μ-PBWT: a lightweight r-indexing of the PBWT for storing and querying UK Biobank Data

Davide Cozzi [1] Massimiliano Rossi [2] Simone Rubinacci [3] Travis Gagie [4] Dominik Köppl [5,6] Christina Boucher [2] Paola Bonizzoni [1,*]

[1] University of Milano-Bicocca, Italy [2] University of Florida, USA [3] University of Lausanne, Switzerland [4] Dalhousie University, Canada [5] Tokyo Medical and Dental University, Japan [6] University of Muenster, Germany

[*] Corresponding author. paola.bonizzoni@unimib.it

Overview

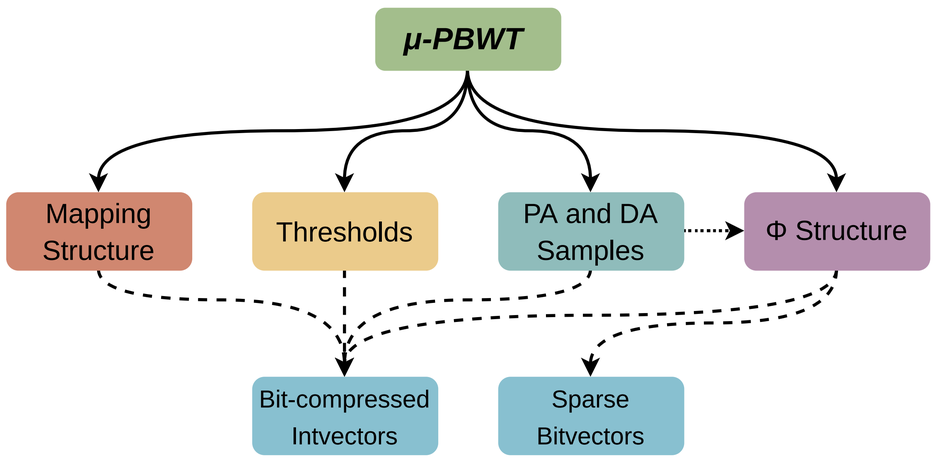
Introduction: Improved haplotype phasing in large cohorts is facilitating the comprehensive collection and study of variations at chromosome-level for genome evolution and clinical applications. The computational challenge moves nowadays to index and query this amount of data in a very efficient way.

Background: The Positional Burrows-Wheeler Transform (PBWT) is a data structure that indexes haplotype sequences and finds maximal haplotype matches in h sequences containing w variation sites in $\mathcal{O}(hw)$-time. However, the PBWT data structure does not allow for queries over panels that consist of several millions of haplotypes since the index must be loaded entirely in RAM, requiring $13hw$ bytes.

Contribution: We introduce μ-PBWT that leverages the r-index techniques for the run-length compressed PBWT (RLPBWT). By keeping in memory only a succinct representation of the RLPBWT, μ-PBWT efficiently computes set maximal matches (SMEMs) over the original panel, by scaling over UK Biobank data and reducing the memory usage up to a factor of 20% compared to the best current PBWT-based indexing. The index produced by μ-PBWT stores high-coverage whole genome sequencing data in about a third of the space of its BCF file.

Method

Our contribution is a novel approach for efficient sampling and storing at run boundaries the prefix (PA)/divergence (DA) arrays and additional information, having that μ-PBWT reduces the space of the PBWT for multiple queries SMEMs-finding from $\mathcal{O}(hw)$-space to $\mathcal{O}(r)$-space, having r as the number of runs in the PBWT.



Experiments

We demonstrate the performance of μ-PBWT by comparing it with the Durbin's PBWT, Syllable-PBWT, and BGT (a PBWT-based compact file format for haplotype sequences) on:

- all autosome panels from the 1000 Genomes Project (5K haplotypes and 1M–6M bi-allelic sites)
- chromosome 20 panel from UK Biobank high-coverage whole genome sequencing data (300K haplotypes and 13M bi-allelic sites)

Discussion

μ-PBWT is a lightweight index for the PBWT data structure for solving the SMEMs-finding problem. Experiments on 1000 Genomes Project data show memory reduction up to 80 times against Durbin's PBWT and a slight index size reduction against Syllable-PBWT, one of the lightest PBWT-based indices. On UK Biobank data, μ-PBWT can store an index in about a third of the space of its binary-format counterpart.

We note that all the indices generated by μ-PBWT are loaded in less than 30 seconds on a commodity laptop, ensuring their practical use.

Future Developments

Memory usage results achieved by μ-PBWT suggest that our approach can scale on large whole genome genotype data for phasing and imputation.

Moreover, other future developments are in the perspective of missing data and a possible parallelization of the index construction.

Availability

- https://github.com/dlcgold/muPBWT
- https://bioconda.github.io/recipes/mupbwt/README.html

References

[1] Davide Cozzi et al. "μ-PBWT: Enabling the Storage and Use of UK Biobank Data on a Commodity Laptop". In: *bioRxiv (updated version to appear in Oxford Bioinformatics as "μ-PBWT: a lightweight r-indexing of the PBWT for storing and querying UK Biobank Data")* (2023).

[2] Paola Bonizzoni et al. "Compressed Data Structures for Population-Scale Positional Burrows–Wheeler Transforms". In: *bioRxiv* (2022).

[3] Richard Durbin. "Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT)". In: *Bioinformatics* 30.9 (2014).

1000 Genomes Project Results

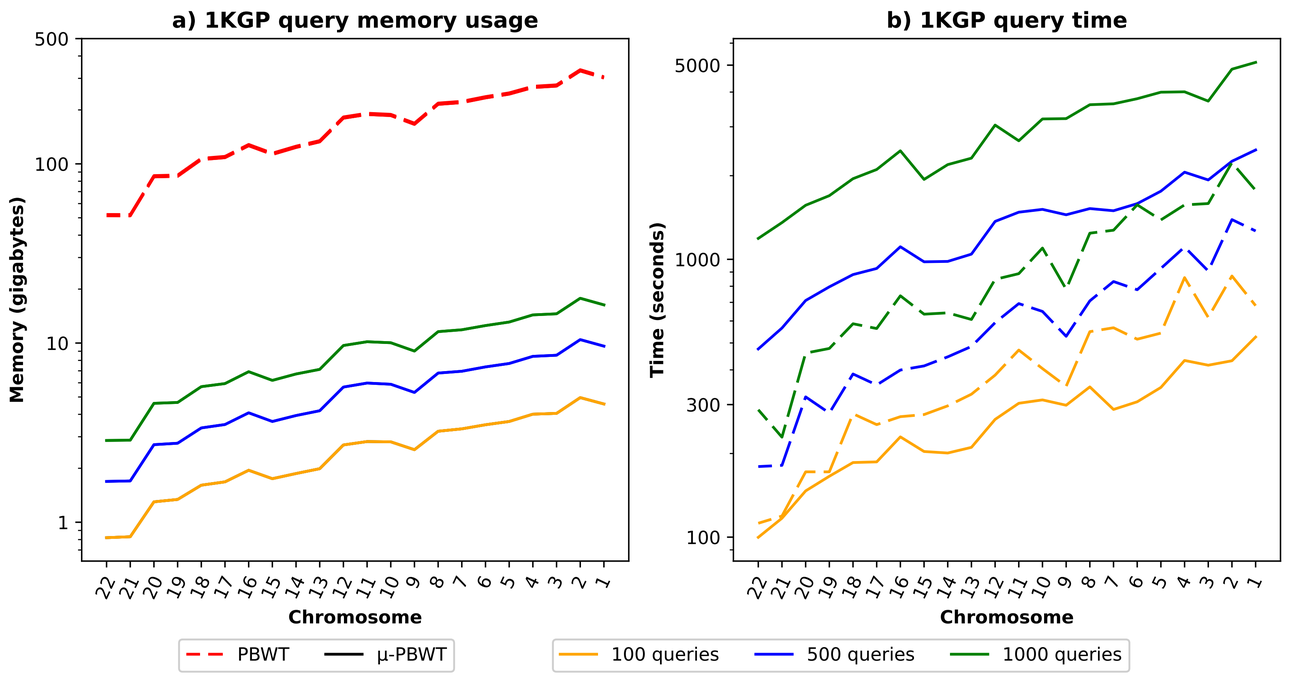


Figure: Space (on the left) and time (on the right) comparison with PBWT on the 1000 Genomes Project data (4K rows and 1M–6M columns) for finding SMEMs with 100/500/1K queries.

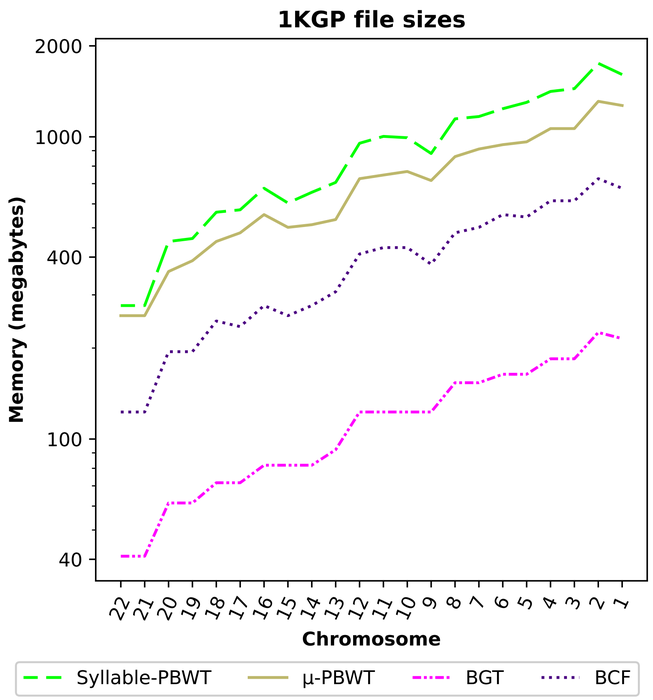


Figure: Comparison among the index sizes of μ-PBWT, Syllable-PBWT, and BGT, against the input file BCF file size on the 1000 Genomes Project data (4K rows and 1M–6M columns).

UK Biobank Results

	GB			hh:mm
Sites	Size BCF	μ-PBWT Build Memory peak		Build Time
865,267	1.9	0.88	2.27	06:25
880,899	2	0.85	2.22	06:28
961,591	2.1	0.77	2.05	07:04
917,468	2	0.73	1.97	06:47
931,010	2	0.71	1.92	06:53
1,919,134	4.2	1.20	3.06	13:54
1,436,549	2.8	0.99	2.63	10:25
1,056,144	2.2	0.76	2.06	07:42
955,970	2	0.79	2.09	06:56
923,178	2	0.80	2.12	06:44
911,452	2	0.81	2.13	06:45
925,442	2	0.84	2.20	06:49
1,096,089	2.4	0.93	2.42	08:00
13,780,193	**29.6**	**11.06**	**29.15**	**13:54**

Table: Results on UK Biobank chromosome 20 data (300K rows, number of columns/sites on the left). Total results in the last row.