Indexing K-mers in Linear Space with Application to SNP Detection

F. Andreace, M. Marcolin, M. Comin

May 9, 2021

Motivation: Sequencing technologies have provided the basis of most modern genome sequencing studies due to their high base-level accuracy and relatively low cost. One of the central obstacles is mapping reads to the human reference genome. The reliance on a single reference human genome could introduce substantial biases in downstream analyses. Moreover, including known variants in the reference makes read mapping, variant calling, and genotyping variant- aware. However, reads mapping is becoming a computationally intensive step for most genomic studies. Alignment-free methods have been used to save compute time and memory by avoiding the cost of full-scale alignment (Vinga and Almeida, Bioinformatics 2003). Recently, alignment-free approaches have been applied to SNP genotyping by (Shajii et al., Bioinformatics 2016) and (Sun and Medvedev, Bioinformatics 2019). They introduce two SNP genotyping tools named LAVA and VarGeno, respectively, which build an index from known SNPs (e.g. dbSNP) and then use approximate k-mer matching to genotype the donor from sequencing data. LAVA and VarGeno are reported to perform 4 to 30 times faster than a standard alignment-based genotyping pipeline while achieving comparable accuracy. However, they require a large amount of memory, about 60GB. In this work, we introduce GenoLight that will address this problem with new efficient data structures.

Methods: The basic idea of LAVA and VarGeno is to build two dictionaries of k-mers, one with all the k-mers from the reference genome, and another with the k-mers covering known SNPs. In the genotyping step, instead of an expensive alignment procedure, the reads are broken down into their constituent k-mers and these k-mers are

searched into the two dictionaries, in order to detect putative SNPs. In LAVA, these dictionaries are implemented with a hash table where all k-mers are explicitly stored, instead, VarGeno uses a Bloom filter. Both these data structures need to be loaded in memory for efficient queries, and they require about 60-63 GB of RAM. However, the size of these dictionaries can be reduced, in fact, most of the information carried by a k-mer is redundant. Given two overlapping k-mers, it is possible to reassemble them into a single (k+1)-mer, thus reducing the storage requirement by k-1 bases. In GenoLight, a set of k-mers, associated with known SNPs, is assembled into a string, or set of strings, that contains all of them. Then, an FM-index is built to further reduce the size of the dictionary and to support fast queries. This procedure allows us to store the whole dictionary in linear form reducing the memory requirements from O(kn) to O(n) without losing information. This problem is closely related to the representation and compression of k-mers sets, which has attracted the attention of many researchers recently (K. Brinda et., Genome Biology 2021) and (A. Rahman, P. Medvedev, Recomb 2020). In GenoLight, in order to reconstruct the SNP position, we need to store for each k-mers in the re-assembled dictionary: its direction, forward or reverse; the original position in the reference; and the associated SNP. Then, the reads' k-mers are reached into the re-assemble k-mers dictionary, and its auxiliary structure, in order to detect candidate SNPs.

Results: The analysis focuses on two fundamental points, the time taken to complete the genotyping process, and the accuracy of the results obtained by the tools: GenoLight, LAVA, VarGeno, and the standard alignment-based genotyping pipeline. The datasets used in this study are two sets of real reads from the 1000 Genomes Project: SRR622461 (low coverage 6X) and SRR622457 (high coverage 10X). We used an up-to-date high-quality genotype annotation generated by the Genome in a Bottle Consortium, widely used for benchmarking, and dbSNP and Affymertix as reference SNPs datasets. We consider the standard alignment-based genotyping pipeline as the reference to compare the three algorithms. As expected, the standard pipeline is always more time-consuming while requiring only 4 GB of memory. As shown in Table 1, the precision of all tools is very close to the standard pipeline. Only Lava shows a lower precision in some cases. VarGeno is always the fastest method, thanks to the Bloom filter, and Lava is the slowest, while the memory required is similar (57-63 GB). GenoLight reports precision results comparable to VarGeno, but it uses between 5 to 10 times less memory than the other two tools. GenoLight is a new tool for reads genotyping that achieves almost the precision levels and the memory usage of the standard pipeline while being significantly faster, thus resolving the high memory issue of VarGeno and Lava, while maintaining good

time performances.

Dataset	SNP db	Algorithm	Precision	Time (minutes)	RAM (GB)
Low Coverage	None	Standard Pipeline	0.9305	1929	4
	dbSNP	VarGeno	0.9116	59	63.2
		Lava	0.8199	506	61
		GenoLight	0.9124	406	12.5
	Affymetrix	VarGeno	0.9353	44	58.9
		Lava	0.9352	318	57.6
		GenoLight	0.9347	295	6
High Coverage	None	Standard Pipeline	0.9691	1888	4
	dbSNP	VarGeno	0.9497	80	63.242
		Lava	0.8459	723	60
		GenoLight	0.9512	563	12.5
	Affymetrix	VarGeno	0.9771	55	59
		Lava	0.9746	435	57.6
		GenoLight	0.9767	396	6

Table 1: Performance results of the genotyping for all tools on various datasets.