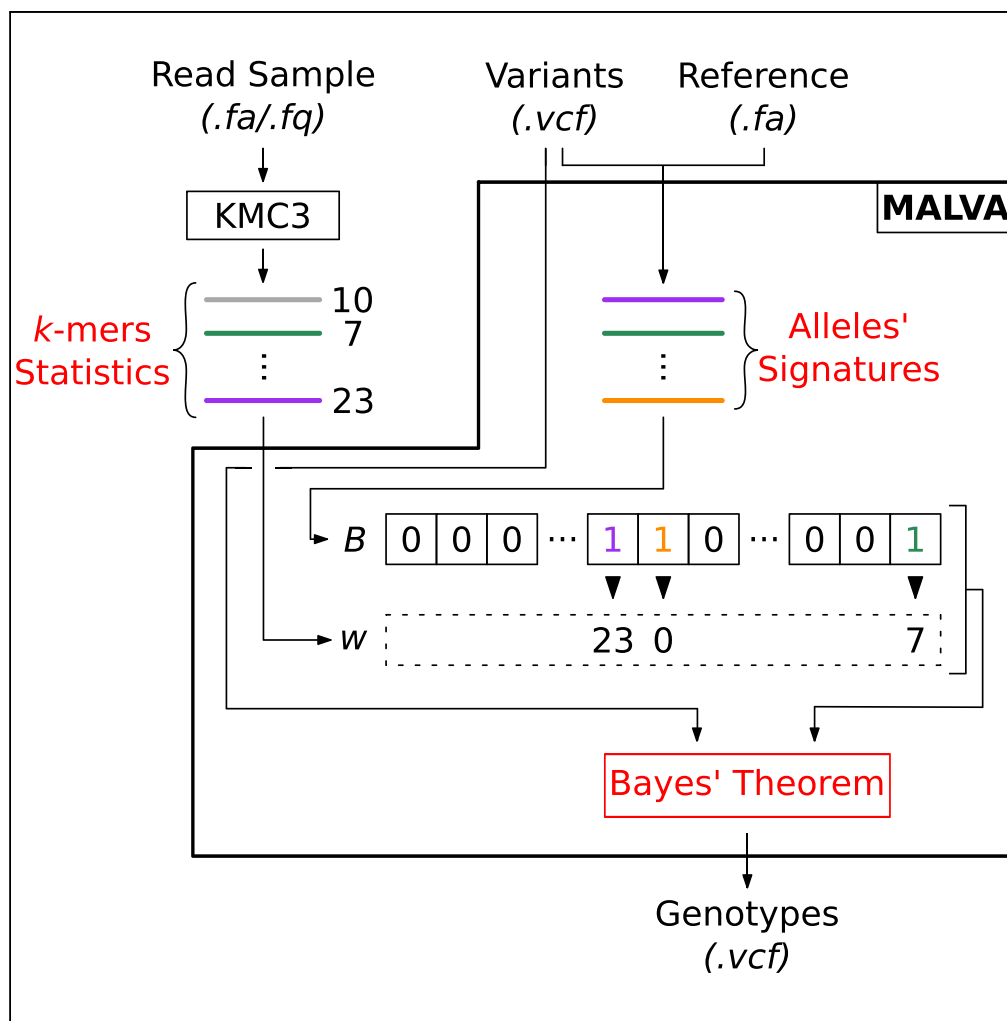


Special Issue: RECOMB-Seq 2019

Article

# MALVA: Genotyping by Mapping-free ALlele Detection of Known VARIants



Luca Denti, Marco Previtali, Giulia Bernardini, Alexander Schönhuth, Paola Bonizzoni

marco.previtali@unimib.it

HIGHLIGHTS

MALVA is able to genotype multi-allelic SNPs and indels without mapping reads

MALVA calls correctly more indels than the most widely adopted genotyping pipelines

Mapping-free approaches are as accurate as alignment-based ones, while being faster

Denti et al., iScience 18, 20–27 August 30, 2019 © 2019 The Authors. <https://doi.org/10.1016/j.isci.2019.07.011>



Special Issue: RECOMB-Seq 2019

## Article

# MALVA: Genotyping by Mapping-free ALlele Detection of Known VAriants

Luca Denti,<sup>1,3</sup> Marco Previtali,<sup>1,3,4,\*</sup> Giulia Bernardini,<sup>1</sup> Alexander Schönhuth,<sup>2</sup> and Paola Bonizzoni<sup>1</sup>

## SUMMARY

The amount of genetic variation discovered in human populations is growing rapidly leading to challenging computational tasks, such as variant calling. Standard methods for addressing this problem include read mapping, a computationally expensive procedure; thus, mapping-free tools have been proposed in recent years. These tools focus on isolated, biallelic SNPs, providing limited support for multi-allelic SNPs and short insertions and deletions of nucleotides (indels). Here we introduce MALVA, a mapping-free method to genotype an individual from a sample of reads. MALVA is the first mapping-free tool able to genotype multi-allelic SNPs and indels, even in high-density genomic regions, and to effectively handle a huge number of variants. MALVA requires one order of magnitude less time to genotype a donor than alignment-based pipelines, providing similar accuracy. Remarkably, on indels, MALVA provides even better results than the most widely adopted variant discovery tools.

## INTRODUCTION

The discovery and characterization of sequence variations in human populations is crucial in genetic studies. A prime challenge is to efficiently analyze the variations of a freshly sequenced individual with respect to a reference genome and the available genomic variations data. To reach this goal, the standard pipeline includes aligning sequenced reads with software like BWA (Li and Durbin, 2009) and Bowtie (Langmead and Salzberg, 2012) and then calling the genotypes (e.g., with GATK [McKenna et al., 2010] or BCFtools [Li, 2011]); such an approach, though, can be highly time consuming and is thus impractical for clinical applications, where time is often an issue. Typically, in diploid organisms variant calling requires SNPs and indel detection and the identification of the pairs of alleles for each position of the studied genome, called genotype.

Assembly-based methods such as Cortex (Iqbal et al., 2012) and discoSNP++ (Peterlongo et al., 2017) form another line of research: the main idea of such tools consist in assembling the reads in a de Bruijn graph and then analyzing the bubbles in this graph to detect the variants. However, since read assembly is a computationally expensive task, such tools are still highly time consuming. Recent tools for genotyping and variant calling like GraphTyper (Eggertsson et al., 2017) and vg (Sirén, 2017), which are based on a graph representation of a pan-genome to avoid biases introduced by considering only the information included in a set of genomes (Computational Pan-Genomics Consortium, 2016), are nevertheless heavy in both computational space and time. Moreover, the size of indexes of variation graphs may be subjected to an exponential growth in the number of variants included, and indexes typically require a great deal of computational resources to be updated with newly discovered variants. When the task is to call the genotype in positions where variants have been previously annotated, alignment-free methods come to the aid. This is a typical case in a medical setting, 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.

Recent mapping-free genotyping tools such as LAVA (Shajii et al., 2016) and VarGeno (Sun and Medvedev, 2018) are word-based methods that, given a list of known SNP loci, call SNPs as either mutant or wild-type up to an order of magnitude faster than the usual alignment-based methods. A major shortcoming of these tools is the large memory requirement, which can easily exceed hundreds of GB of RAM. Their strategy is to create a dictionary for both the reference genome and the SNP list that maps each  $k$ -mer to the positions at which it appears and then to call variants from the reads by evaluating  $k$ -mers frequency. FastGT (Pajuste et al., 2017) is yet another  $k$ -mer-based method to genotype sequencing data: it strongly relies

<sup>1</sup>Department of Computer Sciences, Systems and Communications, University of Milan-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

<sup>2</sup>Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands

<sup>3</sup>These authors contributed equally

<sup>4</sup>Lead Contact

\*Correspondence: marco.previtali@unimib.it  
<https://doi.org/10.1016/j.isci.2019.07.011>



on a pre-compiled database of biallelic single nucleotide variants (SNVs) and corresponding  $k$ -mers, obtained by subjecting the  $k$ -mers that overlap known SNVs to several filtering steps. Such filters remove from the database the SNPs for which unique  $k$ -mers (i.e., not occurring elsewhere in the reference genome) are not observed, those that are closely located (i.e., that are less than  $k$  bases apart), and others; after the filtering steps, only 64% of biallelic SNVs survive and are therefore identifiable. These tools implement strategies to represent and analyze SNPs that improve the time performance but, on the other hand, do not allow to model indels and close variants.

Short insertions and deletions of nucleotides (*indels*) are believed to represent around 16%–25% of human genetic polymorphism (Mills et al., 2006). The presence of indels can be associated with a number of human diseases (Hasan et al., 2015; Ku et al., 2010; Mullaney et al., 2010): for instance, cystic fibrosis (Mullaney et al., 2010), lung cancer (Sequist et al., 2008), Mendelian disorders (MacArthur and Tyler-Smith, 2010), and Bloom syndrome (Kaneo et al., 1996) are all known to be closely correlated to indels. Indels are particularly challenging to call from NGS data, because mapping is more difficult when the reads overlap with indels (Montgomery et al., 2013).

Recently, mapping-free strategy was also applied to the discovery of *de novo* variants (Standage et al., 2019), i.e., variants that exist in a child but do not exist in both its parents.

In this paper we introduce MALVA, a rapid, lightweight, alignment-free method to genotype known (i.e., previously characterized) variants, including indels and close SNPs, in a sample of reads. MALVA is a word-based method: each allele of each known variant is assigned a *signature* in the form of a set of  $k$ -mers, which allows to efficiently model indels and close variants. The genotypes will be called according to the frequency of such signatures in the input reads. Based on the well-known Bayes' formula, we also design a new rule to genotype multi-allelic variants (i.e., variants such that more than one alternate allele is known): even if such variants are trickier to genotype than biallelic ones, we are still able to achieve high precision and recall, as revealed in the real-data experiments we conducted. MALVA directly analyzes a sample leveraging on the information of the variants included in a VCF file, which is the standard format released by the 1000 Genomes Project (Sudmant et al., 2015) (1KGP from now on). To the best of our knowledge, MALVA is the first mapping-free tool able to call indels. Moreover, it proved to be the only such tool capable of handling the huge number of variants included in the latest version of the VCF released by the 1KGP.

## RESULTS

In this section we will describe the implementation details of MALVA and we will provide an experimental analysis on real data. All the analyses were performed on a 64-bit Linux (Kernel 4.4.0) system equipped with four 8-core Intel Xeon 2.30 GHz processors and 256 GB of RAM.

We performed an experimental analysis on real data to evaluate the real feasibility of our method, comparing MALVA to one mapping-free method, one assembly-based approach, and two different alignment-based pipelines. Among the mapping-free methods proposed in the literature we chose VarGeno, as it is an improved version of LAVA that provides better efficiency and accuracy (Sun and Medvedev, 2018). For completeness, we included in our evaluation `discoSNP++` (assembly-based) and the two most widely used alignment-based pipelines, denoted by `BCFtools` and `GATK`, respectively. The pipeline denoted by `BCFtools` consists of an alignment step performed with `BWA-MEM` (Li and Durbin, 2009) followed by a variant discovering step performed using `BCFtools` (Li, 2011). The latter consists of an alignment step performed with `BWA-MEM` and a variant discovering step performed with `GATK` (McKenna et al., 2010), as recommended by the *GATK Best Practices* (DePristo et al., 2011).

MALVA was run setting  $k_s$  equal to 47,  $k_c$  equal to 53,  $\epsilon$  equal to 0.1%, and Bloom filters size equal to 8 GB and considering the genotype data and the *a priori* frequencies of the alleles of the EUR population, since the individual under analysis is part of it.

We tested the tools using the Illumina WGS dataset of the well-studied NA12878 individual provided by the Genome In A Bottle (GIAB) consortium (Zook et al., 2014). We chose this individual because the variant calls provided are highly reliable and can be effectively used to assess the precision and the recall of the considered methods. We downloaded the alignments of its 30x downsampled version and used

Dataset	Tool	P <sub>SNP</sub>	R <sub>SNP</sub>	P <sub>INDEL</sub>	R <sub>INDEL</sub>	Time (hh:mm)	RAM (GB)
HalfGenome	MALVA	93.8%	91.1%	86.0%	81.4%	04:33	30
HalfGenome	VarGeno	97.5%	88.1%	39.5%	0.1%	02:31	52
HalfGenome	discoSnp++	89.5%	39.3%	80.8%	24.2%	07:45	7
HalfGenome	BCFtools	91.2%	94.8%	44.9%	55.4%	24:35	6
HalfGenome	GATK	91.7%	95.1%	53.2%	79.9%	34:43	32
FullGenome	MALVA	92.5%	90.0%	85.0%	80.6%	09:18	39
FullGenome	VarGeno	–	–	–	–	–	–
FullGenome	discoSnp++	86.9%	37.7%	80.0%	22.6%	14:19	9
FullGenome	BCFtools	91.6%	94.4%	44.7%	54.6%	54:23	9
FullGenome	GATK	92.1%	94.7%	53.2%	79.2%	73:36	33

**Table 1. Accuracy and Efficiency Results on the HalfGenome and FullGenome Datasets**

For each dataset, we reported the values of Precision (P) and Recall (R) obtained by the considered tools on both SNPs and indels. The efficiency results are shown in terms of wall clock time and peak memory usage. VarGeno could not complete the analysis of the FullGenome dataset; thus, we did not report its results on this dataset. See also Figure S1.

SAMtools (Li et al., 2009) to extract the corresponding FASTQ file, obtaining 696,168,435 150-bp-long reads. As reference genome and set of known variants, we used the GRCh37 primary assembly and the VCF files provided by Phase 3 of the 1KGP (Sudmant et al., 2015). These VCF files contain a total of 84,739,838 variants, the phased genotype information of 2,504 individuals, and the *a priori* frequency of each allele of each variant of five populations. As stated earlier, from this VCF, in our evaluation MALVA extracted and considered only the individuals from the EUR population, for a total of 502 individuals. We note that we also removed the NA12878 individual from the input to better analyze MALVA and its capability in genotyping an unknown individual.

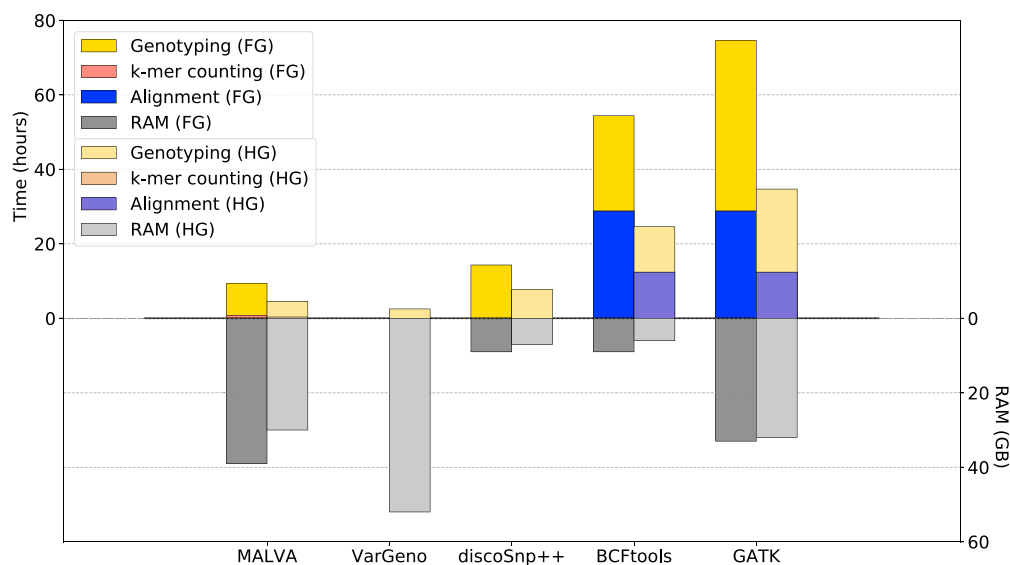
We note that VarGeno requires a different formatting of the fields describing the *a priori* frequencies of the alleles than the ones in the VCF file provided by the 1KGP. Thus, we formatted the input files as required before running VarGeno.

VarGeno could not complete the analysis of this dataset, from now on denoted by FullGenome, on our server. To test whether VarGeno crashed owing to excessive memory usage, we tried to run it on the same instance on a cluster with 1 TB of RAM, but nevertheless it could not complete the analysis, crashing after 20 min. To include VarGeno in our evaluation, we chose 12 chromosomes to create a smaller dataset, denoted by HalfGenome, that thus contains some half of the variants and the reads of the FullGenome dataset.

Each tool was evaluated in terms of variant calling accuracy and efficiency (wall time and memory usage). We note that some steps of the previously cited tools can use multiple threads to improve the time performance (namely, KMC3 for MALVA, discoSnp++, BWA-MEM for BCFtools and GATK, and the variant discovery steps of GATK). Whenever we had this choice, we provided four threads to each tool. We used hap.py (Krusche et al., 2019), the tool developed for the evaluation of variant callers in the recent PrecisionFDA Truth Challenge (<https://precision.fda.gov/challenges/truth>), and the `usr/bin/time` system tool to gather the required data.

Table 1 shows the results obtained by the considered tools on both the FullGenome and the HalfGenome datasets. We point out that hap.py computes precision and recall considering only non-reference VCF records (i.e., non 0/0 calls). A qualitative representation of these results is available in Figure S1 of the Supplemental Information.

As expected, MALVA, VarGeno, and discoSnp++ are faster than the tested alignment-based approaches, i.e., BCFtools and GATK. Indeed, MALVA, VarGeno, and discoSnp++ required 4.5, 2.5, and 7.5 h to analyze the HalfGenome dataset, respectively, whereas BCFtools and GATK required 24.5 and 34.5 h.



**Figure 1. Time and RAM Required by Each Tool to Analyze Both Datasets**

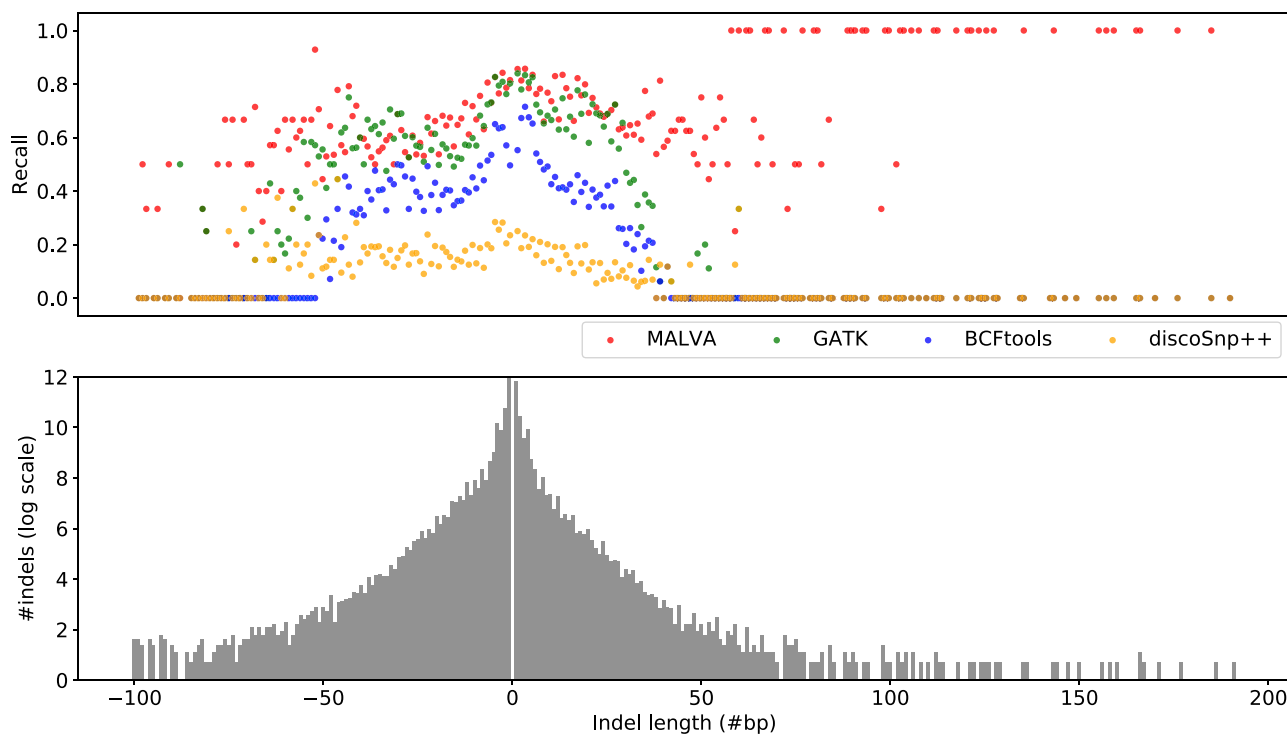
The running times are partitioned by steps performed, whereas the RAM usage represents the peak memory of the entire process. For ease of presentation, we denoted the FullGenome dataset as FG and the HalfGenome dataset as HG. Note that we did not include VarGeno running time and RAM usage on the FullGenome dataset since it crashed after 20 min.

We note that half of the time required by BCFTools and one-third of the time required by GATK was spent running BWA-MEM, which completed its task in 12.5 h (using four threads). The same trend applies to the analysis of the FullGenome dataset, on which each tool required roughly twice the time required on the HalfGenome dataset. A qualitative representation of the running time and the memory usage of each tool is shown in Figure 1.

For what concerns the memory usage, BCFTools proved to be the least memory intensive approach, requiring less than 10 GB of RAM on both datasets to map the reads and less than 1 GB of RAM to call the variants. MALVA and GATK showed similar memory requirements, with GATK showing almost no difference between the two analyses and MALVA increasing the memory consumption by only 23% for the bigger dataset. VarGeno required slightly less than twice the amount of memory required by MALVA on the HalfGenome dataset.

Precision and recall of all the tools varied little over the two datasets, proving that the number of variants and reads only slightly affects their accuracy. As expected, BCFTools and GATK achieved the best recall for non-homozygous reference SNPs owing to the mapping step, which provides a more precise coverage of the alleles and allows to better discern repeated regions of the reference genome. discoSnp++ achieved the lowest recall, whereas VarGeno obtained 3% less recall than MALVA, which in turn called correctly 91.1% of the SNPs. On the other hand, MALVA, discoSnp++, BCFTools, and GATK achieved comparable precision on SNPs, whereas VarGeno obtained the highest one. Overall, on non-homozygous reference SNPs, VarGeno seems to be the most conservative tool among those tested, as it prefers not to call SNPs when there is any uncertainty. On the contrary, MALVA, in avoiding the loss of any potentially interesting information, deliberately prefers to detect any potential alternate allele in the donor, at the cost of a slight loss in precision.

Remarkably, on indels MALVA obtained significantly better recall than BCFTools and discoSnp++ and better precision than any other tool. As expected, since the method of VarGeno is not designed to manage indels, it was only able to genotype a negligible percentage of them. On the other hand, discoSnp++ achieved a high precision but it was only able to call less than a quarter of the total indels. Finally, BCFTools showed a very low precision and recall on indels, whereas GATK achieved a recall similar to MALVA but a low precision. The low precision achieved by the alignment-based tools is mainly due to the difficulties in aligning reads that overlap with indels. We also performed a more detailed analysis on the



**Figure 2. Influence of Indel Size on the Recall Achieved by the Four Considered Tools on the FullGenome Dataset**

The histogram shows the frequency distribution (on logarithmic scale) of the indels with respect to their length. The scatterplot shows the recall of the tools with respect to the indel size.

influence of indel size on the recall obtained by the tools on the FullGenome dataset. As shown in Figure 2, MALVA proved to be the only tool able to call long indels (more than ~40/50 bases), whereas the other tools are limited to short indels (that are also the most common ones). In any case, MALVA outperformed the other tools even on these shorter indels. We did not include VarGeno in this analysis since its recall on indels was lower than 1% even on the HalfGenome dataset.

Overall, MALVA proved to be an accurate and efficient alternative to mapping-based pipelines for variant calling, achieving good results both on SNPs and indels. The experimental evaluation shows the usefulness of the formalization of signature of an allele, of the extension to multi-allelic SNPs and indels, and of the ability to manage variants in dense genomic regions. A more in-depth comparison of MALVA and VarGeno is provided in the next section.

### Comparison of MALVA and VarGeno Output

To assess whether the mapping-free approaches under analysis produce some systematic error, we considered the HalfGenome dataset and performed a more in-depth analysis of the SNPs genotyped by the two tools. For each tool, Figure 3 reports the number of correct genotypes output, grouping them in *homozygous reference* (i.e., 0|0), *heterozygous reference* (i.e., 0|1, 0|2, and so on), *homozygous alternate* (i.e., 1|1, 2|2, and so on), and *heterozygous alternate* (i.e., 1|2, 1|3, 2|3, and so on). As stated in the previous section, we recall that the precision and recall output by hap.py do not consider homozygous reference genotypes; thus, the analysis we present in this section allows us to better understand the behavior of the tools. Since VarGeno is not able to manage indels, we decided to not include them in this analysis.

Consistently with the precision and recall results of hap.py, MALVA detects between 5% and 10% more correct variants than VarGeno in all classes, at the cost of producing more erroneous calls. We note that overall VarGeno filters out 2,004,259 of the 39,796,878 SNPs in the truth.

Both tools show a similar pattern in erroneous calls. More precisely, erroneously genotyped homozygous reference variants were mostly genotyped as heterozygous reference and, vice versa, erroneously

		MALVA							VarGeno				
Real GT	HetAlt	54	98	165	987	0			0	0	0	0	1304
	HomoAlt	25405	24898	696563	117	0			2238	7811	632822	0	104112
	HetRef	97699	1056136	15588	132	0			78451	950511	12049	0	128544
	HomoRef	37809325	68037	1657	17	0			36087283	20507	947	0	1770299
		HomoRef	HetRef	HomoAlt	HetAlt	Uncalled			HomoRef	HetRef	HomoAlt	HetAlt	Uncalled
		Given GT							Given GT				

**Figure 3. Comparison between Real Genotype (Provided by the 1000 Genomes Project) and Genotype Called by MALVA and VarGeno**

HomoRef stands for Homozygous Reference, HetRef stands for Heterozygous Reference, HomoAlt stands for Homozygous Alternate, HetAlt stands for Heterozygous Alternate, and Uncalled means that the given variant was not called by the tool.

genotyped heterozygous reference variants were mostly genotyped as homozygous reference. On the other hand, erroneous homozygous alternate variants in the donor were mostly genotyped as heterozygous reference by `VarGeno`, whereas `MALVA` evenly distributed the errors between homozygous reference and heterozygous reference calls. Finally, erroneous heterozygous alternate variants in the donor were mostly genotyped as homozygous alternate variants by `MALVA`, meaning that the method proposed in this paper was able to detect the fact that the allele was not the reference allele but it called one of the two alternate alleles of the donor erroneously. [Figure 3](#) shows a comparison between real genotype (provided by the 1000 Genomes Project) and genotype called by `MALVA` and `VarGeno` on SNPs. [Figure 4](#) shows the same data row-normalized.

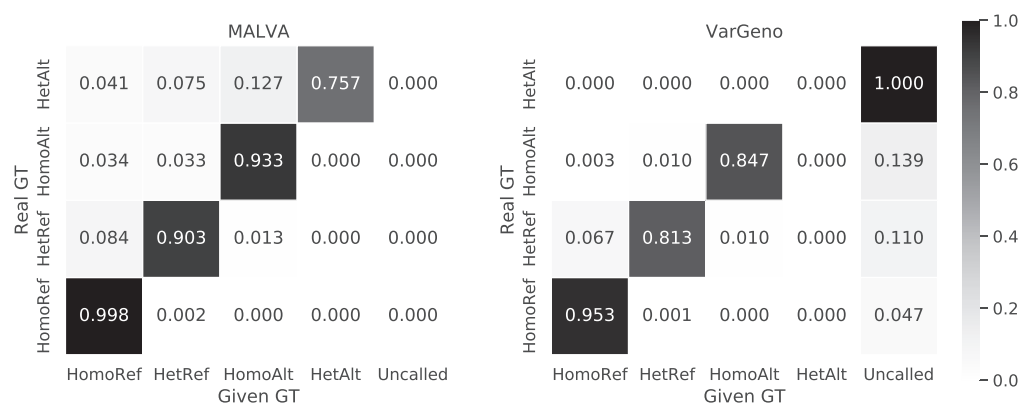
Overall, the errors produced by both tools were “partial” errors in the sense that they rarely mis-call both alleles of the donor.

## DISCUSSION

In this article, we presented `MALVA`, the first efficient mapping-free genotyping tool that is able to handle multi-allelic variants and indels. We compared `MALVA` with `VarGeno`, the state-of-the-art mapping-free genotyping tool, and showed that our method is less memory intensive, achieves better recall, handles dozens of millions of variants effectively, and provides correct genotypes even for indels. We also compared our tool with two variant discovery pipelines, namely, `GATK` and `BCFtools`, showing that `MALVA` is an order of magnitude faster while achieving better accuracy on indels and similar accuracy on SNPs.

`MALVA` proved to be able to efficiently manage a huge amount of variants like those provided by the 1000 Genome Project (about 80 million variants) and to handle multi-allelic variants and indels. These fundamental features allow our method to exploit the whole information in input, without filtering out any data that might be crucial in successive analyses. Most notably, `MALVA`'s ability to genotype indels allows one to apply mapping-free techniques to many clinical contexts, including screens for genetic predispositions for disease linked to the presence of indels ([Rimmer et al., 2014](#); [Warren et al., 1987](#)).

Future steps will be devoted to improving the efficiency of `MALVA` by exploiting the parallel architecture of modern machines and to extending the method to genotype trios. Another possible future direction consists in designing a mapping-free method for genotyping known variants using long reads such those produced by the latest PacBio and Oxford Nanopore technologies. Indeed, `MALVA` is specifically designed for dealing with Illumina short-read data and cannot be directly applied to long-read data owing to their higher error rate.



**Figure 4. Comparison between Real Genotype (Provided by the 1000 Genomes Project) and Genotype Called by MALVA and VarGeno, Normalized by Rows**

HomoRef stands for Homozygous Reference, HetRef stands for Heterozygous Reference, HomoAlt stands for Homozygous Alternate, HetAlt stands for Heterozygous Alternate, and Uncalled means that the given variant was not called by the tool.

### Limitations of the Study

MALVA achieves high accuracy and efficiency by analyzing only a subset of the sequences in input, i.e., the  $k$ -mers centered on the alleles. Although using the concept of signature proved to be effective in this context, in some edge cases two alleles might share the same signature and our method will not be able to discern between the two. A simple solution to reduce the occurrences of different alleles sharing the same signature is to increase the value of  $k$ . Unfortunately, increasing  $k$  beyond 40–50 has two main drawbacks: (1) it is computationally expensive and (2) owing to errors, the probability that such  $k$ -mers appear in the input reads decreases. To face this limitation, future works should (1) investigate the effect of using multiple  $k$ -mers spanning each allele and (2) exploit the  $k$ -mers flanking the potential occurrence of an allele in the read.

### METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

### SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2019.07.011>.

### ACKNOWLEDGMENTS

We acknowledge the support of the Fondo di Ateneo Quota Competitiva (FAQC) of the University of Milan-Bicocca.

### AUTHOR CONTRIBUTIONS

Conceptualization, M.P.; Methodology, G.B., L.D., and M.P.; Software, L.D. and M.P.; Validation, L.D.; Formal Analysis, G.B., L.D., and M.P.; Writing - Original Draft, G.B., L.D., and M.P.; Writing - Review & Editing, G.B., P.B., L.D., M.P., and A.S.; Visualization, L.D.; Supervision, P.B. and A.S.; Project Administration, P.B., M.P., and A.S.; Funding Acquisition, P.B.

### DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 11, 2019

Revised: June 5, 2019

Accepted: July 8, 2019

Published: August 30, 2019



## REFERENCES

- Computational Pan-Genomics Consortium (2016). Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* *19*, 118–135.
- DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* *43*, 491–498.
- Eggertsson, H.P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., Masson, G., Zink, F., Hjorleifsson, K.E., Jonasdottir, A., Jonasdottir, A., et al. (2017). GraphTyper enables population-scale genotyping using pangene graphs. *Nat. Genet.* *49*, 1654–1660.
- Hasan, M.S., Wu, X., and Zhang, L. (2015). Performance evaluation of indel calling tools using real short-read data. *Hum. Genomics* *9*, 20.
- Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de bruijn graphs. *Nat. Genet.* *44*, 226–232.
- Kaneo, T., Tahara, S., and Matsuo, M. (1996). Non-linear accumulation of 8-hydroxy-2'-deoxyguanosine, a marker of oxidized DNA damage, during aging. *Mutat. Res.* *316*, 277–285.
- Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., Francisco, M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* *37*, 555–560.
- Ku, C.S., Loy, E.Y., Salim, A., Pawitan, Y., and Chia, K.S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. *J. Hum. Genet.* *55*, 403–415.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* *9*, 357–359.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* *27*, 2987–2993.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* *25*, 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics* *25*, 2078–2079.
- MacArthur, D.G., and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* *19*, R125–R130.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* *20*, 1297–1303.
- Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res.* *16*, 1182–1190.
- Montgomery, S.B., Goode, D.L., Kvikstad, E., Albers, C.A., Zhang, Z.D., Mu, X.J., Ananda, G., Howie, B., Karczewski, K.J., Smith, K.S., et al. (2013). The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res.* *23*, 749–761.
- Mullaney, J.M., Mills, R.E., Pittard, W.S., and Devine, S.E. (2010). Small insertions and deletions (indels) in human genomes. *Hum. Mol. Genet.* *19*, R131–R136.
- Pajuste, F.D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M., and Remm, M. (2017). FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Sci. Rep.* *7*, 2537.
- Peterlongo, P., Riou, C., Drezen, E., and Lemaitre, C. (2017). DiscoSnp++: de novo detection of small variants from raw unassembled read set(s). *BioRxiv*. <https://doi.org/10.1101/209965>.
- Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R., WGS500 Consortium, Wilkie, A.O., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly-and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* *46*, 912–918.
- Sequist, L.V., Martins, R.G., Spigel, D., Grunberg, S.M., Spira, A., Janne, P.A., Joshi, V.A., McCollum, D., Evans, T.L., Muzikansky, A., et al. (2008). First-line gefitinib in patients with advanced non-small-cell lung cancer harboring somatic EGFR mutations. *J. Clin. Oncol.* *26*, 2442–2449.
- Shajii, A., Yorukoglu, D., William Yu, Y., and Berger, B. (2016). Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics* *32*, i538–i544.
- Sirén, J. (2017). Indexing variation graphs. In 2017 Proceedings of the nineteenth workshop on algorithm engineering and experiments (ALENEX), (SIAM), pp. 13–27.
- Standage, D.S., Brown, C.T., and Hormozdiari, F. (2019). Kevlar: a mapping-free framework for accurate discovery of de novo variants. *BioRxiv*. <https://doi.org/10.1101/549154>.
- Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* *526*, 75–81.
- Sun, C., and Medvedev, P. (2018). Toward fast and accurate SNP genotyping from whole genome sequencing data for bedside diagnostics. *Bioinformatics* *35*, 415–420.
- Warren, S.T., Zhang, F., Licameli, G.R., and Peters, J.F. (1987). The fragile x site in somatic cell hybrids: an approach for molecular cloning of fragile sites. *Science* *237*, 420–423.
- Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* *32*, 246–251.

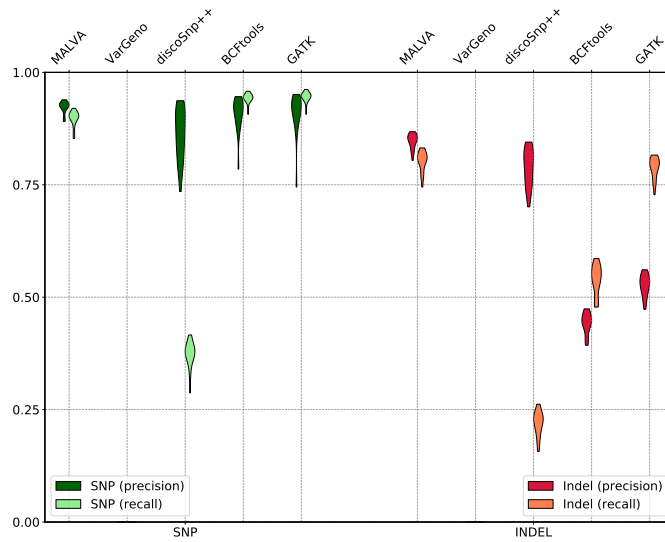
**ISCI, Volume 18**

**Supplemental Information**

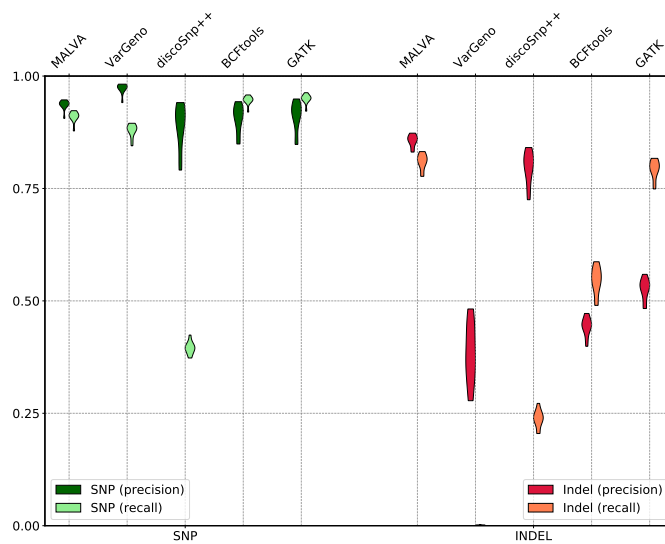
**MALVA: Genotyping by Mapping-free ALlele**

**Detection of Known VARIants**

**Luca Denti, Marco Previtali, Giulia Bernardini, Alexander Schönhuth, and Paola Bonizzoni**



(a) FullGenome dataset



(b) HalfGenome dataset

Figure S1: Qualitative representation of the accuracy results, Related to Table 1. Each violin plot represents the precision and the recall (computed with `hap.py`) achieved by the considered tools on both SNPs and indels. This is a qualitative representation of the information summarized by the table in the main document.

## S1. Transparent Methods

We will first introduce some preliminary definitions and then we will describe the approach we propose (MALVA) for the mapping-free genotyping of known variants.

### S1.1. Preliminaries

Let  $\Sigma$  be an ordered and finite alphabet of size  $\sigma$  and let  $t = c_1, \dots, c_k$ , where  $c_j \in \Sigma$  for  $j = 1, \dots, k$ , be an ordered sequence of  $k$  characters drawn from  $\Sigma$ , we say that  $t$  is a  $k$ -mer. When a  $k$ -mer originates from a double stranded DNA, it is common to consider it and its reverse-complemented sequence as the same  $k$ -mer, and to say that the one that is lexicographically smaller among the two is the *canonical* one. In the following, we will abide by this definition and whenever we refer to a  $k$ -mer we implicitly refer to its canonical form. Moreover, to avoid  $k$ -mers being equal to their reverse-complement, we will only consider odd values of  $k$ .

A Bloom filter (Bloom, 1970) is a probabilistic space-efficient data structure that represents a set of elements and allows approximate membership queries. The result of such queries may be a false positive but never a false negative. Bloom filters are usually represented as the union of a bitvector of length  $m$  and a set of  $h$  hash functions  $\{H_1, \dots, H_h\}$ , each one mapping one element of the universe to one integer in  $\{1, \dots, m\}$ . Using these data structures, the addition of an element  $e$  to the set is performed by setting to 1 the bitvector's cells in positions  $\{H_1(e), \dots, H_h(e)\}$ , while testing if an element is in the set boils down to checking whether the same positions are all set to 1. Due to collisions of the hash functions, an element can be reported as present in the set even if it is absent. Nevertheless, the false positive rate of a Bloom filter of a set of  $n$  elements, with  $h$  hash functions and an array of  $m$  bits is  $(1 - e^{-\frac{hn}{m}})^h$ ; therefore, to increase the size of the Bloom filter decreases the false positive rate. Due to their simplicity and efficiency, Bloom filters have been applied to multiple problems in bioinformatics, such as representing de Bruijn graphs (Chikhi and Rizk, 2013) and counting  $k$ -mers in a sample (Melsted and Pritchard, 2011).

Let  $B$  be a bitvector, the  $\mathbf{rank}_1$  function reports, for each position  $i \in \{1, \dots, |B| + 1\}$ , the number of 1s from the beginning of  $B$  to  $i$  (excluded); we refer to such value as  $\mathbf{rank}_1(i, B)$ . Clearly,  $\mathbf{rank}_1(i, B)$  is not defined for  $i \leq 0$  and for  $i > |B| + 1$ ,  $\mathbf{rank}_1(1, B)$  is 0, and  $\mathbf{rank}_1(|B| + 1, B)$  is the number of 1s in  $B$ . By using succinct support data structures and by a linear time preprocessing step, it is possible to answer  $\mathbf{rank}_1$  queries in constant time for any position of the bitvector (Vigna, 2008).

The difference between the genetic sequence of two unrelated individual of the same species is estimated to be smaller than 0.1% (Venter et al., 2001); therefore, it is common to represent the DNA sequence of an individual as a set of differences from a *reference* genome. Indeed, thorough studies (Consortium et al., 2015, 2003; consortium et al., 2015) of the variations across different individuals encode such information as a VCF (Variant Calling Format) file (Danecek et al., 2011). In the following, we will call *variant* the information encoded by a data line of a VCF file. Besides the genotype data, we are interested in the information carried by the second, fourth, fifth, and eighth field of a VCF line, namely: (i) field POS that is the position of the variant on the reference, (ii) field REF that is the reference allele starting in position POS, (iii) field ALT that is a list of alternate alleles that in some sample replace the reference allele, and (iv) field INFO that is a list of additional information describing the variant. From the latter list we will get the frequencies of reference and alternate alleles, which are needed to call the genotype of a given individual. We denote with  $\mathbf{POS}(v)$ ,  $\mathbf{REF}(v)$ ,  $\mathbf{ALT}(v)$ ,  $\mathbf{FREQ}(v)$ , and  $\mathbf{GTD}(v)$  the reference position, reference allele, list of alternate alleles, list of allele frequencies, and genotype data of a variant  $v$ , respectively. The variants we take into account are SNPs (*i.e.*, both REF and all the elements of ALT are single base nucleotides) and indels (REF and at least one element of ALT are not of the same length). Moreover, given an allele  $a$  (either reference or alternate) of some variant  $v$ , we refer to its sequence of nucleotides as  $\mathbf{SEQ}(a)$ , *i.e.*,  $\mathbf{SEQ}(a)$  is the string that represents  $a$ .

Let  $R$  be a reference genome and let  $V$  be a VCF file that describes all the known variants of  $R$ . Since the genotype data provides information on the

alleles expressed in each genome, another way of thinking of a VCF file is as an encoding of a set of genomes  $\mathcal{G}$ . Each haplotype of the genomes in  $\mathcal{G}$  can be reconstructed by modifying  $R$  according to the genotype information associated to each variant. For ease of presentation, in the following we use the term genome and haplotype interchangeably, although each genome of a polyploid organism is composed of multiple haplotypes.

Let  $\mathcal{G}$  be the set of genomes encoded by a VCF file and let  $a$  be an allele of some variant  $v$ , we denote by  $\mathcal{G}^a \subseteq \mathcal{G}$  the subset of genomes that include  $a$ . We say that a variant  $v$  is *k-isolated* if there is no other known variant within a radius of  $\lfloor k/2 \rfloor$  from the center of any of its alleles, as formally stated in the following definition.

**Definition 1 (*k-isolated variant*).** A variant  $v$  is *k-isolated* if, for all  $a \in \text{ALLELES}(v)$  and  $g \in \mathcal{G}^a$ , there is no variant  $v' \neq v$  with an allele  $a' \in \text{ALLELES}(v')$  such that  $g \in \mathcal{G}^{a'}$  and either  $|\text{BEGIN}_g(a') - \text{CENTER}_g(a)| \leq \lfloor k/2 \rfloor$  or  $|\text{CENTER}_g(a) - \text{END}_g(a')| \leq \lfloor k/2 \rfloor$ , where  $\text{ALLELES}(v) = \text{REF}(v) \cup \text{ALT}(v)$ ,  $\text{BEGIN}_g(a)$  is the position of the first base of  $a$  in  $g$ ,  $\text{END}_g(a)$  the position of the last base, and  $\text{CENTER}_g(a)$  the position of the  $\lceil \frac{|a|}{2} \rceil$ -th base of  $a$  in  $g$ .

The procedure we will present in the next section is heavily based on the concept of *signature* of an allele. Intuitively, a signature of the allele  $a$  of a variant  $v$  is the  $k$ -mer centered in  $a$  in some genome  $g$  in  $\mathcal{G}^a$ . Note that, depending on the genomes encoded by the VCF file (specifically, if variants less than  $k$  bases apart are known), an allele might have multiple signatures. Moreover, if  $\text{SEQ}(a)$  is longer than  $k$  bases, the previous definition is not well formed, since there is no  $k$ -mer that can be centered in  $a$ . In this case, we define the signature of  $a$  as the set of its substrings of length  $k$ . The following definition formalizes the notion of signature of an allele.

**Definition 2 (*Signature of an allele*).** Let  $\mathcal{G}$  be the set of all the genomes encoded by a VCF file  $V$  and let  $k$  be an odd positive value. Let  $v$  be a variant in  $V$ , let  $a$  be one of the alleles of  $v$ , and let  $\mathcal{G}^a \subseteq \mathcal{G}$  be the set of the genomes that include  $a$ . If  $\text{SEQ}(a)$  is longer than  $k$  bases, we say that the signature of  $a$

	1	2	3	4	5	6	7	8	9	10	11	12	13	
$\mathcal{R}$	A	G	A	T	C	C	T	G	C	G	A	A	G	

	Pos	Ref	Alts	Donors		
$v_1$	5	C	AAA	0 1	0 0	1 1
$v_2$	7	T	G	0 1	0 1	0 1
$v_3$	10	G	A,C	0 0	1 2	2 0

Variant	Allele	Signatures
$v_2$	$a_0 = T$	{ {TCCTGCG}, {TCCTGCA}, {AACTGCC} }
	$a_1 = G$	{ {AACGGCG}, {TCCGGCC} }

Figure S2: Signatures of the alleles of variant  $v_2$ .  $\mathcal{R}$  is the reference sequence and the table on the right is a VCF information associated to it, representing 3 variants: an indel ( $v_1$ ), a bi-allelic SNP ( $v_2$ ), and a multi-allelic SNP ( $v_3$ ). The last columns of the VCF file carry the genotype information of 3 individuals. The table at the bottom reports the signatures of each allele of variant  $v_2$ . Note that there are only 5 signatures although 6 haplotypes are encoded by the VCF file since the second haplotype of the first and third individual are the same. We highlighted in red the genotype information associated to the second haplotype of the second genome and the corresponding signature.

is the set of all the substrings of length  $k$  of  $\text{SEQ}(a)$ . If  $\text{SEQ}(a)$  is shorter than  $k$  bases, we say that  $\{x\text{SEQ}(a)y\}$  is the signature of  $a$  in a genome  $g$  in  $\mathcal{G}^a$  if: (i)  $x\text{SEQ}(a)y$  is a  $k$ -mer, (ii)  $|x| = \lfloor \frac{k-|\text{SEQ}(a)|}{2} \rfloor$ , (iii)  $|y| = \lceil \frac{k-|\text{SEQ}(a)|}{2} \rceil$ , (iv)  $x$  is a suffix of the sequence that precedes  $a$  in  $g$ , and (v)  $y$  is a prefix of the sequence that follows  $a$  in  $g$ .

We will refer to the set of all the possible signatures of an allele  $a$  as  $\text{SIGN}(a)$  and we say that  $k$  is the *length of the signature*. An example of signatures of an allele is shown in Figure S2. Notice that the same  $k$ -mer may appear in the signature of more than one allele.

In the following we will leverage on the definition of signature of an allele to detect its presence in an individual without mapping the reads to the reference genome. More precisely, we will analyze whether the  $k$ -mers of a given signature are present in the reads and use such information as an *hint* of the presence of the allele. Unlike other approaches (Pajuste et al., 2017), Definition 2 admits the presence of the alleles of multiple variants in a single signature, allowing MALVA to manage variants that are not  $k$ -isolated. Indeed, the set of signatures of an allele represents all the genomic regions where the allele appears in the genomes encoded by the VCF file.

### *S1.2. MALVA's approach*

In this section we will describe MALVA, the method we designed to genotype a set of known variants directly from a read sample. The general idea of MALVA is to use the frequencies of the signatures of a variant in the sample to call its genotype. The method works under the assumption that given a sample of reads from a genome with standard coverage depth, if an allele is included in the genome then at least one of its signatures must exist as substrings in multiple reads (depending on the coverage depth and the length of the signature). We leverage on this concept to genotype known variants directly from the input reads.

MALVA takes as input a reference genome, a VCF file representing all its known variants, and a read sample; it outputs a VCF file containing the most probable genotype for each variant. The main method is composed of four steps.

In the first step, MALVA computes the set of signatures of length  $k_s$  of all the alternate alleles of all the variants in VCF and stores them in the set **ALTSIG**. In the same step, the signatures of the reference alleles are computed and stored in a second set named **REFSIG**. For each  $k_s$ -mer  $t$  of a signature  $s$  two weights, one representing the number of occurrences of  $t$  in an alternate allele signature and one representing the number of occurrences of  $t$  in a reference allele signature, are stored. We will refer to these two values as  $w_t^A$  and  $w_t^R$ , respectively.

We note that for small values of  $k_s$  the probability that the  $k_s$ -mers that constitute a signature appear in other regions of the genome is high. Since in the following steps MALVA exploits the signatures' sets of the alleles of each variant to call the genotypes, the presence of conserved regions of the reference genome identical to some signature could lead the tool to erroneously genotype some variants. To get rid of a large amount of wrong calls, in the second step MALVA makes use of the context around the allele to distinguish its signatures from such regions. More precisely, if a  $k_s$ -mer of a signature of an alternate allele appears somewhere in the reference genome, MALVA extracts the context of length  $k_c$  (with  $k_c > k_s$ ) covering the reference genome region and collects such  $k_c$ -mers in a third set (**REPCTX**).



In the third step, MALVA extracts all the  $k_c$ -mers from the sample along with the number of its occurrences. For each  $k_c$ -mer  $t_c$  that occurs  $w$  times in the sample, the  $k_s$ -mer  $t_s$  that constitutes the center of  $t_c$  is extracted. If  $t_s$  is found in REFSIG,  $w_{t_s}^R$  is increased by  $w$ . Moreover, if  $t_c$  is not found in REPCTX and if  $t_s$  is in ALTSIG,  $w_{t_s}^A$  is increased by  $w$ . Otherwise, if  $t_c$  is in REPCTX,  $w_{t_s}^A$  is not updated since, although its central  $k_s$ -mer is identical to some  $k_s$ -mer of a signature of an alternate allele of some variant, it is indistinguishable from another region of the genome not covering the variant. We note that when  $w_{t_s}^A$  is not updated, our method might miss a variant in the donor and report a false negative, although for large values of  $k_c$  this would rarely occur. The rationale behind this choice is to avoid biases due to  $k_c$ -mers in conserved regions of the reference genome, preferring not to include an alternate allele in the output whenever ambiguities arise.

Finally, in the fourth step, MALVA uses the weights computed in the previous step to call the genotypes.

In the rest of this Section we will detail each one of the four steps of MALVA.

*Signature computation.* The first step consists of building the signatures of the alleles of all the variants and adding them either to ALTSIG, if they are the signatures of an alternate allele, or to REFSIG, if they are the signature of the reference allele. If a variant  $v$  is  $k_s$ -isolated, we build  $1 + |\text{ALT}(v)|$  signatures, one for each allele of  $v$ . Otherwise, there are some genomes in  $\mathcal{G}$  in which there is at least another allele of a variant that lays within a radius of  $\lfloor k_s/2 \rfloor$  nucleotides from the center of the allele of  $v$ . In practice, this means that we have to look at the genotype data of the variants within such radius: for each allele  $a$  of  $v$  we reconstruct the  $k_s$  bases long portions of the genomes in  $\mathcal{G}^a$  that constitute the signatures of  $a$ .

As pointed out in Definition 2, if  $|\text{SEQ}(a)| \geq k_s$ , the signature of  $a$  is the set of  $k_s$ -mers that appear in  $\text{SEQ}(a)$ . In this case we extract all such  $k_s$ -mers and add them either to REFSIG or ALTSIG. Otherwise, if  $|a| < k_s$ , we build the  $k_s$  bases long substrings of each genome in  $\mathcal{G}^a$  centered in  $a$  by scanning the VCF

file and reconstructing the sequences according to the genotype information it includes. More precisely, let  $a$  be an allele of a variant  $v$  and let  $V = \{v_1, \dots, v_n\}$  be the set of variants such that, for all  $1 \leq i \leq n$ : (i)  $v_i \neq v$ , (ii) there exists an allele  $a_j$  in  $\text{ALLELES}(v_i)$  such that  $a$  and  $a_j$  are both included in some genome  $g$ , and (iii) either ( $\text{END}(a_j) < \text{BEGIN}(a)$  and  $\text{CENTER}(a) - \lfloor k_s/2 \rfloor \leq \text{END}(a_j)$ ) or ( $\text{END}(a) < \text{BEGIN}(a_j)$  and  $\text{CENTER}(a) + \lfloor k_s/2 \rfloor \geq \text{BEGIN}(a_j)$ ) in  $g$ .

Given  $a$ , we use the genotype information stored in the VCF file to retrieve the haplotypes in which it is included, *i.e.*, a subset of the haplotypes in  $\mathcal{G}^a$ , and build the set  $V$ . Using  $V$  we gather all the alleles that precede and succeed  $a$  in the selected haplotypes and we use them, together with the reference sequence, to reconstruct *on the fly* the  $k_s$ -mer that covers  $a$ , by interposing reference substrings and allele sequences. Doing so, we don't need to reconstruct the whole haplotypes but we only analyze and reconstruct the required  $k_s$ -mers when needed.

Once all the  $k_s$ -mers have been constructed, they are added to **REFSIG** if  $a$  is the reference allele, to **ALTSIG** if it is an alternate allele.

*Detection of repeated signatures.* This step is aimed to detect and store in set **REPCTX** all the  $k_c$ -mers of the reference sequence whose central  $k_s$ -mer is included in some signature of some alternate allele,  $k_c > k_s$ . **REPCTX** will be used in a further step to discard alternate alleles that might be erroneously reported as expressed by **MALVA** only because they cannot be told apart from other identical regions of the reference sequence. To compute **REPCTX**, we extract all the  $k_c$ -mers of the reference sequence and test whether their central  $k_s$ -mer is in **ALTSIG**. If so, we add the  $k_c$ -mer to **REPCTX** to report that the  $k_s$ -mer is indistinguishable from some  $k_s$ -mer that is included in the signature of an alternate allele. The set **REPCTX** is then used in the next step as illustrated below. An example comprising the first two steps is shown in Section S1.3.

*Alleles' signatures weights computation.* In the third step, **MALVA** computes how many times the  $k_s$ -mers of each signature appear in the dataset. First, **MALVA** extracts all the  $k_c$ -mers of the read sample and tests their existence in **REPCTX** to

check whether their central  $k_s$ -mer cannot be told apart from some repetition in the reference genome. Then, given a  $k_c$ -mer  $t_c$  that occurs  $w$  times in the read sample, the  $k_s$ -mer  $t_s$  that constitutes its center is extracted. If  $t_s$  is found in REFSIG, *i.e.*,  $t_s$  is the signature of the reference allele of some variant, the weight  $w_{t_s}^R$  is increased by  $w$ . Moreover, if  $t_c$  is not found in REPCTX and  $t_s$  is in ALTSIG, *i.e.*,  $k_s$ -mer  $t_s$  is uniquely associated to an alternate allele of some variant, the weight  $w_{t_s}^A$  is increased by  $w$ . Conversely, if  $t_c$  is in REPCTX,  $w_{t_s}^A$  is not updated. The last scenario happens when  $t_s$  is identical to the signature of an alternate allele of some variant (indeed,  $t_s$  is in ALTSIG), but even the enlarged context  $t_c$  (and consequently  $t_s$ ) appears somewhere else in the reference genome.

*Genotype calling.* In the last step, MALVA uses the allele frequencies stored in the INFO field of the VCF file and the weights of the signatures computed in the previous step to call the genotype of each variant. To this aim, we extend the approaches proposed in the literature for bi-allelic variants (specifically, the one introduced in LAVA (Shajii et al., 2016)) to multi-allelic variants. While the approaches designed for genotyping bi-allelic variants only need to compute the likelihood of three genotypes, our technique must consider a larger number of possible genotypes.

Let  $v$  be a variant with  $n-1$  alternate alleles. The number of possible distinct genotypes is  $\binom{n}{2} + n = \frac{n(n+1)}{2}$ , that is one *homozygous reference* genotype,  $\binom{n}{2}$  *heterozygous* genotypes, and  $n-1$  *homozygous alternate* genotypes. We will refer to the homozygous reference genotype as  $G_{0,0}$ , to the heterozygous genotypes as  $G_{i,j}$  with  $0 \leq i < j \leq n-1$ , and to the homozygous alternate genotypes as  $G_{i,i}$  with  $1 \leq i \leq n-1$ . Following well-established techniques (Shajii et al., 2016; McKenna et al., 2010; Li, 2011), we compute the likelihood of each genotype  $G_{i,j}$  by means of the Bayes' theorem. Given the observed coverage  $C$ , we compute the posterior probability of each genotype as:

$$P(G_{i,j}|C) = \frac{P(G_{i,j})P(C|G_{i,j})}{P(C)}$$

that, by the law of total probability, can be expressed as:

$$P(G_{i,j}|C) = \frac{P(G_{i,j})P(C|G_{i,j})}{\sum_{p=0}^{n-1} \sum_{q=p}^{n-1} P(G_{p,q})P(C|G_{p,q})}$$

To calculate this probability, we compute the *a priori* probabilities of each genotype  $G_{i,j}$  ( $P(G_{i,j})$ ) and the *conditional probability* of the observed coverage given the considered genotype ( $P(C|G_{i,j})$ ). The Hardy-Weinberg equilibrium equation ensures that for each variant  $v$ ,  $(\sum_{i=0}^{n-1} f_i)^2 = 1$ , where  $f_i = \text{FREQ}(v)[i]$ , *i.e.*, the frequency of the  $i$ -th allele of  $v$ . We recall that  $\text{FREQ}(v)$  is stored in the INFO field of the VCF file. The *a priori* probability of each genotype  $G_{i,j}$  is therefore computed as follows:

$$P(G_{i,j}) = \begin{cases} f_i^2 & \text{if } i = j \\ 2f_i f_j & \text{otherwise} \end{cases}$$

To compute the conditional probability  $P(C|G_{i,j})$ , it is first necessary to compute the *coverages* of the alleles of the variant. Without loss of generality, let  $a_0$  be the first allele of the variant, *i.e.*,  $a_0$  is the reference allele with index 0. We recall that  $\text{SIGN}(a_0)$  is the set of signatures of allele  $a_0$  and that each signature is a set of one or more  $k$ -mers. We also recall that, in the previous step, for each  $k$ -mer  $t$  that belongs to some signature we computed two weights, namely  $w_t^R$  and  $w_t^A$ . Given a signature  $s \in \text{SIGN}(a_0)$ , we define its weight as the mean of the weights associated to the  $k$ -mers it contains, *i.e.*,  $\frac{\sum_{t \in s} w_t^R}{|s|}$  where  $|s|$  denotes the number of  $k$ -mers contained in signature  $s$ . Since the same allele may exhibit more signatures, we define the coverage  $c_0$  of allele  $a_0$  as the maximum value among the weights of its signatures, *i.e.*,  $\max\{\frac{\sum_{t \in s} w_t^R}{|s|} : s \in \text{SIGN}(a_0)\}$ . This formula can be easily modified to compute the coverage of an alternate allele ( $c_i$  for  $i \geq 1$ ) by switching  $w_t^R$  with  $w_t^A$ . The coverage  $c_i$  of an allele  $a_i$  of a variant is thus computed as follows:

$$c_i = \begin{cases} \max\{\frac{\sum_{t \in s} w_t^R}{|s|} : s \in \text{SIGN}(a_0)\} & \text{if } i = 0 \\ \max\{\frac{\sum_{t \in s} w_t^A}{|s|} : s \in \text{SIGN}(a_i)\} & \text{otherwise} \end{cases}$$

By extending the approach adopted in (Shajii et al., 2016), we consider each  $P(C|G_{i,j})$  to be multinomially distributed. Given a homozygous genotype  $G_{i,i}$ ,

we assume to observe the  $i$ -th allele, which is the correct one, with probability  $1 - \varepsilon$  (where  $\varepsilon$  is the expected error rate) whereas the other  $n - 1$  alleles (the erroneous ones) with probability  $\frac{\varepsilon}{n-1}$  each. Hence, we compute the conditional probability of an homozygous genotype as:

$$P(C|G_{i,i}) = \binom{c_i + C_E}{c_i} (1 - \varepsilon)^{c_i} \left(\frac{\varepsilon}{n-1}\right)^{C_E}$$

where  $C_E$  is the total sum of the coverages of the erroneous alleles, *i.e.*,  $C_E = \sum_{j \in \{0, \dots, n-1\} \setminus \{i\}} c_j$ . For what concerns heterozygous genotypes, we assume to observe the correct alleles, *i.e.*, the  $i$ -th and the  $j$ -th allele, with equal probability  $\frac{1-\varepsilon}{2}$  whereas the other  $n - 2$  erroneous alleles with probability  $\frac{\varepsilon}{n-2}$  each. We compute the conditional probability of an heterozygous genotype as follows:

$$P(C|G_{i,j}) = \binom{c_i + c_j + C_E}{c_i + c_j} \binom{c_i + c_j}{c_i} \left(\frac{1-\varepsilon}{2}\right)^{c_i} \left(\frac{1-\varepsilon}{2}\right)^{c_j} \left(\frac{\varepsilon}{n-2}\right)^{C_E}$$

where, again,  $C_E$  is the sum of the coverages of the erroneous alleles, *i.e.*,  $C_E = \sum_{p \in \{0, \dots, n-1\} \setminus \{i,j\}} c_p$ .

Finally, after computing the posterior probability of each genotype, MALVA outputs the genotype with the highest likelihood.

### *S1.3. Example of $k$ -mers weight computation*

In this section we present an example of computation of the weights associated with the signatures'  $k_s$ -mers. Figure S3 shows an example composed of three variants and two reads. In this example the values of  $k_s$  and  $k_c$  are set to 7 and 11, respectively. Subfigure (a) shows the 26-bases long reference sequence. Subfigure (b) reports on the left two bi-allelic variants ( $v_1$  and  $v_2$ ) and one multi-allelic variant ( $v_3$ ), and on the right the signatures of each allele of  $v_2$ . Subfigure (c) shows the elements of **ALTSIG** and **REFSIG** related to  $v_2$ . We note that the second signature in **ALTSIG** is composed of a single  $k_s$ -mer ( $t_s$ , equal to **TCCGGCG**) that appears in the reference genome, starting from position 17. Thus, the  $k_c$ -mer starting in position 15 and ending in position 25 ( $t_c$ , equal to **GATCCGGCGAA**) is added to **REPCTX**. Subfigure (d) shows two 11-bases long reads including  $t_s$ , extracted from position 3 and 15 of the donor. Clearly, only

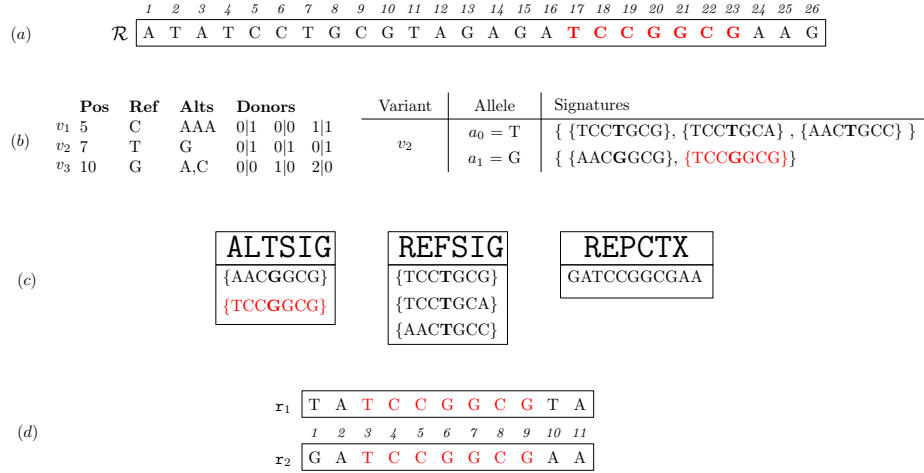


Figure S3: Example with 3 variants and two reads. Subfigure (a) shows a reference genome of 26 bases, Subfigure (b) reports 3 variants and the signatures of each allele of variant  $v_2$ , Subfigure (c) reports the subsets of **ALTSIG**, **REFSIG**, and **REPCTX** including the elements related to  $v_2$ , and Subfigure (d) presents two reads of length 11.

$r_1$  should contribute to the detection of the alternate allele of  $v_2$  in the donor, since  $r_2$  was sequenced from another position of the genome (*i.e.*,  $w_{t_s}^A$  should be equal to 1 in this case). To this aim, **REPCTX** comes to an aid; indeed, when analyzing  $r_1$  the  $k_c$ -mer covering  $t_s$  is extracted (*i.e.*, the whole read) and its inclusion in **REPCTX** is tested. Since **TATCCGGCGTA** is not in **REPCTX** and  $t_s$  is in **ALTSIG**,  $w_{t_s}^A$  is increased by one. On the other hand, since **GATCCGGCGAA** is in **REPCTX**, the occurrence of  $t_s$  in  $r_2$  is not considered in  $w_{t_s}^A$ , thus avoiding to erroneously overestimate the frequency of allele  $a_1$  of  $v_2$ .

We note that on one hand this approach allows us to avoid overestimating the frequencies of some alternate allele but, on the other hand, it produces two major side effects. The first one is that some allele might be underestimated by **MALVA**; indeed, if the  $k_c$ -mer covering an alternate allele in a donor is equal to a  $k_c$ -mer in the genome it will not be detected. The second side effect is that **MALVA** might overestimate the frequency of some allele due to identical signature. Indeed, suppose that the signature of some alternate allele  $a_i$  of another variant  $v_j \neq v_2$  is equal to the signature of alternate allele  $a_1$  of variant

$v_2$ . It is obvious that the weights of the  $k_s$ -mers of the two signatures will be identical and that the occurrences of both the alleles will concur towards their final value, overestimating it.

Although the two side effects pose some limit to the method proposed in this paper, they arise rarely and we think they are a fair price to pay to avoid biases introduced by the reference genome.

#### *S1.4. Implementation details.*

MALVA is implemented in C++ and it is freely available at <https://github.com/AlgoLab/malva>. Bloom filters were implemented as the union of a bitvector and a single hash function  $H$ . Although it is not conventional, in most cases to use a single hash function has similar results as using multiple ones, as noticed by other authors (Sun et al., 2017; Sun and Medvedev, 2018). To check this claim, while developing the tool we tested whether using multiple hash functions would improve the results by extending the Bloom filters to count-min sketches (Cormode and Muthukrishnan, 2005). As expected, the deterioration of the performance far outweighed the gain in precision and recall (that was less than 0.1%). Moreover, to use a single hash function allows us to store  $w_t^R$  and  $w_t^A$  efficiently for each  $k$ -mer  $t$ . Indeed, note that once all the signatures of all the alternate alleles have been added to ALTSIG, the latter is only used to check whether some  $k_s$ -mer is part of a signature, *i.e.*, it becomes static. By representing ALTSIG as a Bloom filter  $B_{\text{ALTSIG}}$  we can create an integer array CNTS of size  $\text{rank}_1(|B_{\text{ALTSIG}}| + 1, B_{\text{ALTSIG}})$  to store the weights of each  $k$ -mer compactly and, if a  $k$ -mer  $t$  of a signature  $s$  is in ALTSIG (*i.e.*, if  $B_{\text{ALTSIG}}[H(s)] = 1$ ) we can access its weight by accessing  $\text{CNTS}[\text{rank}_1(H(s), B_{\text{ALTSIG}})]$ . In a nutshell, after adding all the alternate alleles to  $B_{\text{ALTSIG}}$ , we *freeze* it, build a rank data structure over it, compute the number of ones, and create the CNTS array of the correct size. Similarly, we implemented REPCTX as a Bloom filter  $B_{\text{REPCTX}}$  using a single hash function. Conversely, REFSIG was implemented as a simple hash table, because the number of elements it stores is usually smaller than the number of elements stored in ALTSIG. The bitvectors, the rank data structure,

and the CNTS array were implemented using the `sds1-lite` library (Gog et al., 2014). We pose an upper limit of 255 to the value of each cell of the CNTS array, so as to store each counter using only 8 bits.

Finally, instead of scanning all the  $k_e$ -mers in the read sample, we used KMC3 (Dugosz et al., 2017) to efficiently extract them and counting their occurrences. Therefore, in step 3 MALVA parses the output of KMC3 and updates the counts for each  $k_s$ -mer accordingly.

### *S1.5. Data and Software Availability*

MALVA is freely available at <https://github.com/AlgoLab/malva>. Information and instruction on how to replicate the performed experiments are available at [https://github.com/AlgoLab/malva\\_experiments](https://github.com/AlgoLab/malva_experiments).

## **Supplemental References**

Bloom, B.H., 1970. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM* 13, 422–426.

Chikhi, R., Rizk, G., 2013. Space-efficient and exact de bruijn graph representation based on a bloom filter. *Algorithms for Molecular Biology* 8, 22.

Consortium, .G.P., et al., 2015. A global reference for human genetic variation. *Nature* 526, 68.

Consortium, I.H., et al., 2003. The international hapmap project. *Nature* 426, 789.

consortium, U., et al., 2015. The uk10k project identifies rare variants in health and disease. *Nature* 526, 82.

Cormode, G., Muthukrishnan, S., 2005. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms* 55, 58–75.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin,



- R., Group, .G.P.A., 2011. The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.
- Dugosz, M., Kokot, M., Deorowicz, S., 2017. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33, 2759–2761.
- Gog, S., Beller, T., Moffat, A., Petri, M., 2014. From theory to practice: Plug and play with succinct data structures, in: 13th International Symposium on Experimental Algorithms, (SEA 2014), pp. 326–337.
- Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernyt-sky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al., 2010. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research* 20, 1297–1303.
- Melsted, P., Pritchard, J.K., 2011. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* 12, 333.
- Pajuste, F.D., Kaplinski, L., Möls, M., Puurand, T., Lepamets, M., Remm, M., 2017. FastGT: an alignment-free method for calling common SNVs directly from raw sequencing reads. *Scientific Reports* 7, 2537.
- Shajii, A., Yorukoglu, D., William Yu, Y., Berger, B., 2016. Fast genotyping of known SNPs through approximate k-mer matching. *Bioinformatics* 32, i538–i544.
- Sun, C., Harris, R.S., Chikhi, R., Medvedev, P., 2017. Allsome sequence bloom trees, in: *Research in Computational Molecular Biology - 21st Annual International Conference, RECOMB 2017*, pp. 272–286.
- Sun, C., Medvedev, P., 2018. Toward fast and accurate SNP genotyping from whole genome sequencing data for bedside diagnostics. *Bioinformatics* 35, 415–420.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al., 2001. The sequence of the human genome. *science* 291, 1304–1351.

Vigna, S., 2008. Broadword implementation of rank/select queries, in: *Experimental Algorithms, 7th International Workshop, WEA 2008*, pp. 154–168.