

1 FastGT: an alignment-free method for calling common SNVs 2 directly from raw sequencing reads

3

4 Fanny-Dhelia Pajuste^{1*}, Lauris Kaplinski^{1*}, Märt Möls^{1,2}, Tarmo Puurand¹, Maarja Lepamets¹ & Mairo
5 Remm^{1,3}

6

7 *These authors contributed equally to this work.

8 ¹Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

9 ²Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia

10 ³Correspondence to maido.remm@ut.ee

11

12 **We have developed a computational method that counts the frequencies of unique k -mers in**
13 **FASTQ-formatted genome data and uses this information to infer the genotypes of known**
14 **variants. FastGT can detect the variants in a 30x genome in less than 1 hour using ordinary low-**
15 **cost server hardware. The overall concordance with the genotypes of two Illumina “Platinum”**
16 **genomes¹ is 99.96%, and the concordance with the genotypes of the Illumina**
17 **HumanOmniExpress is 99.82%. Our method provides k -mer database that can be used for the**
18 **simultaneous genotyping of approximately 30 million single nucleotide variants (SNVs),**
19 **including >23,000 SNVs from Y chromosome. The source code of FastGT software is available at**
20 **GitHub (<https://github.com/bioinfo-ut/GenomeTester4/>).**

21

22 Next-generation sequencing (NGS) technologies are widely used for studying genome variation.
23 Variants in the human genome are typically detected by mapping sequenced reads and then performing
24 genotype calling²⁻⁵. A standard pipeline requires 40-50 h to process a human genome with 30x
25 coverage from raw sequence data to variant calls on a multi-thread server. Mapping and calling are
26 state-of-the-art processes that require expert users familiar with numerous available software options. It
27 is not surprising that different pipelines generate slightly different genotype calls⁶⁻¹⁰. Fortunately,
28 inconsistent genotype calls are associated with certain genomic regions only¹¹⁻¹³, whereas genotyping
29 in the remaining 80-90% of the genome is robust and reliable.

30

31 The use of k -mers (substrings of length k) in genome analyses has increased because computers can
32 handle large volumes of sequencing data more efficiently. For example, phylogenetic trees of all
33 known bacteria can be easily built using k -mers from their genomic DNA¹⁴⁻¹⁶. Bacterial strains can be
34 quickly identified from metagenomic data by searching for strain-specific k -mers¹⁷⁻¹⁹. K -mers have also
35 been used to correct sequencing errors in raw reads²⁰⁻²³. One recent publication has described an
36 alignment-free SNV calling method that is based on counting the frequency of k -mers²⁴. This method
37 converts sequences from raw reads into Burrows-Wheeler transform and then calls genotypes by
38 counting using a variable-length unique substring surrounding the variant.

39

40 We developed a new method that offers the possibility of directly genotyping known variants from
41 NGS data by counting unique k -mers. The method only uses reliable regions of the genome and is
42 approximately 1-2 orders of magnitude faster than traditional mapping-based genotype detection. Thus,
43 it is ideally suited for a fast, preliminary analysis of a subset of markers before the full-scale analysis is
44 finished.

45

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

55

56

57 **RESULTS**

58

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

60

61 The crucial component of FastGT is a pre-compiled flat-file database of genomic variants and
62 corresponding k -mer pairs that overlap with each variant. Every bi-allelic single nucleotide variant
63 (SNV) position in the genome is covered by k k -mer pairs, where pair is formed by k -mers
64 corresponding to two alternative alleles (Figure S4). FastGT relies on the assumption that at least a
65 number of these k -mer pairs are unique and appear exclusively in this location of the genome;
66 therefore, the occurrence counts of these unique k -mer pairs in sequencing data can be used to identify
67 the genotype of this variant in a specific individual.

68

69 The database of variants and unique k -mers is assembled by identifying all possible k -mer pairs for
70 each genomic variant and subjecting them to several steps of filtering. The filtering steps remove
71 variants for which unique k -mers are not observed and variants that produce non-canonical genotypes
72 (non-diploid in autosomes and non-haploid in male X and Y chromosomes) in a sequenced test-set of
73 individuals. Filtering of k -mers was performed using high coverage NGS data of 50 individuals from
74 Estonian Genome Project (published elsewhere). The filtering steps are described in Methods section
75 and in Supplementary File (Figure S5).

76

77 Although one k -mer pair is theoretically sufficient for genotyping, mutations occasionally change the
78 genome sequence in the neighborhood of an SNV, effectively preventing the detection of the SNV by a

79 chosen k -mer. If the mutation is allele-specific, then the wrong genotype could be easily inferred.
80 Therefore, we use up to three k -mer pairs per variant to prevent erroneous calls caused by the
81 occasional loss of k -mers because of rare mutations.

82

83 In the current study, we compiled a database of all bi-allelic SNVs from dbSNP and tested the ability of
84 FastGT to detect these SNVs with 25-mers. After the filtering steps, 30,238,283 (64%) validated and
85 bi-allelic SNVs remained usable by FastGT. We also used a subset of autosomal SNV markers present
86 on the Illumina HumanOmniExpress microarray for a concordance analysis. In this set, 78% of the
87 autosomal markers from this microarray were usable by FastGT. The number of SNV markers that
88 passed each filtering step is shown in Table S1.

89

90

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

92

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

96

97 The database of genomic variants and corresponding k -mers is stored as a text file. The frequencies of
98 k -mers listed in the database are counted by `gmer_counter`. It uses a binary data structure, which
99 stores both k -mer sequences and their frequencies in computer memory during the counting process. A
100 good compromise between memory consumption and lookup speed is achieved by using adaptive radix
101 tree (see Supplementary File for detailed description of the data structure). The first 14 nucleotides of a
102 k -mer form an index into a table of sparse bitwise radix trees that are used for storing the remaining
103 sequence of the k -mers. Two bytes per k -mer are allocated for storing frequencies. The current
104 implementation of `gmer_counter` accepts k -mers with lengths between 14 and 32 letters. The
105 frequencies of up to three k -mer pairs from `gmer_counter` are saved in a text file that is passed to
106 `gmer_caller`, which infers the genotypes based on k -mer frequencies and prints the results to a text
107 file.

108

109

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

111

112 `Gmer_caller` uses the Empirical Bayes classifier for calling genotypes from k -mer frequency data,
113 which assigns the most likely genotype to each variant. Allele frequency distributions are modeled by
114 negative binomial distribution, described by eight parameters (see description in the Supplementary
115 File). The model parameters are estimated separately for each analyzed individual using k -mer counts
116 of 100,000 autosomal markers. The model allows us to estimate the most likely copy number for both
117 alleles. Given the observed allele counts, `gmer_caller` calculates the probability of genotypes by
118 applying the Bayes rule. As we do not require allele copy numbers to sum to 2 we can also call mono-,

119 tri-, or tetra-allelic genotypes (which might correspond to deletions and duplications) in addition to
120 traditional bi-allelic (diploid) genotypes (Figure 2). The model parameters can be saved and re-used in
121 subsequent analyses of the same dataset.

122

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

128

129

130 **Assessment of genotype calling accuracy through simulations**

131

132 In order to test the performance of FastGT, we generated simulated raw sequencing reads from the
133 reference genome and analyzed the ability of the Bayesian classifier of FastGT to reproduce genotypes
134 of the reference genome (see [Methods for detailed description of data simulation methods](#)). Throughout
135 this paper we denote A as reference allele and B as alternative allele. In this simulation, the reference
136 genome was assumed to be homozygous in all positions. Thus, the correct genotype for all 30,238,283
137 tested markers would be AA genotype. The fraction of AA genotypes recovered from simulated reads
138 varied between 98.94% (at 5x coverage) and 99.95% (at 20x coverage). The fraction of uncalled
139 markers was between 0.001% (at 20x coverage) and 1.036% (at 5x coverage). The fraction of AB calls
140 was in range of 0.02% to 0.05% at all coverages. The results are shown in Table 1.

141 The performance of calling AB and BB genotypes cannot be estimated from the reference genome. We
142 created simulated genomes using genotypes from 5 sequenced individuals, each from different
143 population (Yoruban, Chinese Han, CEPH, Puerto Rican and Estonian). This analysis helped to test the
144 performance of Bayesian classifier (`gmer_caller`) on calling the AB and BB variants from real-life
145 data. Secondly, this analysis indicates whether the selection of markers that was done using Estonian
146 individuals introduces any population-specific bias in genotype calling. The sensitivity (fraction of
147 correctly called AB and BB variants) was strongly affected by coverage (61% at 5x coverage, 99.8% at
148 20x coverage), but remained almost identical for individuals from different populations: 99.7 – 99.8%
149 at 30x coverage (Figure 3). The specificity (fraction of correctly called AA calls) was more uniform
150 over different coverage levels and remained between 99.60% to 99.95%. These results show that our
151 set of 30 million markers is usable for studying different populations without strong bias in sensitivity
152 or specificity.

153

154

155 **Assessment of genotype calling accuracy through concordance analysis**

156

157 The accuracy of FastGT genotype calls was analyzed by comparing the results to genotypes reported in
158 two Illumina Platinum individuals, NA12877 and NA12878, which were sequenced to 50x coverage.

159 These are high-confidence variant calls derived by considering the inheritance constraints in the
160 pedigree and the concordance of variant calls across different methods¹. We determined genotypes for
161 30,238,283 millions of markers from the FastGT database using raw sequencing data from the same
162 individuals and compared them to genotypes shown in the Platinum dataset.

163

164 The overall concordance of bi-allelic FastGT genotypes with genotypes from two Platinum genomes is
165 99.96%. The concordance of the non-reference (AB or BB) calls was 99.93%. The distribution of
166 differences between the two sets for different genotypes is shown in Table 2. All of the genotypes
167 reported in the Platinum datasets were bi-allelic; thus, we included only bi-allelic FastGT genotypes in
168 this comparison. The fraction of uncertain (no-call) genotypes in the FastGT output was 0.24%. The
169 uncertain genotypes are primarily mono-allelic (A) and tri-allelic (AAA) genotypes that might
170 correspond to deletions or insertions in a given region. However, non-canonical genotypes in the
171 default output are not reported, and they are replaced by NC (“no call”). All of the genotypes and/or
172 their likelihoods can be shown in `gmer_caller` optional output.

173

174 We also compared the genotypes obtained by the FastGT method with the data from the Illumina
175 HumanOmniExpress microarray. We used 504,173 autosomal markers that overlap our whole-genome
176 dataset (Table S2), and the comparison included ten individuals from the Estonian Genome Center for
177 whom both microarray data and Illumina NGS data were available. In these 10 individuals, the
178 concordance between the genotypes from the FastGT method and microarray genotypes was 99.82%
179 (Table 2), and the concordance of non-reference alleles was 99.69%. The fraction of mono-allelic and
180 tri-allelic genotypes (no-call genotypes) in 10 test individuals is rather low (<0.01% of all markers),
181 indicating that our conservative filtering procedure is able to remove most of the error-prone SNVs.

182

183

184 **Markers from Y chromosome**

185

186 FastGT is able to call genotypes from the Y chromosome (chrY) for 23,832 markers that remain in the
187 whole-genome dataset after all filtering steps. The genotypes on chrY cannot be directly compared with
188 the Platinum genotypes because chrY calls were not provided in the VCF file of the Platinum
189 individuals. To assess the performance of chrY genotyping, we compared our results to the genotypes
190 of 11 men from the HGDP panel²⁵ (<http://cdna.eva.mpg.de/denisova/>). The overall concordance of the
191 haploid genotype calls of FastGT and the genotype calls in these VCF files was 99.97%. The fraction
192 of non-canonical genotypes (no-calls) in the FastGT output was 1.27% (Table S3).

193

194 We also tested the concordance of chrY genotypes in seven father-son pairs in CEPH pedigree 1463
195 (<http://www.ebi.ac.uk/ena/data/view/ERP001960>). We assume that changes in chrY genotypes should
196 not occur within one generation. Only one marker (rs199503278) showed conflicting genotypes in any
197 of these father-son pairs. A visual inspection revealed problems with the reference genome assembly in
198 this region, which resulted in conflicting *k*-mer counts and conflicting genotypes from different *k*-mer

199 pairs of the same SNV. This marker was removed from the dataset because it had a high likelihood of
200 causing similar problems in other individuals.

201

202

203 **Effect of genome coverage on FastGT performance**

204

205 We also studied how the genome sequencing depth affects the performance of FastGT. The Platinum
206 genomes have a coverage depth of approximately 50x, but in most study scenarios, sequencing to a
207 lower coverage is preferred because it optimizes costs. For this analysis, we compiled different-sized
208 subsets of FASTQ sequences from the Platinum individual NA12878 and measured the concordance
209 between called genotypes and genotypes from the Platinum dataset. We observed that the concordance
210 rate of non-reference genotypes (AB and BB) declines significantly as the coverage drops below 20x
211 (Figure 4).

212

213

214 **Relationship between k -mer length and number of usable variants**

215

216 An obvious question is how the k -mer length affects the performance of FastGT. We used 25
217 nucleotides long k -mers throughout this article, but FastGT is able to use other k -mer lengths between
218 16 and 32 as well. We tested how many markers from dbSNP would remain usable for FastGT at
219 different values of k . From 47 millions validated markers 7-17% markers are removed in filtering step
220 1 due to closely located SNVs (Figure 5). In filtering step 2 markers are removed if they have no
221 unique k -mer pairs in the expanded reference genome. As expected, a rather large number of markers
222 are eliminated from the dataset if k -mers shorter than 20 nucleotides are used. However, the number of
223 usable markers does not increase significantly for k larger than 24. Thus, k values between 24 and 32
224 should be equally suitable for analyzing the human genome with FastGT. We have not compared the
225 accuracy of genotype calls of different k -mer lengths. However, we expect it to be relatively
226 independent of k -mer length. Two main factors that might influence the accuracy (concordance) of
227 genotypes are non-specific counts from shorter k -mers and drop of effective coverage of k -mers. The
228 effective coverage is negatively correlated with k -mer length. This negative correlation is caused by the
229 higher chance of accumulating sequencing errors within longer k -mers and by the end effects of the
230 reads (lower number of long k -mers per sequencing read). On the other hand, shorter k -mers are more
231 likely to pick up non-specific sequences due to sequencing errors and unknown variations in human
232 genomes. Overall, these effects influence the effective coverage of k -mers and are only critical if
233 genome coverage is low or if k -mer is shorter than 20 nucleotides. At high coverage (>20x) conditions
234 the k -mer length should not have significant influence to genotype accuracy.

235

236

237 **Time and memory usage**

238

239 The entire process of detecting 30 million SNV genotypes from the sequencing data of a single
240 individual (30x coverage, 2 FASTQ files, 115GB each) takes approximately 40 minutes on a server
241 with 32 CPU cores. Most of this time is allocated to counting *k*-mer frequencies by `gmer_counter`.
242 The running time of `gmer_counter` is proportional to the size of the FASTQ files because the
243 speed-limiting step of `gmer_counter` is reading the sequence data from a FASTQ file. However, the
244 running time is also dependent on the number of FASTQ files (Figure 6) because simultaneously
245 reading from multiple files is faster than processing a single file. Genotype calling with
246 `gmer_caller` takes approximately 2-3minutes with 16 CPU cores.

247
248 The minimum amount of required RAM is determined by the size of the data structure stored in
249 memory by `gmer_counter`. We have tested `gmer_counter` on Linux computer with 8 GB of
250 RAM. However, server-grade hardware (multiple CPU cores and multiple fast hard drives in RAID) is
251 required to achieve the full speed of `gmer_counter` and `gmer_caller`.

252

253

254 **METHODS**

255

256 **Compilation of database of unique *k*-mers**

257

258 A *k*-mer length of 25 was used throughout this study, and the *k*-mers for genotyping were selected by
259 the following filtering process (see also Figure S4). First, the validated single nucleotide variants
260 (SNVs), as well as the validated and common indels, were extracted from the dbSNP database build
261 146^{26,27}. Indels were used for testing the uniqueness of *k*-mers only; they are not included in the
262 database of variants. For every bi-allelic SNV from this set, two sequences surrounding this SNV
263 location were created: the sequence of the human reference genome (GRCh37) and the sequence
264 variant corresponding to the alternative allele. The sequences were shortened to eliminate any possible
265 overlap with neighboring SNVs or common indels. Essentially, this filtering step removed all of the
266 SNVs that were located between two other SNVs (or indels) with less than 25bp between them. This
267 step was chosen to avoid the additional complexity of counting and calling algorithms because of the
268 multiple combinations of neighboring SNV alleles. For all these SNVs that had variant-free sequences
269 of at least 25bp, the sequences were divided into 25-mer pairs.

270 In the second filtering step, we tested the uniqueness of the 25-mers compiled in the previous step. The
271 uniqueness parameter was tested against the “expanded reference genome,” which is a set of 25-mers
272 from the reference genome plus all possible alternative 25-mers containing the non-reference alleles of
273 the SNVs and indels. A *k*-mer pair is considered unique if both *k*-mers occur no more than once in the
274 “expanded reference genome”. All non-unique *k*-mer pairs were removed from the list. The
275 `Glistcompare` tool²⁸, which performs set operations with sorted *k*-mer lists, was used in this step.
276 The *k*-mer pairs demonstrating uniqueness even with one mismatch were preferred. This constraint was
277 added to reduce the risk of forming an identical *k*-mer by a rare point mutation or a sequencing error.

278 In the third step, the k -mers were further refined using the k -mer frequencies and genotypes in a set of
279 sequenced individual genomes. For this purpose, the k -mer counts and genotypes were calculated for
280 all SNVs of 50 random individuals whose DNA was collected and sequenced during the Center of
281 Translational Genomics project at the University of Tartu. Twenty-five men and 25 women were used
282 for filtering the autosomal SNVs; for chrX and chrY, 50 men were used. The sequencing depth in these
283 individuals varied between 21 and 45. Three different criteria were used for removing k -mer pairs and
284 SNVs in this step. First, we excluded all chrY markers that had k -mer frequency higher than 3 in more
285 than one woman. Second, autosomal k -mers showing abnormally high frequencies (greater than 3 times
286 the median count) in more than one individual were removed. Third criterion was based on unexpected
287 genotypes: the SNVs that produced a non-canonical allele count in more than one individual out of 50
288 were removed from the dataset. The non-canonical allele count is any value other than two alleles in
289 autosomes or a single allele in male chrX and chrY. The criteria used in filtering step 3 should remove
290 SNVs located in the regions that are unique in the reference genome, but frequently duplicated or
291 deleted in real individuals.

292 The final set contained 30,238,283 SNVs usable by FastGT, with 6.8% (2,063,839) located in protein-
293 coding regions. A detailed description of the filtering steps used in this article is shown in Figure S5.
294 The number of markers removed in each step is shown in Table S1.

295

296 **Statistical framework**

297

298 The statistical framework for Empirical Bayes Classifier implemented in `gmer_caller` is described
299 in Supplementary File.

300

301 **Generating and analyzing simulated data**

302

303 FastGT was tested on simulated reads. Simulated sequencing reads were generated using WgSim
304 (version 0.3.1-r13) software from samtools package³. The following parameters were used:
305 `base_error_rate=0.005`, `outer_distance_between_the_two_ends=500`, `standard_deviation=50`,
306 `length_of_the_first_read=100`, `length_of_the_second_read=100`. We used the base error rate 0.005
307 (0.5%) because this is similar to error rate typically observed in Illumina HiSeq sequencing data. We
308 estimated average error rate in the raw reads of high-coverage genomes from Estonian Genome Center
309 by counting the fraction of erroneous k -mers. The error rates in 100 individuals varied between 0.0030
310 and 0.0082, with average 0.0048 (CI95%=0.0002). Previous studies have reported similar overall error
311 rate in raw reads generated by Illumina HiSeq, varying between 0.002 and 0.004^{29,30}. The sequencing
312 reads were simulated with different coverages: about 5, 10, 20, 30 and 40. The number of read pairs
313 generated were 80 million, 160 million, 320 million, 480 million and 640 million respectively.

314 Reads were generated from standard reference genome, version GRCh37. For Figure 3 the reads were
315 also simulated using real SNV information for 5 individuals from 5 different populations (CEU, CHS,
316 YRI, PUR and EST). The following individuals were used in simulations: HG00512 (CHS), NA19238

317 (YRI), HG00731 (PUR), NA12877 (CEU) and V00055 (EST). Their sequencing data was retrieved
318 from 1000G project repository at
319 ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/hgsv_sv_discovery/data/ (CHS, YRI and
320 PUR), from the <ftp://ftp.sra.ebi.ac.uk/vol1/ERA172/ERA172924/bam/> (CEU) or from the Estonian
321 Genome Center. For each of these individuals, the standard reference genome was used as base and the
322 corresponding SNV genotypes from their VCF files were added to generate the reads with realistic
323 variants. The SNV genotypes were calculated from BAM or CRAM files using Genome Analysis
324 Toolkit version 3.6⁴.
325 The sensitivity and specificity were calculated for each individual and for each coverage. True positive
326 was defined as AB or BB genotype that was correctly called by FastGT in simulated data. True
327 negative values were defined as correctly called AA genotypes. The genotypes called from sex
328 chromosomes were not used for sensitivity and specificity calculations.

329

330 **Testing genotype concordance**

331

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

336

337 High-quality genotypes were retrieved from the Illumina Platinum Genomes FTP site at
338 <ftp://ussd-ftp.illumina.com/hg38/2.0.1/>.

339 BAM-format files of NA12877 and 12878 were downloaded from
340 ftp://ftp.sra.ebi.ac.uk/vol1/ERA172/ERA172924/bam/NA12877_S1.bam and
341 ftp://ftp.sra.ebi.ac.uk/vol1/ERA172/ERA172924/bam/NA12878_S1.bam.

342

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

347

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

352

353 The 5x, 10x 20x, 30x, and 40x data points for Figure 4 were created using random subsets of reads
354 from raw FASTQ files of 50x coverage from the Platinum individual NA12878.

355

356 **Code availability**

357

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

363

364

365 **DISCUSSION**

366

367 FastGT is a flexible software package that performs rapid genotyping of a subset of previously known
368 variants without a loss of accuracy. Another similar approach of genotype calling has been published
369 before²⁴. Both methods need to pre-process the reference genome and personal short-read data. Our
370 method pre-processes the genome by selecting the SNVs and compiling the database of k -mers that can
371 be used for calling these SNVs. The short-read data is pre-processed by counting and storing the k -mer
372 frequencies using `gmer_counter`. The method by Kimura and Koike uses dictionary-based approach
373 for storing both reference sequence and short reads. The dictionary is implemented by means of the
374 Burrows-Wheeler transform (BWT). The main advantage of BWT is the ability of storing and
375 comparing long strings efficiently. Therefore, this method can be used to call all SNVs, including those
376 that are in repeated genomic regions. FastGT uses fixed length k -mer with maximum length of 32. This
377 limits the number of variants that can be called from the human genome. On the other hand, using fixed
378 length k -mers allows faster processing of data due to 64-bit architecture of computer hardware. Thus,
379 FastGT essentially sacrifices calling some SNVs (up to 36%) from difficult genomic regions to
380 minimize data processing time. Another difference between FastGT and the method used by Kimura
381 and Koike is handling of the *de novo* mutations. Kimura and Koike implemented two methods (drop-
382 scan and step-scan) to detect *de novo* variants based on k -mer coverage and/or by local alignment of
383 surrounding region. FastGT has currently no ability to call *de novo* variants and is limited to calling
384 sub-sets of pre-defined variants. Thus, FastGT functions in principle as a large digital microarray with
385 millions of probes.

386

387 Numerous software packages can organize the raw sequencing data of each individual into
388 comprehensive k -mer lists^{28,31-34}, which can be later used for fast retrieval of k -mer counts. However,
389 the compilation of full-genome lists is somewhat inefficient if the lists are only used once and then
390 immediately deleted. FastGT uses adaptive radix tree, which allows us to store frequencies for only the
391 k -mers of interest, instead of for all k -mers from the genome. This approach is particularly useful for
392 genotyping only a small number of variants from each individual. Storing only the frequencies of
393 relevant k -mers avoids the so-called “curse of deep sequencing,” in which a higher coverage genome
394 can overwhelm the memory or disk requirements of the software³⁵. The disk and memory requirements
395 of FastGT are not directly affected by the coverage of sequencing data.

396

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

405

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

413

414 NGS data are usually stored in BAM format, and the original FASTQ files are not retained. In this
415 case, the FASTQ file can be created from available BAM files. This can be performed by a number of
416 software packages (Picard, bam2fq from SAMtools package¹, bam2fastx from TopHat package³⁶). We
417 have tested FastGT software with raw FASTQ files and FASTQ files generated from the BAM-
418 formatted files and did not observe significant differences in the k -mer counts or genotype calls. In
419 principle, care should be taken to avoid multiple occurrences of the same reads in the resulting FASTQ
420 file. Regardless of the method of genome analysis, contamination-free starting material, diligent sample
421 preparation, and sufficient genome coverage are the ultimate pre-requisites for reliable results. The
422 “garbage in, garbage out” principle applies similarly to mapping-based genome analyses and k -mer
423 based genome analyses.

424

425

426 REFERENCES

427

- 428 1. Eberle, M. A. *et al.* A reference dataset of 5.4 million phased human variants validated by
429 genetic inheritance from sequencing a three-generation 17-member pedigree. *bioRxiv* (2016).
- 430 2. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
431 *Bioinformatics* **26**, 589–95 (2010).
- 432 3. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
433 2079 (2009).
- 434 4. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing
435 next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
- 436 5. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
437 357–9 (2012).
- 438 6. Highnam, G. *et al.* An analytical framework for optimizing variant discovery from personal
439 genomes. *Nat. Commun.* **6**, 6275 (2015).
- 440 7. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP
441 and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
- 442 8. O’Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications

- 443 for exome and genome sequencing. *Genome Med.* **5**, 28 (2013).
- 444 9. Pirooznia, M. *et al.* Validation and assessment of variant calling pipelines for next-generation
445 sequencing. *Hum. Genomics* **8**, 14 (2014).
- 446 10. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples.
447 *Bioinformatics* **30**, 2843–51 (2014).
- 448 11. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**,
449 (2012).
- 450 12. Lee, H. & Schatz, M. C. Genomic dark matter: The reliability of short read mapping illustrated
451 by the genome mappability score. *Bioinformatics* **28**, 2097–2105 (2012).
- 452 13. Weisenfeld, N. I. *et al.* Comprehensive variation discovery in single human genomes. *Nat.*
453 *Genet.* **46**, 1350–5 (2014).
- 454 14. Wen, J., Chan, R. H. F., Yau, S.-C., He, R. L. & Yau, S. S. T. K-mer natural vector and its
455 application to the phylogenetic analysis of genetic sequences. *Gene* **546**, 25–34 (2014).
- 456 15. Ondov, B. D. *et al.* *Mash: fast genome and metagenome distance estimation using MinHash.*
457 (2015). doi:10.1101/029827
- 458 16. Haubold, B., Klötzl, F. & Pfaffelhuber, P. andi: fast and accurate estimation of evolutionary
459 distances between closely related genomes. *Bioinformatics* **31**, 1169–75 (2015).
- 460 17. Hasman, H. *et al.* Rapid whole-genome sequencing for detection and characterization of
461 microorganisms directly from clinical samples. *J. Clin. Microbiol.* **52**, 139–46 (2014).
- 462 18. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using
463 exact alignments. *Genome Biol.* **15**, R46 (2014).
- 464 19. Roosaare, M. *et al.* *StrainSeeker: fast identification of bacterial strains from unassembled*
465 *sequencing reads using user-provided guide trees.* (2016). doi:10.1101/040261
- 466 20. Song, L., Florea, L. & Langmead, B. Lighter: fast and memory-efficient sequencing error
467 correction without counting. *Genome Biol.* **15**, 509 (2014).
- 468 21. Marçais, G., Yorke, J. A. & Zimin, A. QuorUM: An Error Corrector for Illumina Reads. *PLoS*
469 *One* **10**, e0130821 (2015).
- 470 22. Lim, E.-C. *et al.* Trowel: a fast and accurate error correction module for Illumina sequencing
471 reads. *Bioinformatics* **30**, 3264–5 (2014).
- 472 23. Zhao, X. *et al.* EDAR: an efficient error detection and removal algorithm for next generation
473 sequencing data. *J. Comput. Biol.* **17**, 1549–60 (2010).
- 474 24. Kimura, K. & Koike, A. Ultrafast SNP analysis using the Burrows-Wheeler transform of short-
475 read data. *Bioinformatics* **31**, 1577–83 (2015).
- 476 25. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual.
477 *Science* **338**, 222–6 (2012).
- 478 26. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology
479 Information. *Nucleic Acids Res.* **44**, D7–19 (2016).
- 480 27. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms
481 and other classes of minor genetic variation. *Genome Res.* **9**, 677–9 (1999).
- 482 28. Kaplinski, L., Lepamets, M. & Remm, M. GenomeTester4: a toolkit for performing basic set
483 operations - union, intersection and complement on k-mer lists. *Gigascience* **4**, 58 (2015).
- 484 29. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51
485 (2013).
- 486 30. Schirmer, M., D’Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: resolving
487 fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**, 125 (2016).
- 488 31. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
489 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
- 490 32. Deorowicz, S., Kokot, M., Grabowski, S. & Debudaj-Grabysz, A. KMC 2: Fast and resource-
491 frugal k-mer counting. *Bioinformatics* **31**, 1569–1576 (2014).
- 492 33. Rizk, G., Lavenier, D. & Chikhi, R. DSK: K-mer counting with very low memory usage.
493 *Bioinformatics* **29**, 652–653 (2013).
- 494 34. Roy, R. S., Bhattacharya, D. & Schliep, A. Turtle: Identifying frequent k-mers with cache-
495 efficient algorithms. *Bioinformatics* **30**, 1950–1957 (2014).
- 496 35. Roberts, A. & Pachter, L. RNA-Seq and find: entering the RNA deep field. *Genome Med.* **3**, 74
497 (2011).
- 498 36. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-
499 Seq. *Bioinformatics* **25**, 1105–1111 (2009).
- 500
501

502 **ACKNOWLEDGEMENTS**
503

504 This work was funded by institutional grant IUT34-11 from the Estonian Ministry of Education and
505 Research, grant SPIGVARENG from the University of Tartu, and the EU ERDF grant No. 2014-
506 2020.4.01.15-0012 (Estonian Center of Excellence in Genomics and Translational Medicine). The cost
507 of the NGS sequencing of the individuals from the Estonian Genome Center was partly covered by the
508 Broad Institute (MA, USA) and the PerMed I project from the TERVE program. The genome data was
509 collected and used with ethical approval Nr. 206T4 (obtained for the project SPIGVARENG). The
510 computational costs were partly covered by the High Performance Computing Centre at the University
511 of Tartu. The authors thank Märt Roosaare, Ulvi Talas, and Priit Palta for performing a critical reading
512 of the manuscript.

513

514 **AUTHOR CONTRIBUTIONS**

515

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

523

524 **COMPETING FINANCIAL INTERESTS**

525

526 The authors declare no competing financial interests.

527 **FIGURES**

528

529 **Figure 1.** Overall principle of k -mer-based genotyping.

530

531 **Figure 2.** Illustration of genotype calling based on the frequencies of two k -mers. The parameters that
532 define boundaries between genotypes are estimated from the k -mer frequency data of each individual.
533 By default, only conventional genotypes are reported in the output. “A” denotes the reference allele,
534 and “B” denotes an alternative allele. The estimated depth of coverage (λ) of the individual used in this
535 example was 38.6.

536

537 **Figure 3.** Performance of Bayesian classifier with simulated data. Sensitivity and specificity of calling
538 alternative alleles is shown for reference genome and simulated genome of individuals from 5 different
539 populations. Populations are abbreviated as follows: EST – Estonian; CHS – Southern Han Chinese;
540 PUR – Puerto Rican; CEU – Utah residents with Northern and Western European Ancestry; YRI –
541 Yoruban from Ibadan, Nigeria. Specificities are nearly identical for all individuals and thus their lines
542 are overlapping with the green dotted line.

543

544 **Figure 4.** Effect of genome coverage on the concordance of genotypes. The accuracy of calling non-
545 reference variants starts to decline as the genome coverage drops below 20x. Only the accuracy of the
546 non-reference allele (genotypes AB and BB) calls declines significantly as the coverage drops because
547 the higher prior probability of the reference allele has a stronger influence on the final decision of the
548 Bayesian classifier in situations where the coverage is low (which increases the bias toward the more
549 common allele).

550

551 **Figure 5.** Effect of k -mer length on filtering SNV markers. Lines show the fraction of remaining
552 markers after filtering steps 1 and 2. The step 1 removes SNVs that have other marker within k
553 nucleotides on both sides. The step 2 removes SNVs that have no unique k -mers in the expanded
554 reference genome. Y-axis indicates the fraction of remaining markers after each filtering step. 100% in
555 this figure corresponds to 46,954,719 SNVs from the dbSNP that were fed to filtering pipeline.

556

557 **Figure 6.** The time spent counting k -mer frequencies is proportional to the genome coverage (because
558 of the larger FASTQ files). `Gmer_counter` is able to read data from multiple files simultaneously;
559 thus, it runs faster if the sequence data are distributed between different files (e.g., files with paired
560 reads).

561 **TABLES**

562

563 **Table 1.** Genotypes retrieved from the simulated reads generated from the reference genome. “A”

564 denotes the allele from the reference genome, and “B” denotes the alternative allele. “NC” is no-call.

		Coverage				
		5x	10x	20x	30x	40x
FastGT genotype calls	AA	28,734,597 (98.943%)	29,025,104 (99.943%)	29,027,872 (99.952%)	29,025,157 (99.943%)	29,011,146 (99.895%)
	AB	6,140 (0.021%)	12,107 (0.042%)	13,567 (0.047%)	11,007 (0.038%)	10,898 (0.038%)
	BB	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	NC	300,941 (1.036%)	4,467 (0.015%)	239 (0.001%)	5,514 (0.019%)	19,634 (0.068%)

565

566

567

568 **Table 2.** Concordance between the autosomal genotypes of two individuals from the Platinum dataset

569 and bi-allelic FastGT genotypes called from the same individuals. The reference allele is denoted by

570 “A” and the alternative allele is denoted by “B” denotes the alternative allele.

571

		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%

572

573

574

BAM file

↓ samtools or
↓ Picard or
↓ bam2fastq

ca 1 - 3 hours for
converting BAM
files to FASTQ files
(if needed)

Premade
database of
allele-specific
k-mers

FASTQ file(s)

bioRxiv preprint doi: <https://doi.org/10.1101/060822>; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

gmer_counter

ca 45 minutes for
k-mer counting

Text file with frequencies
of *k*-mer pairs

gmer_caller

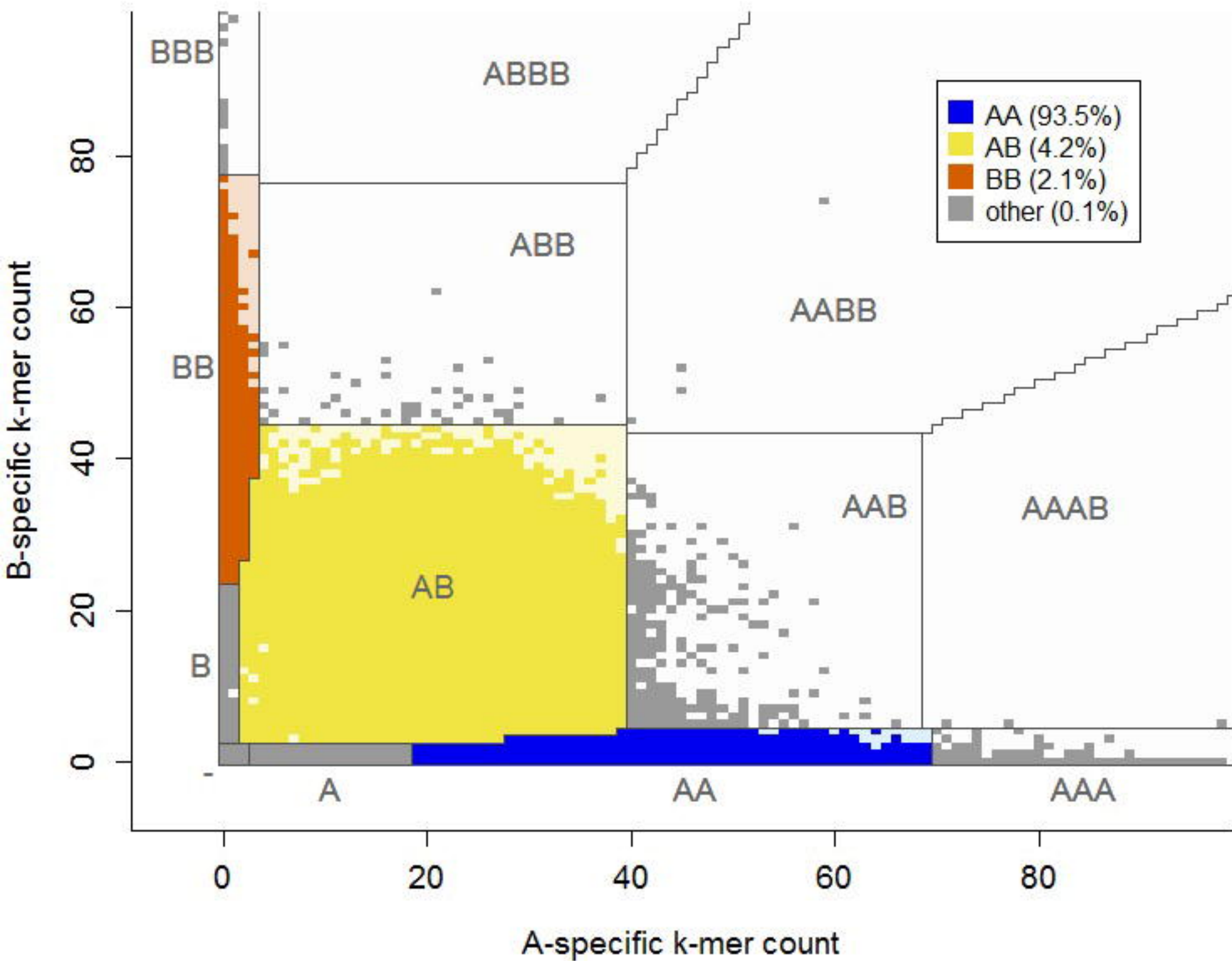
ca 4 minutes for
genotype calling
from counts

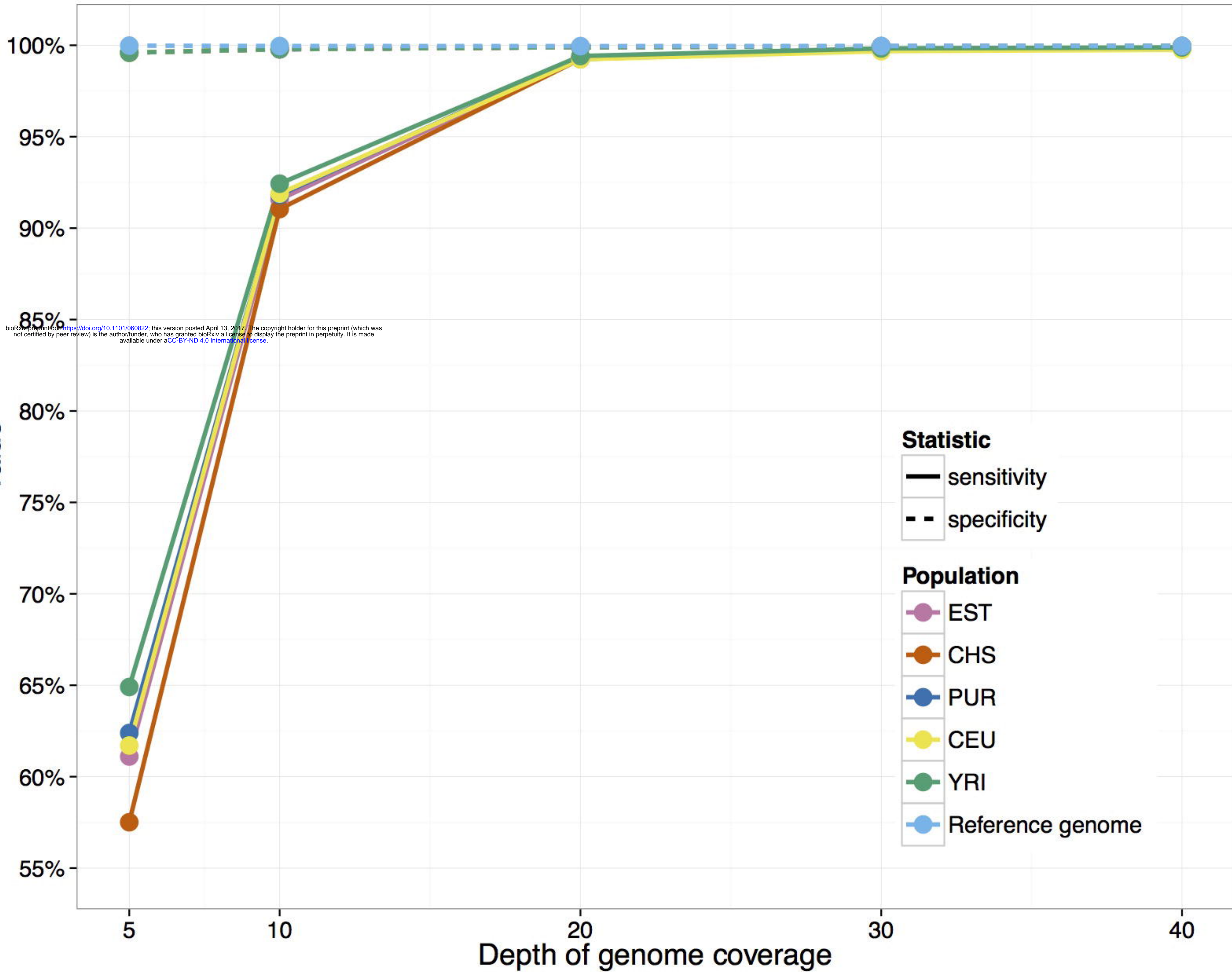
Text file with genotypes and
their probabilities

↓ formatting
↓ scripts

optional step

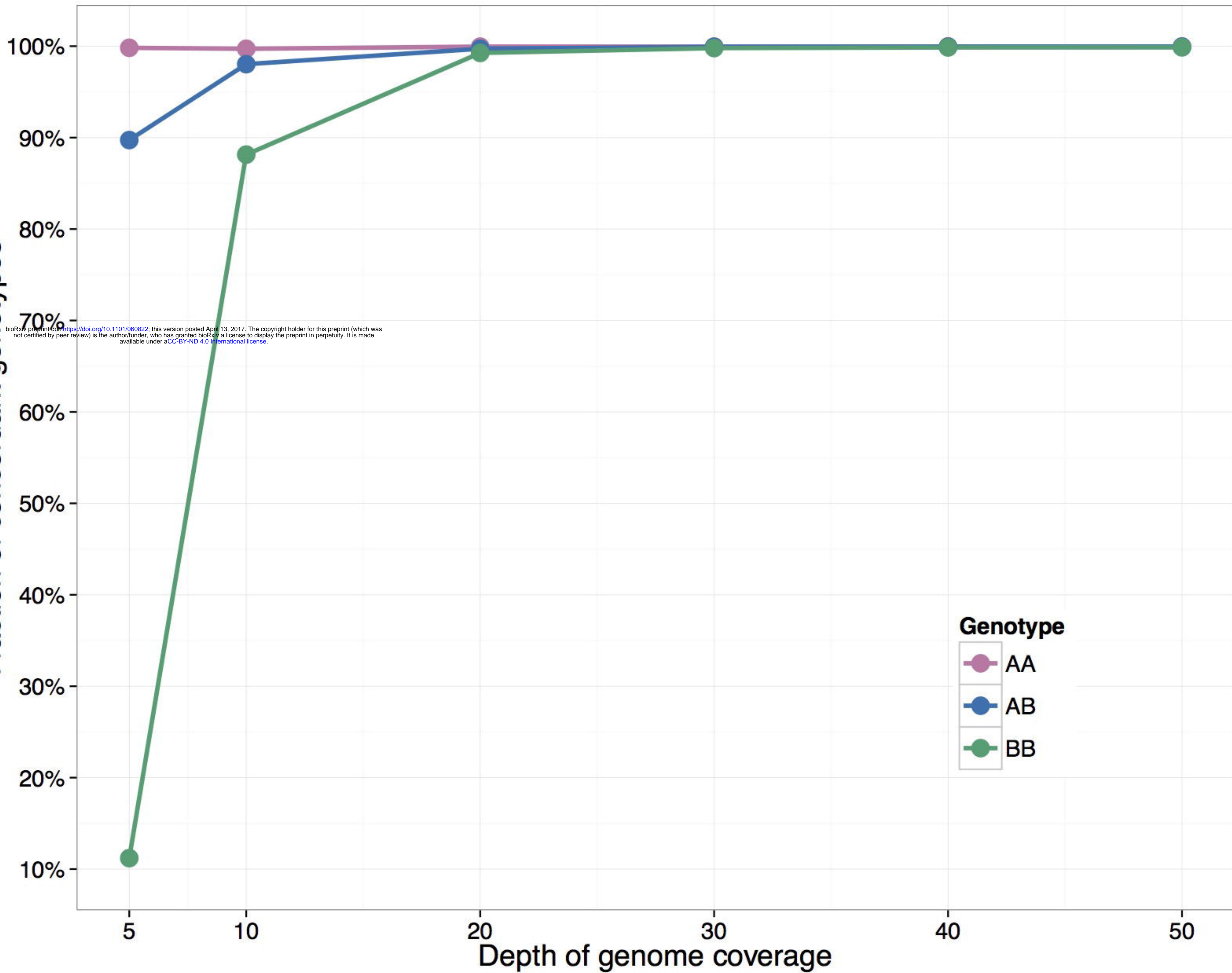
Genotype file in .vcf format



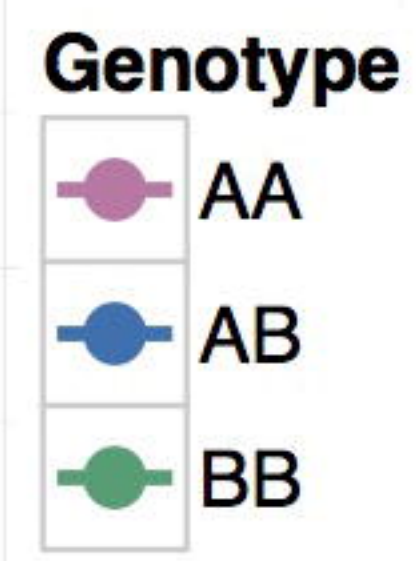


bioRxiv preprint doi: <https://doi.org/10.1101/060822>; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fraction of concordant genotypes

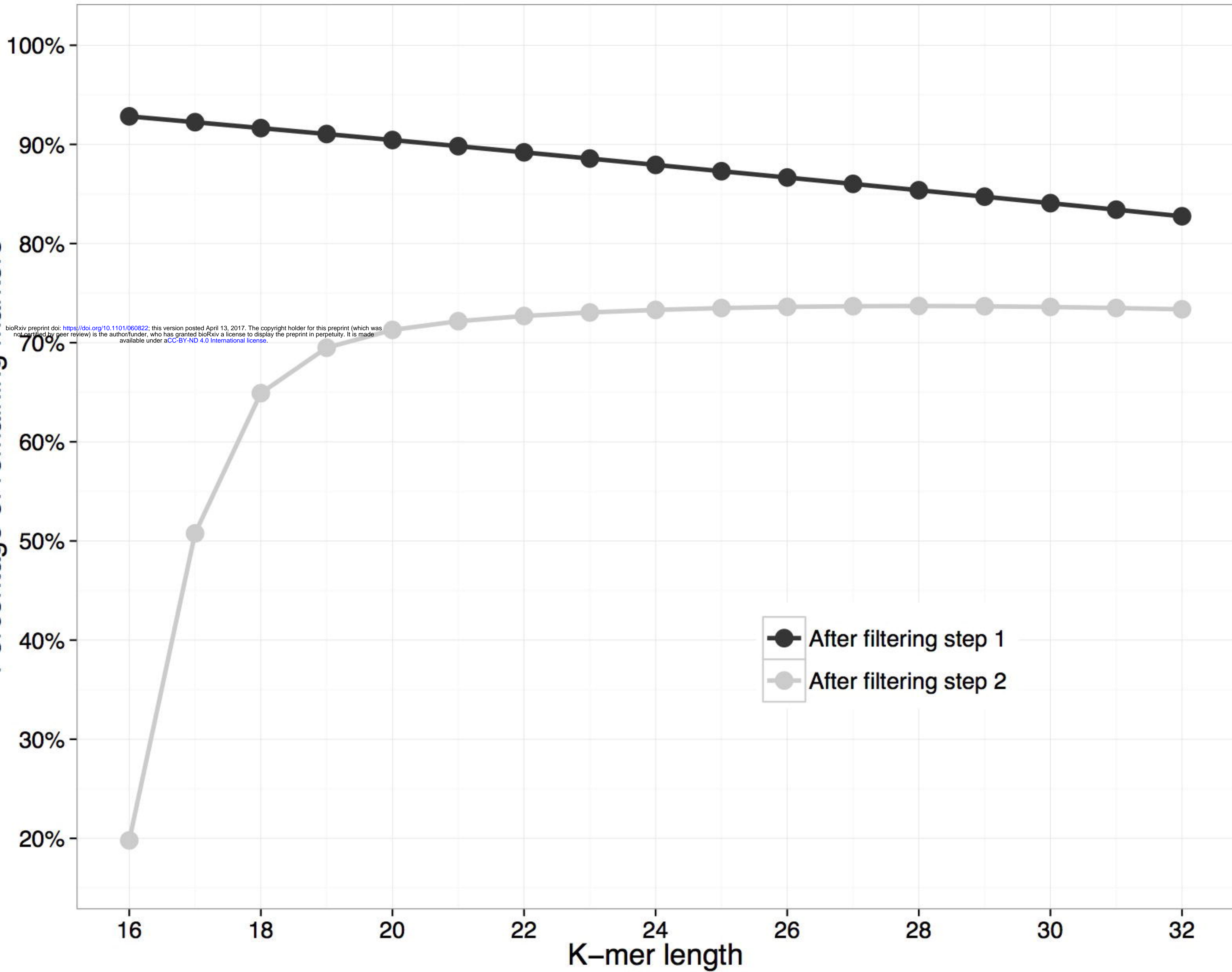


bioRxiv preprint doi: <https://doi.org/10.1101/060822>; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: <https://doi.org/10.1101/060822>; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Percentage of remaining markers



● After filtering step 1
● After filtering step 2

K-mer length

Gmer_counter running time (min)

Nr of FASTQ files

- one
- two
- four

bioRxiv preprint doi: <https://doi.org/10.1101/060822>; this version posted April 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

120
100
80
60
40
20
0

5 10 15 20 25 30 35 40 45 50

Depth of genome coverage

