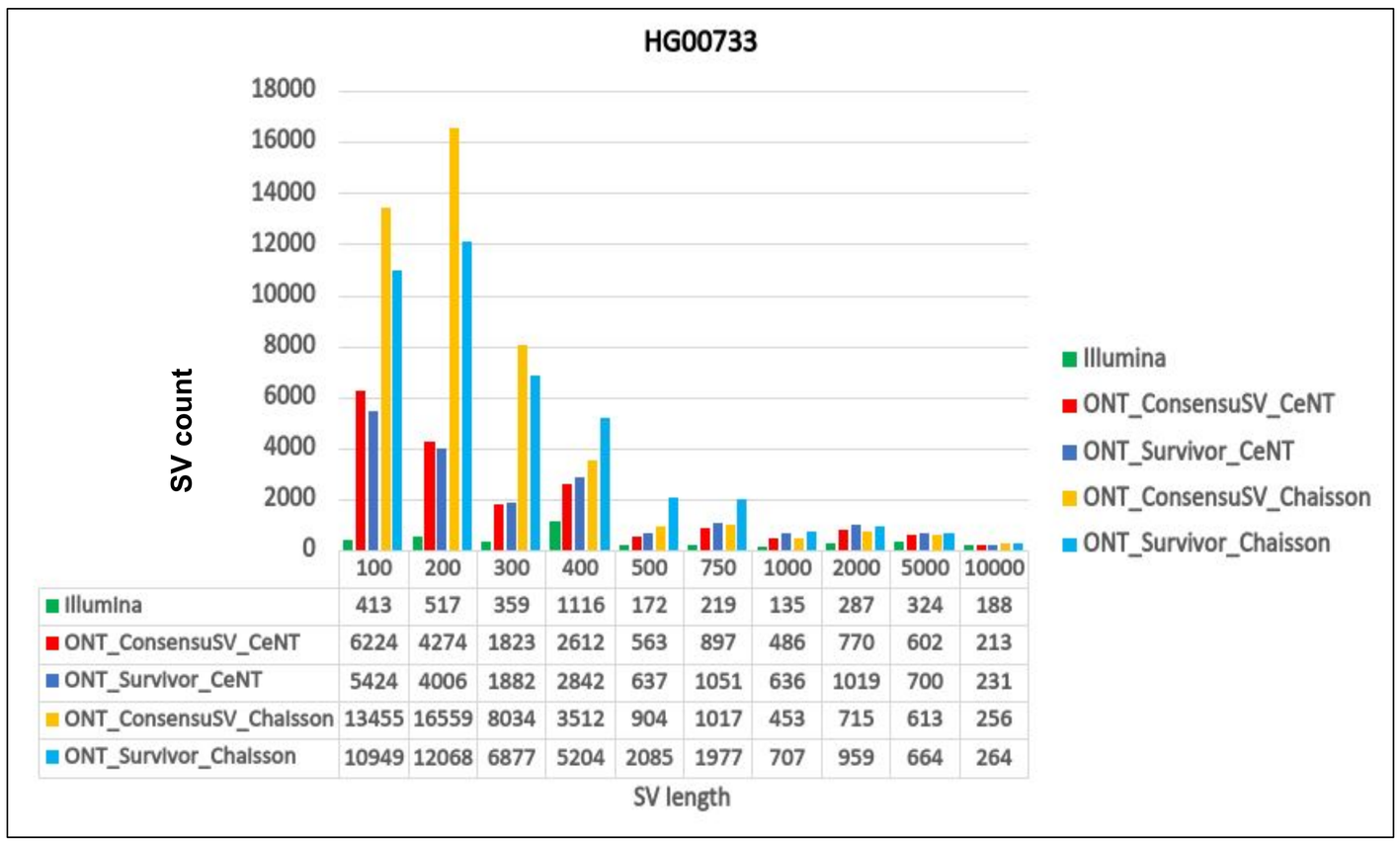
We present a comparative analysis of whole genome sequencing (WGS) of three families from the 1000 Genomes Project. The samples (3 daughters from Trios families) were sequenced using long-read Oxford Nanopore Technology (ONT) and compared with short-read sequencing by Illumina. In our work, we focused on the structural variants (SV) - segments of DNA that are at least 50 base pairs in length, abundant in human genomes as reported by the 1000 Genomes Project. We designed an ensemble-based method that takes structural variants output from multiple SV callers, trains them using neural networks, and benchmarked on the gold set of Structural Variants (Chaisson, 2019) as a truth set from the 1000 Genomes Project.

2. Methods

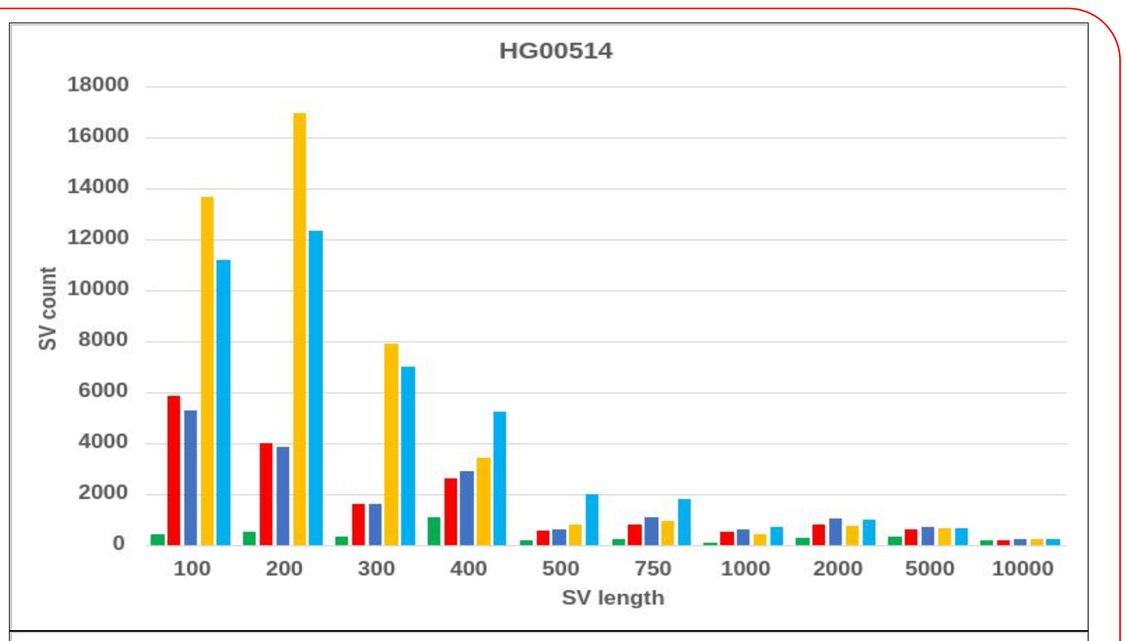


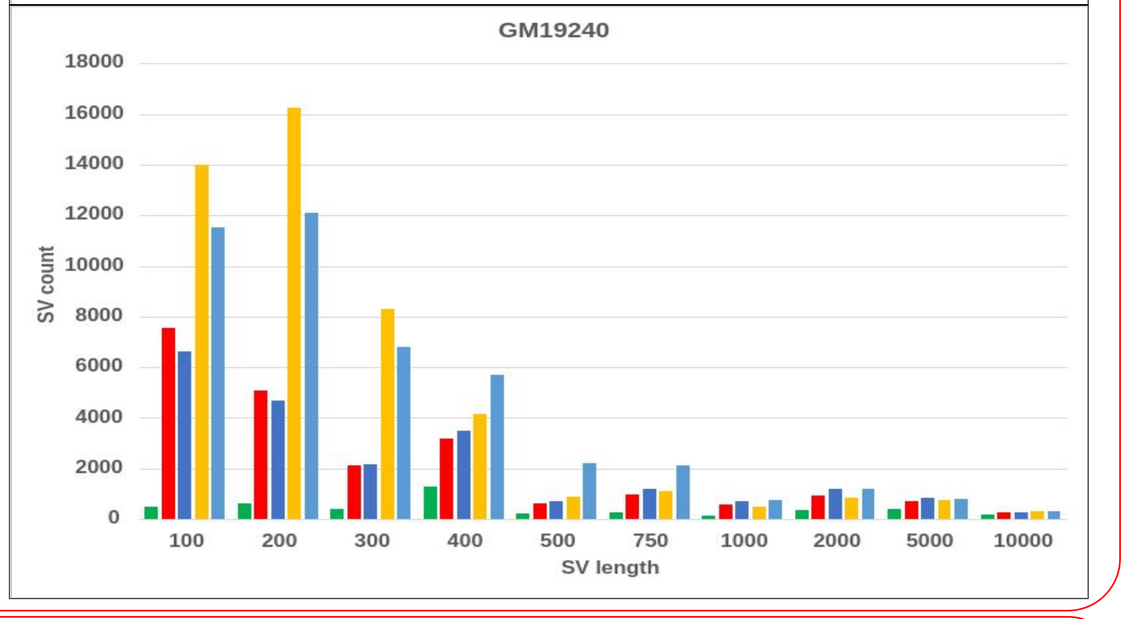


3. Results : Distribution of SV size and count of 3 daughters from Trios



	100	200	300	400	500	750	1000	2000	5000	10000
Illumina	413	517	359	1116	172	219	135	287	324	188
ONT_ConsensuSV_CeNT	6224	4274	1823	2612	563	897	486	770	602	213
ONT_Survivor_CeNT	5424	4006	1882	2842	637	1051	636	1019	700	231
ONT_ConsensuSV_Chaisson	13455	16559	8034	3512	904	1017	453	715	613	256
ONT_Survivor_Chaisson	10949	12068	6877	5204	2085	1977	707	959	664	264





Consensus SVs obtained from our in-house lower coverage (LC) ONT data (12x) are comparable with the ones from deep coverage (DC) ONT sequencing (30x) (Chaisson et al, 2019).

4. Conclusion

- We conclude that with the consensus approach we recovered comparable numbers of large-size SVs for both LC and DC datasets. Our result confirms the previously reported observation that for ONT sequencing where 10X coverage is good enough to identify SV breakpoints using on average 10kb long reads (Sedlazeck et al., 2017, Beyter et al., 2021).
- Finally, upon comparative analysis of SV distribution, we confirm that the Oxford Nanopore Technology long-read sequencing is superior to Illumina short-read across the whole range of structural variants sizes, even with significantly less coverage.

Acknowledgement: This work has been co-supported by Polish National Science Centre (2019/35/O/ST6/02484 and 2020/37/B/NZ2/03757). Computations were performed thanks to the Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology using Artificial Intelligence HPC platform financed by Polish Ministry of Science and Higher Education (decision no. 7054/IA/SP/2020 of 2020-08-28).

National Science Centre Poland · European Funds Smart Growth · FNP Foundation for Polish Science · European Union European Regional Development Fund · Faculty of Mathematics and Information Science