BITS :: Call for Abstracts 2019 - Oral communication

Type	Oral communication
Session	Algorithms in Bioinformatics
Title	MALVA: genotyping by Mapping-free ALlele detection of known VAriants
All Authors	Bernardini G(1), Bonizzoni P(1), Denti L(1), Previtali M(1), Schönhuth A(2)
Affiliation	

(1) Department of Informatics, Systems, and Communication (DISCo). University of Milano-Bicocca. Milano
(2) Centrum Wiskunde & Informatica. Amsterdam, The Netherlands

Motivation

The amount of genetic variation discovered and characterized in human populations is huge, and is growing rapidly with the widespread availability of modern sequencing technologies. Such a great deal of variation data, that accounts for human diversity, leads to various challenging computational tasks, including variant calling and genotyping of newly sequenced individuals.

The standard pipelines for addressing these problems includes aligning sequenced reads with softwares like BWA and Bowtie and then calling the genotypes (e.g. with GATK or BCFtools). Such an approach, though, can be highly time consuming, thus impractical for clinical applications, where time is often an issue.

In medical settings where the discovery of new variants is not desired, but, rather, what is important is to know the genotype at certain loci that are already established to be of medical relevance, alignment-free methods come to the aid. Even though such methods are up to an order of magnitude faster than the usual alignment-based approaches, they only focus on isolated, bi-allelic SNPs, providing limited support for multi-allelic SNPs, indels, and genomic regions with high variant density. Short insertions and deletions of nucleotides (indels), though, are of particular clinical interest: they are believed to represent around 16% to 25% of human genetic polymorphism and they can be associated with a number of human diseases, e.g., cystic fibrosis and lung cancer.

To address the limitations of these approaches, we introduce MALVA, a fast and lightweight mapping-free method to genotype known (i.e., previously characterized) variants directly from a NGS sample. MALVA is the first mapping-free tool that is able to genotype multi-allelic SNPs and indels, even in high density genomic regions, and to effectively handle a huge number of variants such as those provided by the 1000 Genome Project.

Methods

MALVA takes as input a reference genome, a VCF file, and a read sample; it outputs a VCF file containing a genotype for each variant. MALVA leverages on the notion of signature of an allele: each allele is assigned a set of k-mers, which allows to efficiently model SNPs, indels, and close variants. The general idea of MALVA is to use the frequencies of the signatures in the sample to call the genotypes. The method works under the assumption that if an allele is included in the genome then at least one of its signatures must exist as substrings in multiple reads.

The method is composed of three steps.

In the first step, MALVA computes the set of signatures of all the alleles of the input variants and stores them in a Bloom filter. By cleverly iterating over the VCF file and the genotype information associated to each variant, we are able to compute the signatures on the fly, without reconstructing the whole haplotypes.

In the second step, MALVA analyzes the k-mer statistics produced by KMC to assign a weight to each k-mer contained in the Bloom filter built in the previous step. In other words, MALVA assigns to each k-mer the number of times it occurs in the sample. By using these weights, MALVA computes the weight of each signature and then the expected coverage of each allele.

Finally, in the last step, MALVA uses the coverages computed in the previous step to call the genotypes. Based on the well-known Bayes' formula, we design a new rule to genotype multi-allelic variants, i.e., variants with more than one alternate allele. Specifically, we extend the approach proposed in LAVA to multi-allelic variants.

Results

We performed an experimental analysis on real data to evaluate the real feasibility of our method, comparing MALVA to one mapping-free method (VarGeno) and to two different alignment-based pipelines (BWA-MEM followed by BCFtools and by GATK, respectively). Each tool was evaluated in terms of variant calling accuracy and efficiency.

We tested the tools using a 30x Illumina WGS dataset of the well-studied NA12878 individual provided by the GIAB consortium, as the variant calls provided for this individual are highly reliable and can be effectively used to assess the accuracy of the considered methods.

Our experimental evaluation shows that MALVA requires one order of magnitude less time to genotype a donor than alignment-based pipelines, providing similar accuracy on SNPs and much higher accuracy on indels.

Since VarGeno could not complete the analysis of the whole-genome dataset, we created a smaller one by halving the number of chromosomes. From this evaluation, VarGeno results faster than MALVA but more memory expensive. While the two tools achieved similar accuracy on SNPs, thanks to the novel notion of signature of an allele, MALVA was also able to genotype indels. The notion of signature is also convenient for managing close variants. Finally, unlike VarGeno, MALVA proved to be able to genotype multi-allelic variants, thanks to its improved genotyping module.

Info
-
Figure
-

Availability	https://algolab.github.io/malva/

Corresponding Author	
Name, Surname	Luca, Denti
Email	l.denti@campus.unimib.it

Submitted on	25.04.2019