FastGT: from raw sequence reads to 30 million genotypes in less than an hour

Fanny-Dhelia Pajuste[1*], Lauris Kaplinski[1*], Märt Möls[1,2], Tarmo Puurand[1], Maarja Lepamets[1] & Maido Remm[1]

*These authors contributed equally to this work.

[1]Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

[2]Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia

We have developed a computational method that counts the frequencies of unique *k*-mers in FASTQ-formatted genome data and uses this information to infer the genotypes of known variants. FastGT can detect the variants in a 30x genome in less than 1 hour using ordinary low-cost server hardware. The overall concordance with the genotypes of two Illumina "Platinum" genomes is 99.96%, and the concordance with the genotypes of the Illumina HumanOmniExpress is 99.82%. Our method provides *k*-mer database that can be used for the simultaneous genotyping of approximately 30 million single nucleotide variants (SNVs), including >23,000 SNVs from Y chromosome.

Next-generation sequencing (NGS) technologies are widely used for studying genome variation. Variants in the human genome are typically detected by mapping sequenced reads and then performing genotype calling[1–4]. A standard pipeline requires 40-50 h to process a human genome with 30x coverage from raw sequence data to variant calls on a multi-thread server. Mapping and calling are state-of-the-art processes that require expert users familiar with numerous available software options.It is not surprising that different pipelines generate slightly different genotype calls[5–9]. Fortunately, inconsistent genotype calls are associated with certain genomic regions only[10–12], whereas genotyping in the remaining 80-90% of the genome is robust and reliable.

The use of *k*-mers (substrings of length *k*) in genome analyses has increased because computers can handle large volumes of sequencing data more efficiently. For example, phylogenetic trees of all known bacteria can be easily built using *k*-mers from their genomic DNA[13–15]. Bacterial strains can be quickly identified from metagenomic data by searching for strain-specific *k*-mers[16–18]. *K*-mers have also been used to correct sequencing errors in raw reads[19–22]. One recent publication has described a method that calls variants from raw sequencing reads by using only a unique substring surrounding the variant[23].

We developed a new method that offers the possibility of directly genotyping known variants from NGS data by counting unique *k*-mers. The method only uses reliable regions of the genome and is approximately 1-2 orders of magnitude faster than traditional mapping-based genotype detection. Thus,

it is ideally suited for a fast preliminary analysis of a subset of markers before the full-scale analysis is finished.

The method is implemented in the C programming language and is available as the FastGT software package. FastGT is currently limited to the calling of previously known genomic variants because specific *k*-mers must be pre-selected for all known alleles. Therefore, it is not a substitute for traditional mapping and variant calling but a complementary method that facilitates certain aspects of NGS-based genome analyses. In fact, FastGT is comparable to a large digital microarray that uses NGS data as an input.Our method is based on three original components: 1) the procedure for the selection of unique *k*-mers, 2) the customized data structure for storing and counting *k*-mers directly from a FASTQ file, and 3) a maximum likelihood method designed specifically for estimating genotypes from *k*-mer counts.

RESULTS

Compilation of the database of unique *k*-mer pairs

The crucial component of FastGT is a pre-compiled flat-file database of genomic variants and corresponding *k*-mer pairs that overlap with each variant. Every bi-allelic single nucleotide variant (SNV) position in the genome is covered by *k* *k*-mer pairs. FastGT relies on the assumption that at least a number of these *k*-mer pairs are unique and appear exclusively in this location of the genome; therefore, the occurrence counts of these unique *k*-mer pairs in sequencing data can be used to identify the genotype of this variant in a specific individual (Figure S1).

The database of variants and unique *k*-mers is assembled by identifying all possible *k*-mer pairs for each genomic variant and subjecting them to several steps of filtering. The filtering steps remove variants for which unique *k*-mers are not observed and variants that produce non-canonical genotypes (non-diploid in autosomes and non-haploid in male X and Y chromosomes) in a sequenced test-set of individuals. A detailed description of the filtering steps used in this article is shown in Figure S2 and the Supplementary Data. Although one *k*-mer pair is theoretically sufficient for genotyping, mutations occasionally change the genome sequence in the neighborhood of an SNV, effectively preventing the detection of the SNV by a chosen *k*-mer. If the mutation is allele-specific, then the wrong genotype could be easily inferred. Therefore, we use three *k*-mer pairs per variant to prevent erroneous calls caused by the occasional loss of *k*-mers because of rare mutations (Figure S3). The number of pairs per variant is a compromise between the error rate and efficiency because using less than three k-mer pairs would increase the error rate, whereas using more *k*-mer pairs would consume more computer memory and prolong the genotyping.

In the current study, we compiled a database of all bi-allelic SNVs from dbSNP and tested the ability of FastGT to detect these SNVs with 25-mers. After the filtering steps, 30,238,283 (64%) validated and

bi-allelic SNVs remained usable by FastGT. We also used a subset of autosomal SNV markers present on the Illumina HumanOmniExpress microarray for a concordance analysis. In this set, 78% of the autosomal markers from this microarray were usable by FastGT. The number of SNV markers that passed each filtering step is shown in Table S1.

Algorithm and software for *k*-mer-based genotyping

The genotyping of individuals is executed by reading the raw sequencing reads and counting the frequencies of *k*-mer pairs described in the pre-compiled database of variants using the custom-made software `gmer_counter` and `gmer_caller` (Figure 1).

The database of genomic variants and corresponding *k*-mers is stored as a text file. The frequencies of *k*-mers listed in the database are counted by `gmer_counter`. It uses a binary data structure, which stores both *k*-mer sequences and their frequencies in computer memory during the counting process. A good compromise between memory consumption and lookup speed is achieved by combining a sorted table with a suffix tree. The sorted table is used for storing the sequence of the first 14 nucleotides, and the sparse bitwise suffix tree is used for storing the remaining sequence of the *k*-mers. Two bytes per *k*-mer are allocated for storing frequencies. The current implementation of `gmer_counter` accepts *k*-mers with lengths between 14 and 32 letters. The frequencies of up to three *k*-mer pairs from `gmer_counter` are saved in a text file that is passed to `gmer_caller`, which infers the genotypes based on *k*-mer frequencies and prints the results to a text file.

Empirical Bayes' method for inferring genotypes from k-mer counts

`Gmer_caller` uses the Empirical Bayes classifier for calling genotypes from *k*-mer frequency data, which assigns the most likely genotype to each variant. Allele frequency distributions are modeled by negative binomial distribution, described by seven parameters (see Supplementary Material). The model parameters are estimated separately for each analyzed individual using *k*-mer counts of 100,000 autosomal markers. The model allows us to estimate the most likely copy number for both alleles independently. Thus, in addition to calling bi-allelic (diploid) genotypes, we can also call mono-, tri-, or tetra-allelic genotypes, which might correspond to deletions and duplications of one allele (Figure 2). The model parameters can be saved and re-used in subsequent analyses of the same dataset.

The gender of the individual is determined automatically from the sequencing data using the median frequency of markers from the X chromosome (chrX). If the individual is female, only the autosomal model is used in the calling process and Y chromosome (chrY) markers are not called. For men, an additional haploid model of Bayes' classifier is trained for calling genotypes from sex chromosomes. Parameters for the haploid model are estimated using 100,000 markers from chrX.

120
121
122 **Assessment of genotype calling accuracy**
123
124 The accuracy of FastGT genotype calls was analyzed by comparing the results to genotypes reported in
125 two Illumina Platinum individuals, NA12877 and NA12878, which were sequenced to 50x coverage.
126 These are high-confidence variant calls derived by considering the inheritance constraints in the
127 pedigree and the concordance of variant calls across different methods
128 (http://www.illumina.com/platinumgenomes/).
129
130 The overall concordance of bi-allelic FastGT genotypes with genotypes from two Platinum genomes is
131 99.96%. The concordance of the non-reference (AB or BB) calls was 99.93%. The distribution of
132 differences between the two sets for different genotypes is shown in Table 1. All of the genotypes
133 reported in the Platinum datasets were bi-allelic; thus, we included only bi-allelic FastGT genotypes in
134 this comparison. The fraction of uncertain (no-call) genotypes in the FastGT output was 0.24%. The
135 uncertain genotypes are primarily mono-allelic (A) and tri-allelic (AAA) genotypes that might
136 correspond to deletions or insertions in a given region. However, non-canonical genotypes in the
137 default output are not reported, and they are replaced by NC ("no call"). All of the genotypes and/or
138 their likelihoods can be shown in `gmer_caller` optional output.
139
140 We also compared the genotypes obtained by the FastGT method with the data from the Illumina
141 HumanOmniExpress microarray. We used 504,173 autosomal markers that overlap our whole-genome
142 dataset (Table S1), and the comparison included ten individuals from the Estonian Genome Center for
143 whom both microarray data and Illumina NGS data were available.
144
145 In these 10 individuals, the concordance between the genotypes from the FastGT method and
146 microarray genotypes was 99.82% (Table 2), and the concordance of non-reference alleles was
147 99.69%. The fraction of mono-allelic and tri-allelic genotypes (no-call genotypes) in 10 test individuals
148 is rather low (<0.01% of all markers), indicating that our conservative filtering procedure is able to
149 remove most of the error-prone SNVs.
150
151
152 **Markers from Y chromosome**
153
154 FastGT is able to call genotypes from the Y chromosome (chrY) for 23,832 markers that remain in the
155 whole-genome dataset after all filtering steps. The genotypes on chrY cannot be directly compared with
156 the Platinum genotypes because chrY calls were not provided in the VCF file of the Platinum
157 individuals. To assess the performance of chrY genotyping, we compared our results to the genotypes
158 of 11 men from the HGDP panel[24] (http://cdna.eva.mpg.de/denisova/). The overall concordance of the

159 haploid genotype calls of FastGT and the genotype calls in these VCF files was 99.97%.The fraction of
160 non-canonical genotypes (no-calls) in the FastGT output was 1.27% (Table S2).
161
162 We also tested the concordance of chrY genotypes in seven father-son pairs in CEPH pedigree 1463
163 (http://www.ebi.ac.uk/ena/data/view/ERP001960). We assume that changes in chrY genotypes should
164 not occur within one generation. Only one marker (rs199503278) showed conflicting genotypes in any
165 of these father-son pairs. A visual inspection revealed problems with the reference genome assembly in
166 this region, which resulted in conflicting *k*-mer counts and conflicting genotypes from different *k*-mer
167 pairs of the same SNV. This marker was removed from the dataset because it had a high likelihood of
168 causing similar problems in other individuals.
169
170
171 **Effect of genome coverage on FastGT performance**
172
173 We also studied how the genome sequencing depth affects the performance of FastGT. The Platinum
174 genomes have a coverage depth of approximately 50x, but in most study scenarios, sequencing to a
175 lower coverage is preferred because it optimizes costs. For this analysis, we compiled different-sized
176 subsets of FASTQ sequences from the Platinum individual NA12878 and measured the concordance
177 between called genotypes and genotypes from the Platinum dataset. We observed that the concordance
178 rate of non-reference genotypes (AB and BB) declines significantly as the coverage drops below 20x
179 (Figure 3).
180
181
182 **Time and memory usage**
183
184 The entire process of detecting 30 million SNV genotypes from the sequencing data of a single
185 individual (30x coverage, 2 FASTQ files, 115GB each) takes approximately 40 minutes on a server
186 with 32 CPU cores. Most of this time is allocated to counting *k*-mer frequencies by `gmer_counter`.
187 The running time of `gmer_counter` is proportional to the size of the FASTQ files because the
188 speed-limiting step of `gmer_counter` is reading the sequence data from a FASTQ file. However, the
189 running time is also dependent on the number of FASTQ files (Figure 4) because simultaneously
190 reading from multiple files is faster than processing a single file. Genotype calling with
191 `gmer_caller` takes approximately 2-3minutes with 16 CPU cores.
192
193 The minimum amount of required RAM is determined by the size of the data structure stored in
194 memory by `gmer_counter`. We have tested `gmer_counter` on Linux computer with 8 GB of
195 RAM. However, server-grade hardware (multiple CPU cores and multiple fast hard drives in RAID) is
196 required to achieve the full speed of `gmer_counter` and `gmer_caller`.
197
198

METHODS

The methods used for compiling *k*-mer databases, statistical inference and testing the concordance of FastGT genotypes are described in the Supplementary Data.

Code availability

The binaries of FastGT package and *k*-mer databases described in the current paper are available on our website, http://bioinfo.ut.ee/FastGT/. The source code is available at GitHub (https://github.com/bioinfo-ut/GenomeTester4/). `Gmer_counter` and `gmer_caller` are distributed under the terms of GNU GPL v3, and the *k*-mer databases are distributed under the Creative Commons CC BY-NC-SA license.

DISCUSSION

FastGT is a flexible software package that performs rapid genotyping of a subset of previously known variants without a loss of accuracy. FastGT can be compared with a large digital microarray with millions of probes. One of the main strengths of FastGT is the selection of *k*-mers that are truly unique in the human genome. Because evaluating uniqueness in the reference genome alone is insufficient to identify *k*-mers that produce inconsistent results, we considered the use of short variants (SNVs and indels) from dbSNP databases and tested the uniqueness of the *k*-mers against all possible combinations of these variants. Additionally, we tested the expected behavior of the *k*-mers in a set of50 sequenced individual genomes. These procedures were used to compile a database of *k*-mers that directly yields reliable genotypes from sequencing data without the time-consuming mapping of reads. Our filtering procedure is rather conservative because we believe that the reliability of genotypes is more important than the sheer number of markers that can be genotyped.

The other advantage of FastGT is its efficient hybrid data structure for storing *k*-mer sequences and counts in binary format, which allows us to store data for only the *k*-mers of interest instead of for all *k*-mers from the data. This approach is particularly useful for genotyping only a small number of variants from each individual. An alternative method of *k*-mer-based genotyping can be based on full-genome *k*-mer lists. Numerous software packages can organize the raw sequencing data of each individual into comprehensive *k*-mer lists[25–29]. Using pre-compiled lists for each individual might have an advantage in certain situations, such as when the same lists are used repeatedly for different applications. However, the compilation of full-genome lists is somewhat inefficient if the lists are only used once and then immediately deleted. Most of the *k*-mers in the list will not be required for genotyping. For example, our database of 30 million SNVs contains 172 million *k*-mers, which is less than 5% of the *k*-mers present in the typical raw sequence data of an individual genome. Thus, if the lists are deleted immediately after use, it would be more reasonable to store the *k*-mer counts in RAM. Storing only the

relevant k-mers avoids the so-called "curse of deep sequencing," in which a higher coverage genome can overwhelm the memory or disk requirements of the software[30]. The disk and memory requirements of FastGT are not directly affected by the coverage or the amount of sequencing data.

Our analysis focuses on genotyping SNVs. However, FastGT is not limited to identifying SNVs. Any known variant that can be associated with a unique and variant-specific *k*-mer can be detected with FastGT. For example, short indels could be easily detected by using pairs of indel-specific *k*-mers. In principle, large indels, pseudogene insertions, polymorphic Alu-elements, and other structural variants could also be detected by *k*-mer pairs designed over the breakpoints. However, the detection of structural variants relies on the assumption that these variants are stable in the genome and have the same breakpoint sequences in all individuals, which is not always true for large structural variants. The applicability of FastGT for detecting structural variants requires further investigation and testing.

This software has only been used with Illumina sequencing data, which raises the question of whether our direct genotyping algorithm is usable with other sequencing technologies. In principle, *k*-mer counting should work with most sequencing platforms that produce contiguous sequences of at least *k* nucleotides. The uniformity of coverage and the fraction of sequencing errors in raw data are the main factors that influence *k*-mer counting because a higher error rate reduces the number of usable *k*-mers and introduces unwanted noise. The type of error is less relevant because both indel-type and substitution-type errors are equally deleterious for *k*-mer counting.

NGS data are usually stored in BAM format, and the original FASTQ files are not retained. In this case, the FASTQ file can be created from available BAM files, which can be performed by a number of software packages with multiple filtering choices(Picard, bam2fq from SAMtools package[1], bam2fastx from TopHat package[31]). We have tested FastGT software with raw FASTQ files and FASTQ files generated from the BAM-formatted files and did not observed significant differences in the *k*-mer counts or genotype calls. The sequencing strategies and techniques are diverse, and there is no single correct method of extracting the sequences. In principle, care should be taken to avoid multiple occurrences of the same reads in the resulting FASTQ file. Regardless of the method of genome analysis, contamination-free starting material, diligent sample preparation, and sufficient genome coverage are the ultimate pre-requisites for reliable results. The "garbage in, garbage out" principle applies similarly to mapping-based genome analyses and *k*-mer based genome analyses.

REFERENCES

1. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
2. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20,** 1297–303 (2010).
3. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).
4. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

5. Highnam, G. *et al*. An analytical framework for optimizing variant discovery from personal genomes. *Nat. Commun*. **6,** 6275 (2015).
6. Zook, J. M. *et al*. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol*. **32,** 246–251 (2014).
7. O'Rawe, J. *et al*. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*. **5,** 28 (2013).
8. Pirooznia, M. *et al*. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genomics* **8,** 14 (2014).
9. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30,** 2843–51 (2014).
10. Derrien, T. *et al*. Fast computation and applications of genome mappability. *PLoS One* **7,** (2012).
11. Lee, H. & Schatz, M. C. Genomic dark matter: The reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28,** 2097–2105 (2012).
12. Weisenfeld, N. I. *et al*. Comprehensive variation discovery in single human genomes. *Nat. Genet*. **46,** 1350–5 (2014).
13. Wen, J., Chan, R. H. F., Yau, S.-C., He, R. L. & Yau, S. S. T. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **546,** 25–34 (2014).
14. Ondov, B. D. *et al*. *Mash: fast genome and metagenome distance estimation using MinHash*. (2015). doi:10.1101/029827
15. Haubold, B., Klötzl, F. & Pfaffelhuber, P. andi: fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics* **31,** 1169–75 (2015).
16. Hasman, H. *et al*. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J. Clin. Microbiol*. **52,** 139–46 (2014).
17. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. **15,** R46 (2014).
18. Roosaare, M. *et al*. *StrainSeeker: fast identification of bacterial strains from unassembled sequencing reads using user-provided guide trees*. (2016). doi:10.1101/040261
19. Song, L., Florea, L. & Langmead, B. Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol*. **15,** 509 (2014).
20. Marçais, G., Yorke, J. A. & Zimin, A. QuorUM: An Error Corrector for Illumina Reads. *PLoS One* **10,** e0130821 (2015).
21. Lim, E.-C. *et al*. Trowel: a fast and accurate error correction module for Illumina sequencing reads. *Bioinformatics* **30,** 3264–5 (2014).
22. Zhao, X. *et al*. EDAR: an efficient error detection and removal algorithm for next generation sequencing data. *J. Comput. Biol*. **17,** 1549–60 (2010).
23. Kimura, K. & Koike, A. Ultrafast SNP analysis using the Burrows-Wheeler transform of short-read data. *Bioinformatics* **31,** 1577–83 (2015).
24. Meyer, M. *et al*. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–6 (2012).
25. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27,** 764–770 (2011).
26. Deorowicz, S., Kokot, M., Grabowski, S. & Debudaj-Grabysz, A. KMC 2: Fast and resource-frugal k-mer counting. *Bioinformatics* **31,** 1569–1576 (2014).
27. Rizk, G., Lavenier, D. & Chikhi, R. DSK: K-mer counting with very low memory usage. *Bioinformatics* **29,** 652–653 (2013).
28. Roy, R. S., Bhattacharya, D. & Schliep, A. Turtle: Identifying frequent k-mers with cache-efficient algorithms. *Bioinformatics* **30,** 1950–1957 (2014).
29. Kaplinski, L., Lepamets, M. & Remm, M. GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *Gigascience* **4,** 58 (2015).
30. Roberts, A. & Pachter, L. RNA-Seq and find: entering the RNA deep field. *Genome Med*. **3,** 74 (2011).
31. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

ACKNOWLEDGEMENTS

This work was funded by institutional grant IUT34-11 from the Estonian Ministry of Education and Research, grant SP1GVARENG from the University of Tartu, and the EU ERDF project (No. 2014-2020.4.01.15-0012, Estonian Center of Excellence in Genomics and Translational Medicine). The cost of the NGS sequencing of the individuals from the Estonian Genome Center was partly covered by the Broad Institute (MA, USA) and the PerMed I project from the TERVE program. The computational costs were partly covered by the High Performance Computing Centre at the University of Tartu. The authors thank Märt Roosaare, Ulvi Talas, and Priit Palta for performing a critical reading of the manuscript.

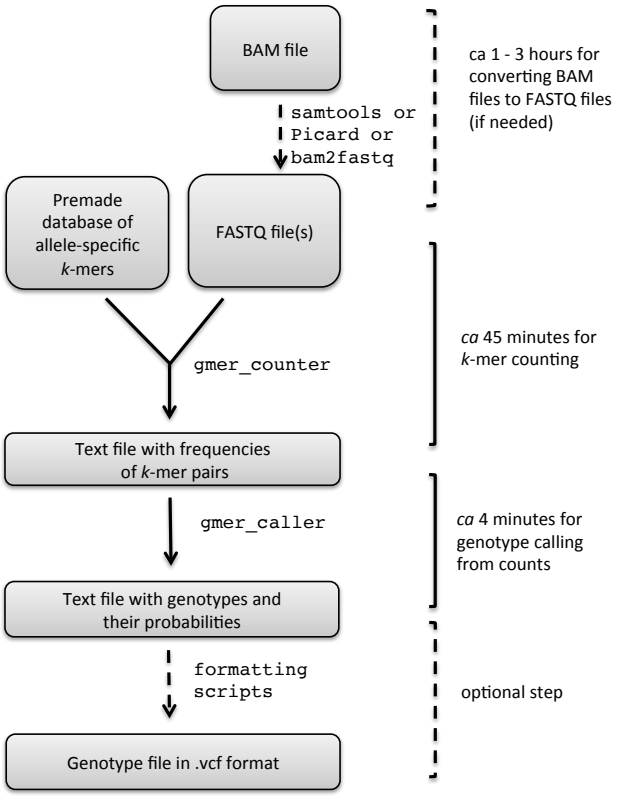
AUTHOR CONTRIBUTIONS

FDP compiled the databases of unique *k*-mers and conducted the genotype concordance analyses. LK invented the data structures and algorithms for `gmer_counter` and `gmer_caller` and implemented their code in C. MM wrote the initial code of the Bayesian classifier for genotype calling and supervised the development of a statistical framework. TP validated the genotyping results by performing a manual analysis of BAM files and providing expertise for NGS data management. ML performed an initial survey of the optimal number of *k*-mer pairs per variant. MR supervised the work and wrote the final version of the manuscript.

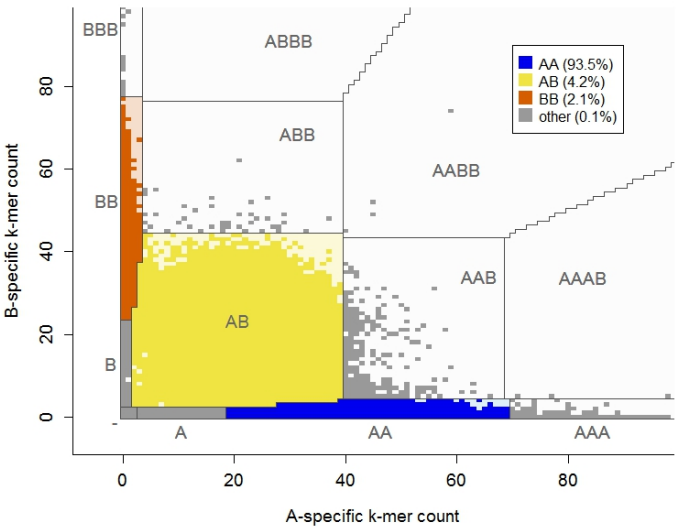
COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

364 **FIGURES**
365
366



367
368 **Figure 1.** Overall principle of *k*-mer-based genotyping.
369
370



371
372
373 **Figure 2**. Illustration of genotype calling based on the frequencies of two *k*-mers. The parameters that
374 define boundaries between genotypes are estimated from the *k*-mer frequency data of each individual.
375 By default, only conventional genotypes are reported in the output. "A" denotes the reference allele,
376 and "B" denotes an alternative allele. The median *k*-mer frequency of the individual used in this
377 example was 38.6.
378

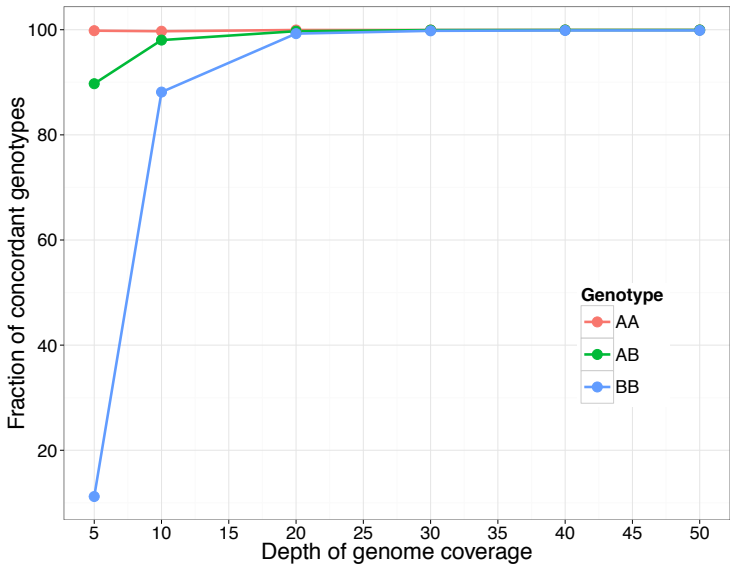


Figure 3. Effect of genome coverage on the concordance of genotypes. The accuracy of calling non-reference variants starts to decline as the genome coverage drops below 20x. Only the accuracy of the non-reference allele (genotypes AB and BB) calls declines significantly as the coverage drops because the higher prior probability of the reference allele has a stronger influence on the final decision of the Bayesian classifier in situations where the coverage is low (which increases the bias toward the more common allele).

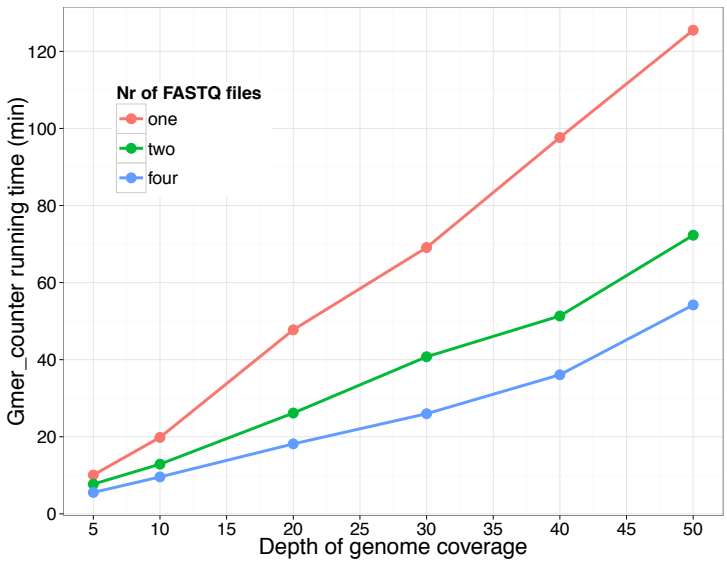


Figure 4. The time spent counting *k*-mer frequencies is proportional to the genome coverage (because of the larger FASTQ files). `Gmer_counter` is able to obtain data from multiple files simultaneously; thus, it runs faster if the sequence data are distributed between different files (e.g., files with paired reads).

394 **TABLES**
395
396 **Table 1**. Concordance between the autosomal genotypes of two individuals from the Platinum dataset
397 and bi-allelic FastGT genotypes called from the same individuals. "A" denotes the allele from the
398 reference genome, and "B" denotes the alternative allele.

		Platinum genotype calls		
		AA	AB	BB
FastGT genotype calls	AA	54,246,425 (93.39%)	987 (0.00%)	68 (0.00%)
	AB	20,041 (0.03%)	2,427,315 (4.18%)	1,516 (0.00%)
	BB	2,261 (0.00%)	156 (0.00%)	1,245,902 (2.14%)
	NC	126,513 (0.22%)	1,787 (0.00%)	10,376 (0.02%)
	concordant (%)	99.96%	99.95%	99.87%

399
400
401
402 **Table 2**. Distribution of all autosomal genotypes inferred by FastGT (rows) from the raw sequencing
403 data of 10 individuals from the Estonian Genome Center and the Illumina HumanOmniExpress
404 microarray genotypes (columns) from the same individuals. The depth of coverage of NGS data in
405 these individuals was between 21 and 35.
406

		Platinum genotype calls		
		AA	AB	BB
FastGT genotype calls	AA	2,750,130 (54.55%)	1,602 (0.03%)	1,204 (0.02%)
	AB	1,695 (0.03%)	1,477,508 (29.31%)	3,580 (0.07%)
	BB	2 (0.00%)	815 (0.02%)	804,828 (15.96%)
	NC	89 (0.00%)	253 (0.01%)	24 (0.00%)
	concordant (%)	99.94%	99.84%	99.41%

407
408

SUPPLEMENTARY METHODS

Compilation of database of unique *k*-mers

A *k*-mer length of 25 was used throughout this study, and the *k*-mers for genotyping were selected by the following filtering process (see also Figure S1). First, the validated single nucleotide variants (SNVs), as well as the validated and common indels, were extracted from the dbSNP database (build 146). Indels were used for testing the uniqueness of *k*-mers only; they are not included in the database of variants. For every bi-allelic SNV from this set, two sequences surrounding this SNV location were created: the sequence of the human reference genome (GRCh37) and the sequence variant corresponding to the alternative allele. The sequences were shortened to eliminate any possible overlap with neighboring SNVs or common indels. Essentially, this filtering step removed all of the SNVs that were located between two other SNVs (or indels) with less than 25bp between them. This step was chosen to avoid the additional complexity of filtering, counting, and calling algorithms because of the multiple combinations of neighboring SNV alleles. For all these SNVs that had variant-free sequences of at least 25bp, the sequences were divided into 25-mer pairs.

In the second filtering step, we tested the uniqueness of the 25-mers compiled in the previous step. The uniqueness parameter was tested against the "expanded reference genome," which is a set of 25-mers from the reference genome plus all possible alternative 25-mers containing the non-reference alleles of the SNVs and indels. A *k*-mer pair is considered unique if both *k*-mers occur no more than once in the "expanded reference genome". All non-unique *k*-mer pairs were removed from the list. The `Glistcompare` tool[29], which performs set operations with sorted *k*-mer lists, was used in this step. The *k*-mer pairs demonstrating uniqueness even with one mismatch were preferred. This constraint was added to reduce the risk of forming an identical *k*-mer by a rare point mutation or a sequencing error.

In the third step, the *k*-mers were further refined using the *k*-mer frequencies in a validated set of sequenced individual genomes. For this purpose, the *k*-mer counts were calculated for all SNVs of 50 random individuals whose DNA was collected and sequenced during the Center of Translational Genomics project at the University of Tartu. Twenty-five men and 25 women were used for filtering the autosomal SNVs; for chrX and chrY, 50 men were used. The sequencing depth in these individuals varied between 21 and 45. The *k*-mers showing abnormally high frequencies (greater than 3 times the median count in at least 2 out of 50 individuals) were removed from the database. In addition, the *k*-mer frequencies were counted for 50 women for chrY, and all the *k*-mers with a count greater than 3 were removed.

In the fourth stage, the remaining SNVs were filtered using the genotyping results from the same set of sequenced individual genomes. The genotypes for the remaining SNVs were calculated for 50 individuals (25 men + 25 women for markers from autosomes and chrX, 50 men for markers from chrY). The SNVs that produced a non-canonical allele count in more than one individual out of 50were removed from the dataset. The non-canonical allele count is any value other than two alleles in autosomes or a single allele in male chrX and chrY.

The final set contained 30,238,283 SNVs usable by FastGT, with 6.8% (2,063,839) located in protein-coding regions.

Statistical framework

The statistical framework for Empirical Bayes Classifier implemented in `gmer_caller` is described in Pajuste_2016_SupplementaryMaterial_S1.pdf

Testing genotype concordance

Version 20160503 of the FastGT package was used throughout this study. For the concordance analysis with the Platinum genotypes, `gmer_counter` and `gmer_caller` were run with the default options. The performance was tested on a Linux server with 32 CPU cores, 512GB RAM, and IBM 6Gbps and SAS 7200rpm disk drives in a RAID10 configuration.

High-quality genotypes were retrieved from the Illumina Platinum Genomes FTP site at ftp://ussd-ftp.illumina.com/hg38/2.0.1/.
BAM-format files of NA12877 and 12878 were downloaded from ftp://ftp.sra.ebi.ac.uk/vol1/ERA172/ERA172924/bam/NA12877_S1.bam and ftp://ftp.sra.ebi.ac.uk/vol1/ERA172/ERA172924/bam/NA12878_S1.bam.

FASTQ files were downloaded from the European Nucleotide Archive at http://www.ebi.ac.uk/ena/data/view/ERP001960. FASTQ files for the chrY genotype comparison were created from the corresponding BAM files using SAMtools bam2fq version 0.1.18. The read length of the Platinum genomes was 101 nucleotides.

Illumina HumanOmniExpress microarray genotypes and Illumina NGS data (read length 151 nt) for individuals V00278, V00328, V00352, V00369, V00402, V08949, V09325, V09348, V09365, and V09381 were obtained from the Estonian Genome Center. For the concordance analysis with the microarray genotypes, `gmer_caller` was run with the microarray markers (504,173) only.

The 5x, 10x 20x, 30x, and 40x data points for Figure3 were created using random subsets of reads from raw FASTQ files of 50x coverage from the Platinum individual NA12878.

SUPPLEMENTARY FIGURES

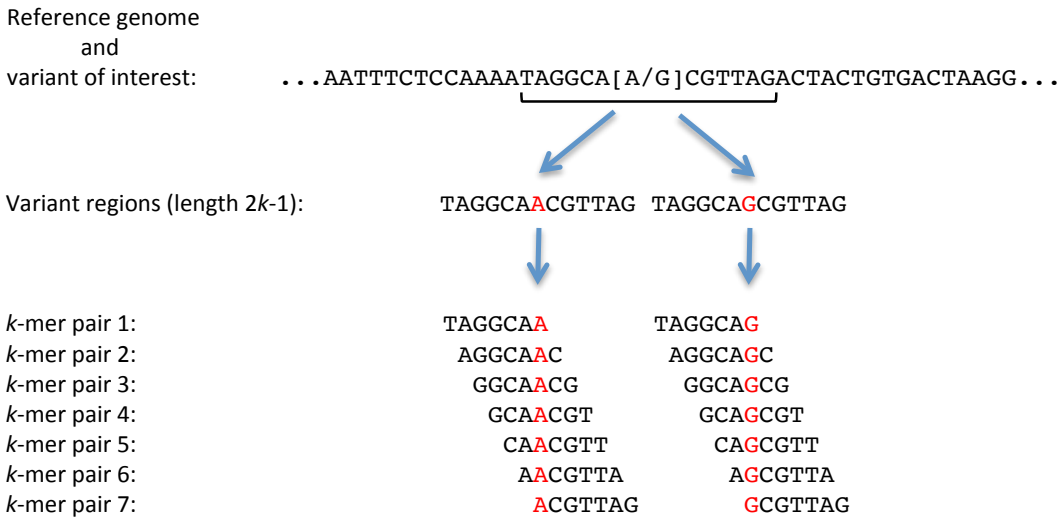


Figure S1. Simplified example of seven *k*-mer pairs (*k*=7) that can be used to distinguish two alleles of an SNV.

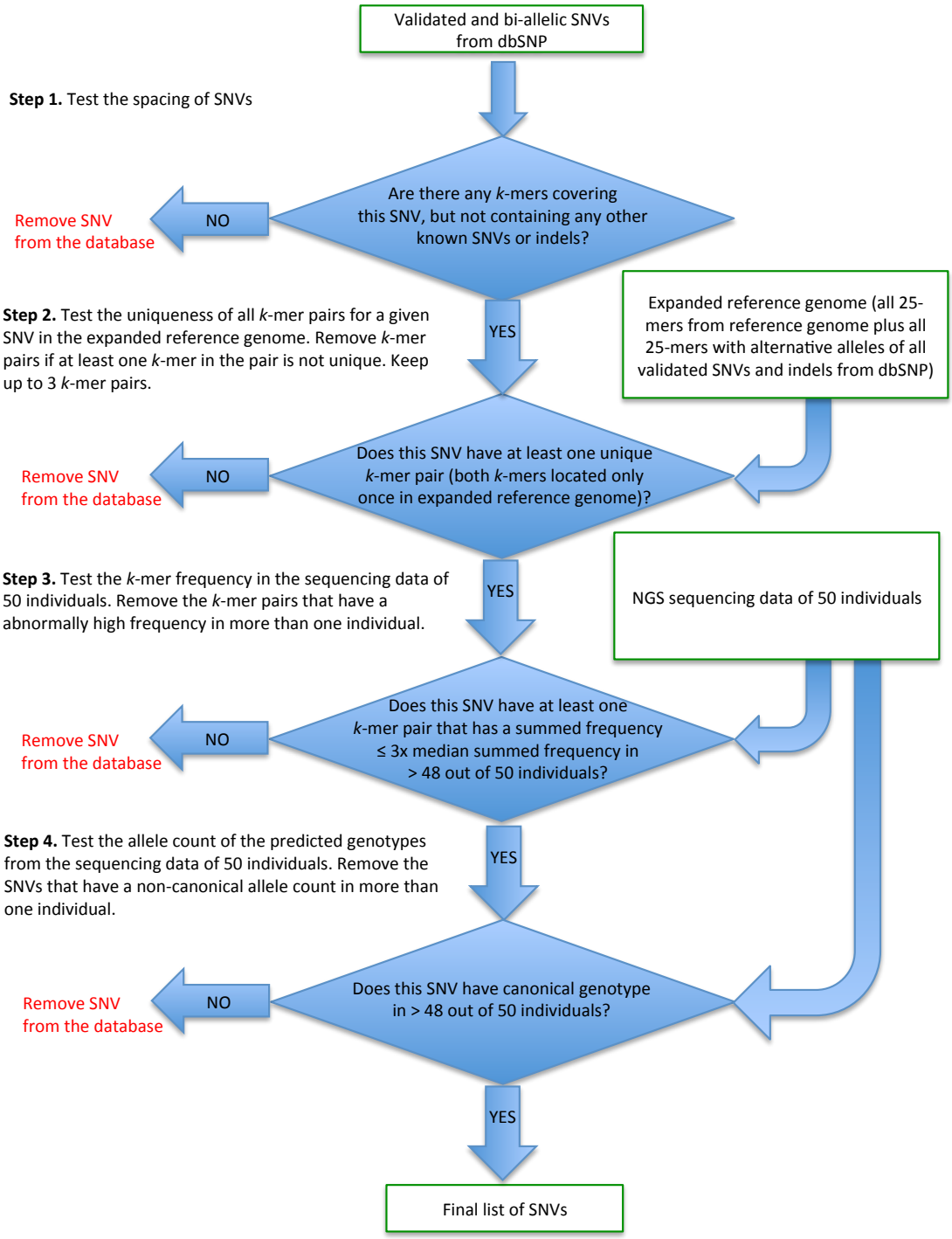


Figure S2. Pipeline for filtering markers.

```
Reference sequence:   ...TAGGCAACGTTAG...
k-mer pair 1:            TAGGCAA
                         TAGGCAG

k-mer pair 4:               GCAACGT
                            GCAGCGT

k-mer pair 7:                  ACGTTAG
                               GCGTTAG
```

For each SNV (shown in green), three *k*-mer pairs located as far away from each other as possible are selected.

```
Diploid genome sequenced to   ...TAGGCAACGTTAG...
coverage depth of ca 30x :    ...TAGTCAGCGTTAG...

k-mer pair 1:            TAGGCAA          15
                         TAGGCAG           0
k-mer pair 4:               GCAACGT       15
                            GCAGCGT        0
k-mer pair 7:                  ACGTTAG    15
                               GCGTTAG    15
```

For rare mutations in the neighborhood of the SNV (shown in red), certain *k*-mer pairs show abnormal frequencies. In this situation, at least one *k*-mer pair should still be usable.

```
Frequencies of k-mer pair 1:   15     0
Frequencies of k-mer pair 4:   15     0
Frequencies of k-mer pair 7:   15    15    AB
```

The *k*-mer pair with the total frequency closest to the median total frequency of all *k*-mer pairs in the entire genome is selected for genotype calling.

Figure S3. Principles of using redundant *k*-mer pairs for genotyping. *K*-mer pairs located as far away from each other as possible are selected. For example, in the case of *k*=7, as shown in this figure, we would prefer to use the 1st, 4th, and 7th *k*-mer pairs. For 25-mers, we prefer to use the 1st, 13th, and 25th *k*-mer pairs. If the most distant *k*-mer pair cannot be used (is not unique or contains SNVs), the next farthest *k*-mer pair is used. The third *k*-mer pair is chosen in the middle at an equal distance from both *k*-mers if possible. Thus, if a rare mutation at one side of the SNV changes the sequence on that side, we expect the *k*-mer pair from the other side to still have the expected counts. Although the frequencies for all three pairs are counted by `gmer_counter`, the genotype calling software `gmer_caller` uses only one pair, which is the pair with a total *k*-mer frequency count that is closest to the median *k*-mer frequency in a given individual.

509 **SUPPLEMENTARY TABLES**
510
511
512 **Table S1.** Number and fraction of usable SNVs remaining after subsequent filtering steps.

Dataset	All SNVs from dbSNP	Autosomal SNVs from HumanOmniExpress
Bi-allelic validated SNVs	46,954,719 (100%)	650,307 (100%)
After filtering step 1 (removal of closely located SNVs)	40,946,100 (87%)	596,806 (92%)
After filtering step 2 (removal of SNVs without unique k-mer pair)	34,463,965 (73%)	594,762 (91%)
After filtering step 3 (removal of k-mer pairs with high k-mer counts)	34,398,367 (73%)	594,736 (91%)
Final set after filtering step 4 (removal of SNVs with abnormal genotypes)	30,238,283 (64%)	504,173 (78%)

513
514
515
516

517 **Table S2**.Differences in all Y chromosome genotypes inferred by FastGT (rows) and the genotypes in
518 the VCF files of 11 men from the HGDPpanel[24].

		VCF genotype calls		
		AA	AB	BB
FastGT genotype calls	A	247,246 (94.42%)	3,797 (1.45%)	38 (0.01%)
	B	43 (0.02%)	148 (0.06%)	7,446 (2.84%)
	NC	3,026 (1.16%)	82 (0.03%)	41 (0.02%)
	concordant (%)	99.98%	0%	99.49%

519
520