Dirk D., Z Liu, Matthew C, Jared T,Zamin I, Richard D, Shane A. Mc, and Thomas M. Keane
Using reference-free Compressed data structures to analyze sequencing from thousands of Human Genomes Genome Res. 2017 Feb; 27(2):300-309

The paper "Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes" addresses about the idea of a Population BWT a compression method to store and index of 2705 Samples from the 1000 Genomes Project. This was work was undertaken to reduce the overall size of the genome by error correction, BWT (Burrows Wheeler Transform), and FM (Ferragina& Manzini) Indexing to decrease query time. The purpose of this paper is to assess the read support for 1000 genomes data for every base position of the two human reference genomes GRCh37 and GRCh38, show how accurately and rapidly SNP and Indel genotyping carried across genomes in population BWT and lastly to carry out non reference queries against the BWT to search for presence of all known viral genomes.

Methods : Firstly, all the sequencing data was downloaded from FASTQ format from the 1000GP ftp site (phase 3). This data was corrected using Cortex Software based on De Bruijin Graph assembler. It merged all the 1000 GP data and one overall cortex graph was created. The reference genome was then parsed to annotate k-mers. Also, any duplicates and reads below 73 and less than Q20 were removed. Then corrected reads were sorted in reverse alphabetical order. Then the experimenters prepared the BWT and FM index. Firstly, based on last 2 bp of reads 16 partitions were created. These partitions were then passed through a software called SGA. This software outputs BWT strings in RLE (Run In Length encoding). Each and every byte produced consist of two parts, first, the five different characters (3 bits in length) and secondly the last five bytes encode for the runs for these character up to 31. Secondly, we know that the basis of FM indexing is last to first column mapping which keeps the query time linear in nature, therefore, all the BWT strings were FM indexed for faster query. In the final step, they created a physical server on the

principle of Queue Data structure which means first request in will be handled first and then the rest. After setting up the server population BWT is now read for analysis. Firstly, they generated 31-mers forward sets from the GRCh37 and GRCh38 at every positions and then queried them against the population BWT both in forward and reverse direction. Similarly, they performed SNP genotyping and Indel genotyping on the population BWT using 1000Gp Illumina chip data and GIAB (genome in a bottle data) respectively. Firstly, SNP genotyping was carried out with population BWT by generating a reference and alternate allele using 99bp of flanking sequence for each site. 34 mers were created and Local Smith Waterman alignment was performed. Secondly, for Indel genotyping they used 100bp flanking sequence and then generated 25 mers from these sequences to query Population BWT for matching reads. Lastly, for viral genome analysis, firstly a Kraken Database was generated by reference genomes of human, bacteria, plasmid and viruses reference genomes. Final set of data was prepared by querying the data with GENBank and classifying the alignments.

Results and Assertions: A new compressed population BWT using RLO (Reverse lexicographical order) was ready for use. To test if they made the right choice regarding the RLO, they perform different arranging techniques to see its impact on the BWT with increasing number of samples. A sub-linear growth was seen when they used RLO and therefore, data could be compressed at the highest level. This verifies their choice of RLO to be optimal for compression.

Secondly, support for human resource assemblies and variation was tested using GRCh37 and GRCh38 and was seen population BWT does cover 99.97%. Then they also tested the read coverage and completeness of BWT and found that during transition from GRCh37 to GRCh 38 the population BWT support was lost (around 3.1Mbp) but 7.5x sequences were gained. This result verifies the completeness and coverage of the population BWT.

Thirdly, they did the reference free population genotyping on the Population BWT and found that it is slightly inconsistent and higher for the population BWT genotyping (1.82%) compared to the GATK (0.81%) and SAMtools (0.73%) which are reference based tools. On the other hand, for heterozygous SNPs, the inconsistency rate of 2% vs. 2.17% for GATK, and 3.41% for SAMtools was observed. In the end, we can assert that population BWT is able to compete well with the previous reference based tools but not as well. Furthermore, they also compared for runtime for genotyping a certain sample on population BWT vs GATK and SAMtools and saw that BWT took five times longer to complete genotyping but produced better quality results for all 31mer queries.

Lastly, in the viral genome analysis, when a non-reference query for was run against their population BWT, it was faster (2days vs 7days). Also, Population BWT corresponded really well to find presence all viral genomes by running non reference queries. It also showed that most of these unknown reads were mostly herpesviruses. This result proved that the population BWT is good enough to be used for some hypothesis based analysis and was much faster.

Critique: The article is not clear in terms of the title. In this paper, the experimenters are able to convey that population BWT is better compression method than the rest of the available methods but there is not enough evidence that it is useful for analysis because the only analysis they perform is the viral genome analysis, and all the rest methods were used to convey that population BWT is fast, rapid and can be used for studies and it's a better method to applied for finding human reference assemblies and variation, SNP and Indel genotyping.

Secondly, there was a discrepancy in terms how they agreed to use 31mers sets of queries for both reference assemblies and the population BWT when doing reference genome analysis. Also, for SNP genotyping they used 34mers queries and similarly, for Indel genotyping they used 25mer

queries against the population BWT. It remains ambiguous for the reader as to why they used these numbers. Even if we decide to agree that they were optimum k-mers, a graph would have been helpful to show how different of k-mers affected the optimality of these studies.

Thirdly, the procedures and methods are not well defined as there are many grey areas in terms of outside software used (Cortex, SGA, preparation of Kraken Databases). It seems like the data went into a black box and came out clean. The parameter and settings used to process the reads were ambiguous. However, the experimenters do point you towards the Git-hub link but the trade-off here is that, the reader might have to go through long ReadMe files. Also, the supplemental data can be improved with adding parameters used whenever an outside software is used. Such ambiguity regarding the software makes the study less replicable.

Overall Viewpoint: Currently, in the field of bioinformatics a lot of data is generated every day, I agree with experimenter's approach as they were able to lower the amount of the space used to store genomic data using BWT and FM indexing. Although, there were some grey areas in terms of querying speed and performance of this approach while genotyping, whereas in terms of coverage and hypothetical study, it did perform really well. Secondly, I do agree with the completeness and coverage of the populations BWT but results show that the population BWT was not able to perform as well when compared to SAMtools and GATK while SNP genotyping and Indel genotyping. For better population scale analysis, more evolution is required in terms of storage of the data but there is no point of storing data in a compression based model if the query time is only as good as what is available today. An all-round approach is what we should look for in the future. It is only possible with better coordination and communication between experts of different fields which is the core of bioinformatics.