MALVA: genotyping by Mapping-free ALlele detection of known Variants

Giulia Bernardini[1], **Paola Bonizzoni**[1], **Luca Denti**[1,*], **Marco Previtali**[1], **Alexander Schönhuth**[2]

[1]Dept. of Informatics, Systems, and Communication (DISCo), University of Milano–Bicocca, Milan (Italy)
[2]Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
*l.denti@campus.unimib.it

The amount of genetic variations discovered and characterized in human populations is huge, and is growing rapidly with the widespread availability of modern sequencing technologies. For example, the 1000 Genomes projects sequenced 2504 individuals and identified more than 84 millions of variants (SNPs and indels). In clinical context where the discovery of new variants is not desired, any investigation can be restricted to the analysis and the genotyping of the set of known variants that are established to be of medical relevance. In such a context, time is often an issue.

State of the art

Alignment-based approaches are very accurate but they are too slow. Assembly-based approaches are a bit faster but they are not adequately accurate. Alignment-free approaches are fast and show great accuracy (when variant discovering is not required).

Main Contribution

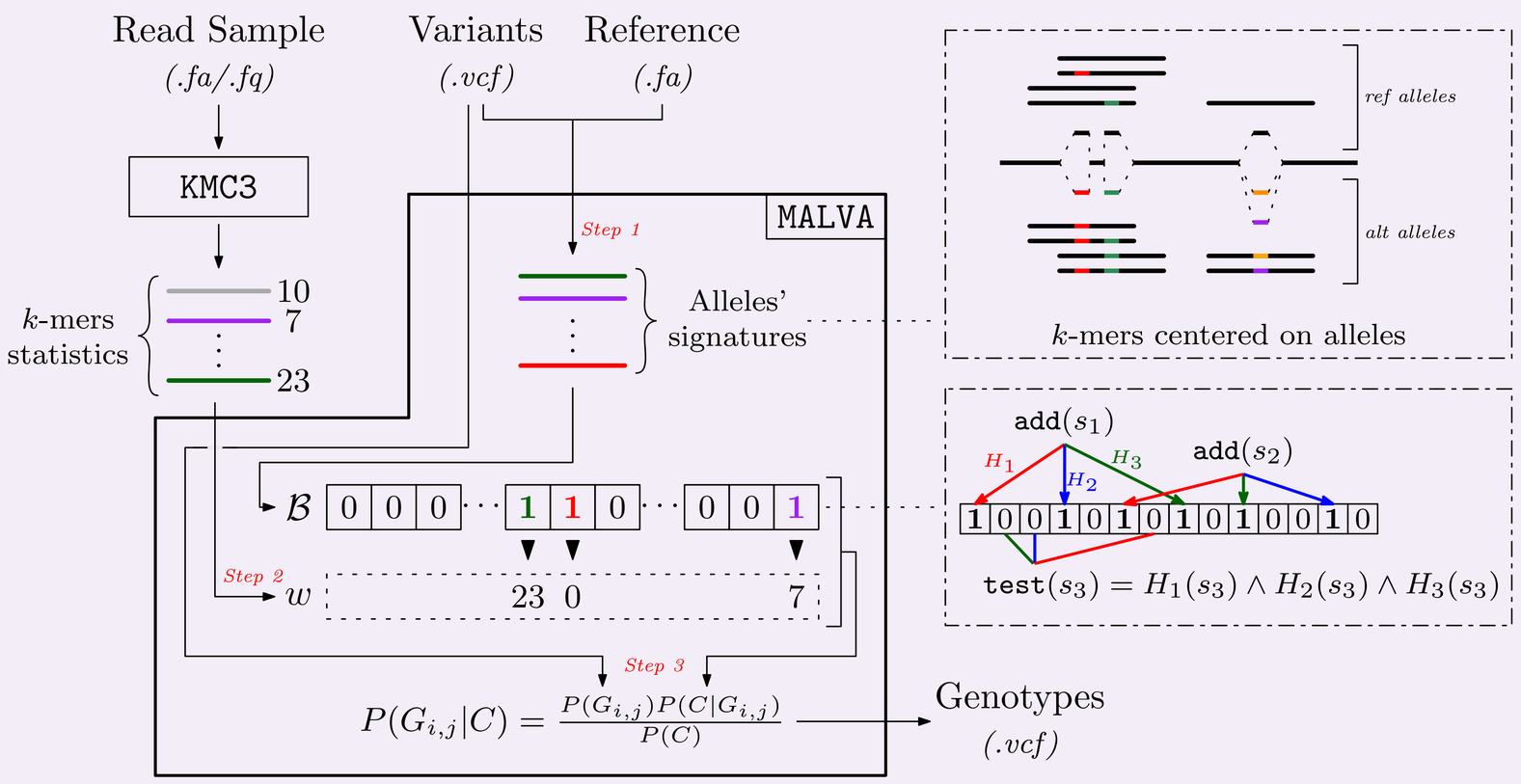
`MALVA` is a fast and lightweight mapping-free method to genotype a set of known variants directly from a sample of reads. `MALVA` is the first mapping-free tool that is able to genotype multi-allelic SNPs and indels, even in high density genomic regions, and to effectively handle a huge number of variants.

Experiments

- GRCh37 primary assembly
- `VCF` from phase3 of the 1000 Genomes Project
- 30x WGS sample from `NA12878`

[+ halved version]

Method

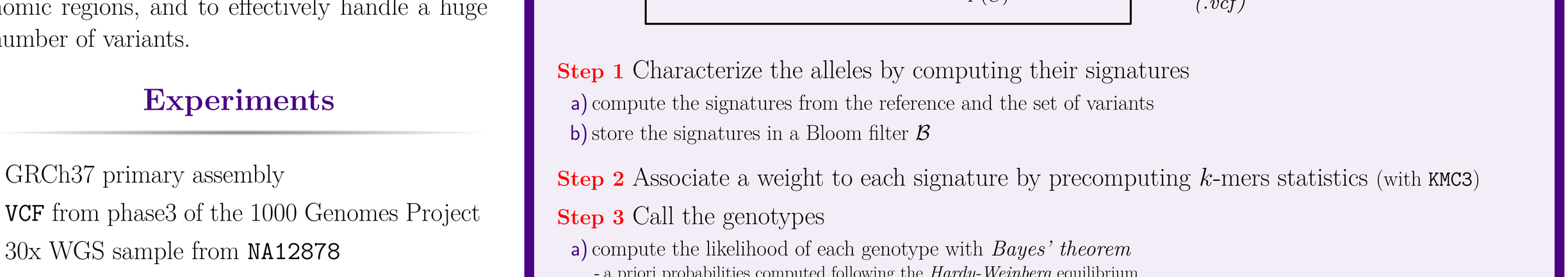


Step 1 Characterize the alleles by computing their signatures
- a) compute the signatures from the reference and the set of variants
- b) store the signatures in a Bloom filter $\mathcal{B}$

Step 2 Associate a weight to each signature by precomputing k-mers statistics (with `KMC3`)

Step 3 Call the genotypes
- a) compute the likelihood of each genotype with *Bayes' theorem*
  - a priori probabilities computed following the *Hardy-Weinberg* equilibrium
  - conditional probabilities computed by extending `LAVA` approach to multi-allelic variants
- b) output the genotype with the highest likelihood

Results



Discussion

- `MALVA` is a valid alternative to alignment-based approaches, *especially when time efficiency is crucial*
- `MALVA` is fast and accurate
- on indels, `MALVA` provides even better results than the most widely adopted variant discovery tools
- `MALVA` is freely available at `https://algolab.github.io/malva/`

install with bioconda

References

- `MALVA`: Bernardini et al., *bioRxiv* (2019)
- `hap.py`: Krusche et al., *Nature biotechnology* (2019)
- `VarGeno`: Sun et al., *Bioinformatics* (2018)
- `DiscoSnp++`: Peterlongo et al., *bioRxiv* (2017)
- `LAVA`: Shajii et al., *Bioinformatics* (2016)
- `BCFtools`: Li, *Bioinformatics* (2011)
- `GATK`: McKenna et al., *Genome research* (2010)
- 1000 Genomes Project: Sudmant et al., *Nature* (2015)