Comparison and Validation of the SNPTools Genotype Likelihood Model Using Low Depth Data

Michael Ben Ezra (lxc844)

2018-01-22

Abstract

High-Throughput sequencing (HTS) yields data which suffer from errors in base-calling leading to mapping errors and eventually to low confidence variant and genotype calling. The problem is exacerbated when dealing with low-depth data due to the potential lack of sampling of both chromosomes in diploid individuals, as well as the general scarcity of data per genomic loci. The SNPTools pipeline proposes improvements in modelling of low depth data, promising more accurate variant and genotype calling. In particular, SNPTools claims to accurately estimate genotype likelihoods for putative variant sites. Genotype likelihoods express the probability of observing the base reads given the genotype and are used in probabilistic frameworks that model and quantify the uncertainty involved in calling variants and genotypes. We validate SNPTools' genotype likelihood model under low-depth data conditions and compare it to the widely used SAMTools and GATK models. We find that SNPTools' genotype likelihood model leads to more accurate genotype calls.

1. Introduction

High-throughput sequencing (HTS) is an inherently error prone process, yielding data which suffer from a variety of potential errors, among them incorrect or low confidence base-calling as well as errors occurring during de novo assembly or alignment to a reference genome.[6] Since the downstream analysis of virtually all genetic and genomic studies rely heavily on genetic variant detection and called genotypes, selecting a statistical model of the data that best represents the true nature of the genetic substrate is crucial to correct interpretation of results.

In deeply sequenced regions of the genome, uncertainty due to low confidence base-calling can be handled heuristically simply by discarding reads with low quality scores without greatly impacting the power to confidently call variants and genotypes, this is due to the ample reads at such loci. However, low-depth data (<5X per site) on which many NGS studies rely is highly susceptible to uncertainty in variant and genotype calling. This is due to the scarcity of data at such loci and the high probability that only one of the two chromosomes is sampled in diploid individuals, making it particularly difficult to detect rare mutations and call heterozygous genotypes with high confidence. In order to maximize and quantify the certainty in genotype and variant calling under low-depth data conditions we must utilize all the data at our disposal (i.e. we can not afford to discard bases simply because they are associated with a low quality score.) Sophisticated probabilistic frameworks that model the errors such as those associated with base calling and mapping therefore use genotype likelihoods - the probability of observing the data (base calls) given a parameter (the genotype.)

The SNPTools pipeline seeks to address some of the analytical challenges in low-coverage genome sequencing data by (1) modelling the weight of each base read using both mapping and base-quality; to be used in both variant calling and genotype likelihood estimation, (2) aggregating reads from all samples to identify alternative alleles (3) calling variant sites by assessing the distribution of alternative alleles across samples to minimize the influence of sequencing errors (4) calling variant sites by assessing the best genotype fit, (5) utilizing millions of putative SNP sites within each BAM to better estimate genotype class binomial probabilities from which to calculate genotype likelihoods.[6]

In this work we seek to compare the SNPTools genotype likelihood model to that of SAMTOOLS and GATK under low-depth sequencing data conditions. To asses the effect of the genotype likelihood model on downstream analysis we examined the accuracy of resultant genotype calls and calculated the site frequency spectrum (SFS) from genotype likelihoods. To judge the accuracy of the genotype likelihoods calculated by SNPTool we examined their performance on real data by calling genotypes and comparing to the supposed true genotypes as specified in the HapMap.

2. Materials and Methods

Github

All scripts employed in data analysis for the purposes of this report can be found at https://github.com/mbenezra/angsd_snptools

1000 Genomes Project Data Set (phase 3)

The 1000 genomes project is the largest public catalogue of of human variation and genotype data. We applied the SNPTools pipeline to 90 individuals from the CEU population.[1]

HapMap data

The International HapMap Project developed a haplotype map of the human genome, describing common patterns of human variation. The publicly available HapMap data provides genotype data for selected sites (n=1222022) of the 90 CEU individuals from the 1000 genomes project. We employ the genotype HapMap data when validating genotype calls.

Analysis of next generation Sequencing Data (ANGSD)

We employ the ANGSD[2] program for genotype likelihood calculations based on the GATK and SAMTOOLS models as well as genotype calling.

GATK Genotype Likelihood Model

GATK[4] calculates the posterior probability of each of the possible 10 diploid genotypes from next generation DNA sequencing reads. GATK incorporates the bases covering each locus and their respective phred quality scores and the derived probability that the read is incorrect e. The prior probability of each of the possible 10 diploid genotypes G given the data D is calculated using Bayes' rule.

$$p(G|D) = \frac{p(G)p(D|G)}{p(D)}$$

$p(D)$ is constant across all genotypes and can be ignored. $p(G)$ is the prior probability of observing a genotype and can be set to a uniform prior or an allele frequency prior.

$p(D|G)$ is the product of the probabilities of each base given the genotype.

$$p(D|G) = \prod_{b \in pileup} p(b|G)$$

The probability of a base given a genotype is calculated as half of the probability of observing the first allele plus half of the probability of observing the other allele.

$$p(b|G) = p(b|\{A_1, A_2\}) = \frac{1}{2}p(b|A_1) + \frac{1}{2}p(b|A_2)$$

The probability of observing a base given an allele is $1 - e$ if the base is equal to the allele since e is the probability that the read is incorrect and therefore $1 - e$ is the probability that the read is correct. The probability of observing any of the other 3 alleles is therefore $e/3$

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases}$$

Since SNPTools employs a bi-allelic assumption by selecting only a single alterntive nucleotide for variant sites, we fix the major and minor alleles and calculate the posterior probabilities of 3 diploid genotypes also when employing the GATK genotype likelihood model.

SAMTOOLS

Similarly to GATK, SAMTOOLS[3] employs a genotype likelihood model that incorporates quality score, however, SAMTOOLS also incorporates quality dependency.

SNPTools

Effective Base Depth (EBD) calculation

SNPTools[6] calculates a read depth pseudo-count termed Effective Base Depth (EBD) that weights reads based on their base quality and mapping quality. This allows SNPTools to assign a lower weight to base reads with high sequencing and mapping errors that may lead to incorrect identification of alternative alleles (this is particularly problematic where read depth is low.)

An EBD value is calculated for each nucleotide of every locus in a sample.

$$EBD_{s,g=A,C,G,T} = \sum_{k}^{K_s} (1 - BaseQuality_k) \times (1 - MappingQuality_k), \text{for all k=g.}$$

SNP Site Discovery (variance ratio statistic)

SNPTools employs a variance ratio statistic that pulls together EBD pseudo-count information from all samples and assess the distribution of alternative allele read counts across the population to call variant sites. Subsequently, SNPTools employs all and only variant sites in modelling sequencing error in base calling when calculating genotype likelihoods. Specific details of the VarianceRatioStatistic are beyond the scope of this report and the mathematical formulations of the test statistic are provided here for reference. Some of this notation will also be relevant in the Genotype Likelihood estiamation section to follow.

$$VarianceRatioStatistic = \frac{\sum_{i=1}^{I} [a_i - e(a_i + r_i)]^2 - Te(1 - e)}{\sum_{i=1}^{I} Min\left\{[a_i - 0(a_i + r_i)]^2, [a_i - \frac{1}{2}(a_i + r_i)]^2, [a_i - 1(a_i + r_i)]^2\right\}}$$

$$a_i = EBD_{i,g=alternativeallele}$$

$$r_i = EBD_{i,g=referenceallele}$$

$$T = \sum_i^I (a_i + r_i)$$

$$e = \frac{\sum_i^I (a_i)}{\sum_i^I (a_i + r_i)}$$

Genotype Likelihood estimation (BAM-specific Binomial Mixture Modelling BBMM)

Genotype data likelihoods can be modelled as a binomial distribution and calculated using the binomial mass function as a series of trials represented by the effective base depth (EBD) counts $r_s + a_s$ for each site with the success binomial probability p_v of a reference read defined for each of the genotype classes rr=Ref/Ref, ra=Ref/Alt and aa=Alt/Alt where v is the genotype class. If sequencing was error free, the binomial probabilities of a reference read for each of the genotype classes would be $p_{rr} = 1$, $p_{ra} = 0.5$ and $p_{aa} = 0$.

$$Binomial(r_s + a_s, p_v) = \begin{bmatrix} r_s + a_s \\ a_s \end{bmatrix} p_v^{a_s} (1 - p_v)^{r_s}$$

In reality the binomial probabilities of a reference read at each of the genotype classes deviate from the theoretical values. Moreover, general operational heterogeneity in a collection of samples (ex: sequencing runs, sequencing centres, etc) contribute to variability that decreases the signal to noise ratio. For this reason SNPTools estimates BAM-specific values of the binomial probabilities. In turn these BAM-specific probabilities are used for calculating the genotype data likelihoods as described above.

In order to estimate the binomial probabilities p_v, SNPTools models each BAM as a flexible binomial mixture. The binomial mixture is a linear superposition of three binomials each representing one of the genotype classes (rr, ra, aa). Each of the three binomials accounts for a certain proportion of the mixture and is assigned a weight w_v, put together the weights add up to 1.

In order to approximate the binomial probabilities p_v, SNPTools employs the Expectation Maximization algorithm. Each site is assigned to a genotype class by way of a latent variable $z_{s,v}$, a site can only be assigned to a single genotype class at a time. See [TB1] for complete derivation of the Expectation Maximization algorithm for mixture models.

The EM algorithm seeks to maximize the log likelihood function with respect to the parameters p_v and w_v.

$$ln(p(r_s, a_s | p_v, w_v)) = \sum_s \sum_{v=rr,ra,aa} ln(w_v * Binomial(r_s + a_s, p_v))$$

Initial values are selected for the parameters p_v and w_v after which the conditional probability of the latent variables given the data and the initial parameters is calculated (E-step). The EM algorithm then calculates new values for the unknown parameters based on the conditional probability of the latent variables (M-step). Once the algorithm converges, the best estimate for the parameters p_v is obtained and genotype likelihoods are calculated as a binomial probability of the data given the binomial probabilities of a reference read at each of the genotype classes.

$$GL_{s,v=rr,ra,aa} = p(a_s | z_{s,v}, p_v) = Binomial(r_s + a_s, p_v)$$

The mathematical formulations for SNPTools above are taken from [6].

3. Results

Depth

To confirm that we are experimenting on low-dept data, we used the SAMTOOLS depth command to output the depth of all sites in each of the 90 CEU population BAM files and plotted the distribution of depths in Figure 1 (see Figures section). The total depth of the collection of CEU samples is also plotted in red. We observe a mean population sequencing depth of approximately 6-7X. Some of the samples exhibit a higher mean sequencing depth of up to approximately 15X while others are as low as approximately 3X.

Comparison of genotype call accuracy and validation of genotype calls

To asses the quality of the genotype likelihoods estimated by SNPTools we called genotypes using each of the genotype likelihood models (SNPTools, SAMTOOLS and GATK) using both a uniform and an allele frequency prior. We plotted the proportion of non-calls vs call accuracy (Figure 2 - see Figures section) for every quality threshold and highlighted the 0.95 quality threshold in the plot (see diamond markers on lines.)

We called genotypes for the variable sites previously identified by SNPTools for which we also have data in the HapMap (n=1222022). We instructed ANGSD to select the major and minor alleles from GL using maximum likelihood.

The comparison paints a clear picture of better performance under an allele frequency prior in general. We also observe that SNPTools genotype likelihoods yield the most accurate genotype calls. At the 0.95 quality threshold SNPTools performs better than both SAMTOOLS and GATK though at the cost of a higher proportion of non-calls.

In Table 1 we summarize the number of correct and incorrect calls at the 0.95 quality threshold for the various callers. We observe that under the allele frequency prior SNPTools makes more correct calls when compared to the uniform prior, at the same time SNPTools also makes fewer incorrect calls.

Table 1: Genotype call totals at 0.95 genotype likelihood confidence level

	Correct Calls	Incorrect Calls
SNPTools (uniform prior)	88479071	810663
SNPTools (frequency prior)	94369752	794119
SAMTOOLS (uniform prior)	89870918	928526
SAMTOOLS (frequency prior)	96052705	944089
GATK (uniform prior)	90271881	991874
GATK (frequency prior)	96400511	1008880

Site Frequency Spectrum (SFS)

We calculate a folded Site Frequency Spectrum (SFS) for 10 random individuals from the CEU population based on the variable sites identified by SNPTools (n=12212852) from genotype likelihoods. We could only calculate the SFS for the variable sites identified by SNPTools due to the fact that SNPTools calculates genotype likelihoods for variable sites only. We fixed the major and the minor alleles to those identified by SNPTools for the purposes of comparison with SAMTOOLS and GATK, the site allele frequencies were then calculated based on individual genotype likelihoods assuming HWE. Since we do not polarise the SFS to an out-group of closely related species (such as the chimp) the SFS contains no information about ancestral or derived states and we instead calculated the folded SFS spectrum. The normalized spectra for polymorphic sites are shown in Figure 3 (see Figures section).

The three genotype likelihood models yield very similar site frequency spectra as well as very similar variability (calculated as proportion of heterozygous sites.) We observe that the three spectra indicate the existence of approximately 40% non-variable sites of the total 12212852 identified as variable by SNPTools.

4. Discussion

In this report we compared SNPTools' genotype likelihood model to that of SAMTools and GATK. We did so by calling genotypes of 90 CEU population individuals from the 1000 genomes project database and comparing with true genotypes in the HapMap. We observed the genotype call accuracy and the proportion of non-calls at the 0.95 confidence threshold and concluded that SNPTools performs better than both GATK and SAMTOOLS though at the cost of calling fewer genotypes at the 0.95 threshold. Additionally, the plot comparing the genotype call accuracy and proportion of non-calls indicates higher accuracy of SNPTools at all thresholds.

We were surprised to observe very similar site frequency spectra due to the three different genotype likelihood models. We hypothesize that due to the highly curated nature of the 1000 genomes projects' data where base qualities have been recalibrated to high accuracy, SNPTools' supposed contribution in better modelling the BAM-Specific error profile was less pronounced. We know that GATK's genotype likelihood model assumes that the base quality scores and correct and in the case of the 1000 genomes project data this turns out to be a good assumption. We recommend running a similar analysis on data where quality scores are less certain. Moreover, since SNPTools' proposes to better model low-depth data, it would also be advisable to re-run the analysis on a data set with a better defined depth profile, while the mean depth of our samples was low, the variance of depths among the samples was high.

5. Acknowledgements

Supervisor: Thorfinn Sand Korneliussen

6. References

TB1. Pattern Recognition and Machine Learning Information Science and Statistics, ISSN 1613-9011 Author Christopher M. Bishop

TB2. An Introduction to Population Genetics: Theory and Applications, Rasmus Nielsen, Mongomery Slatkin.

1. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
2. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014;15:356.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
5. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12(6):443-51.

6. Wang Y, Lu J, Yu J, Gibbs RA, Yu F. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. Genome Res. 2013;23(5):833-42.

7. Figures

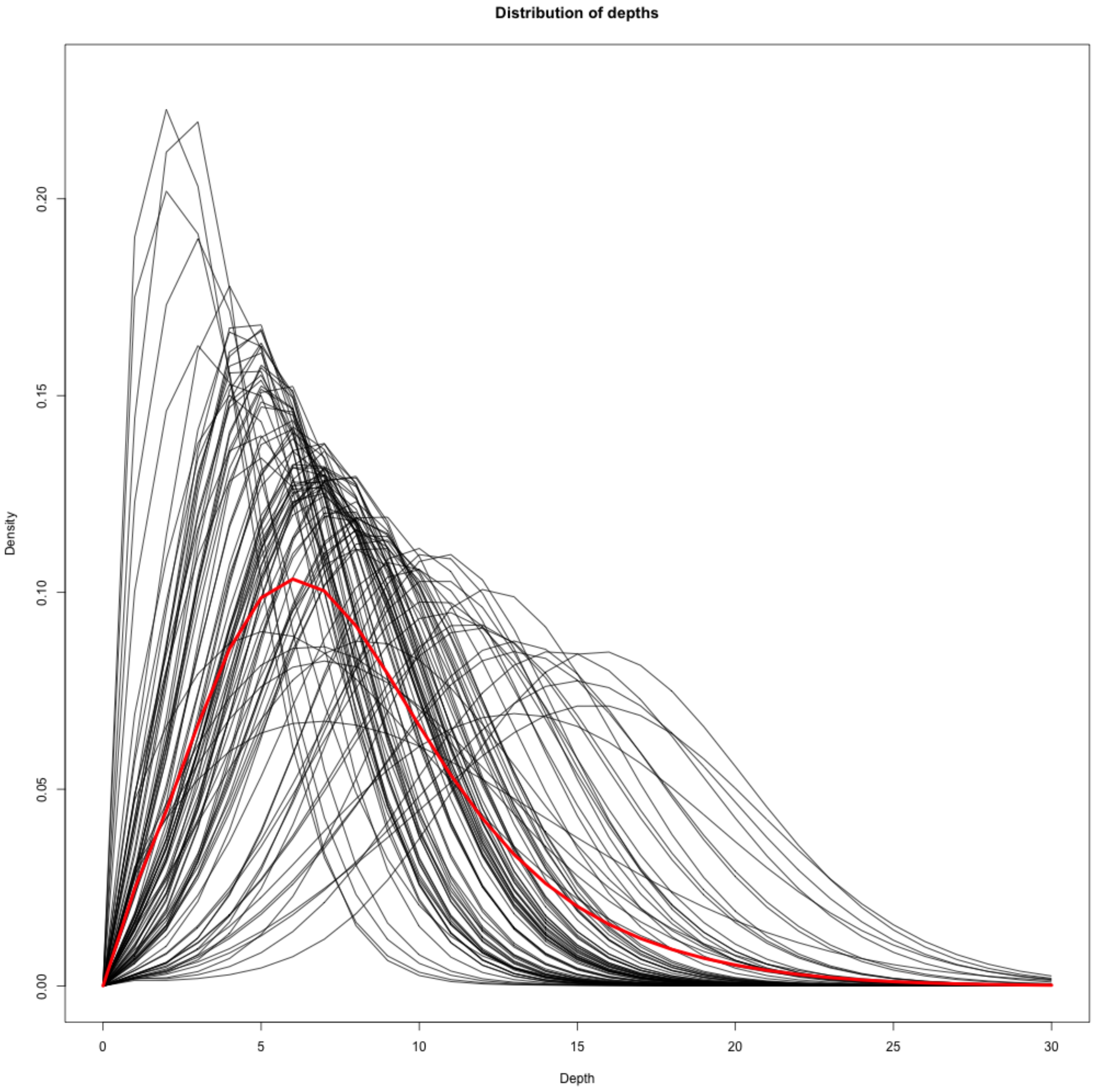

Figure 1: Distribution of depths for each of the 90 CEU individuals and the distribution of depths for all the samples put together (in red)

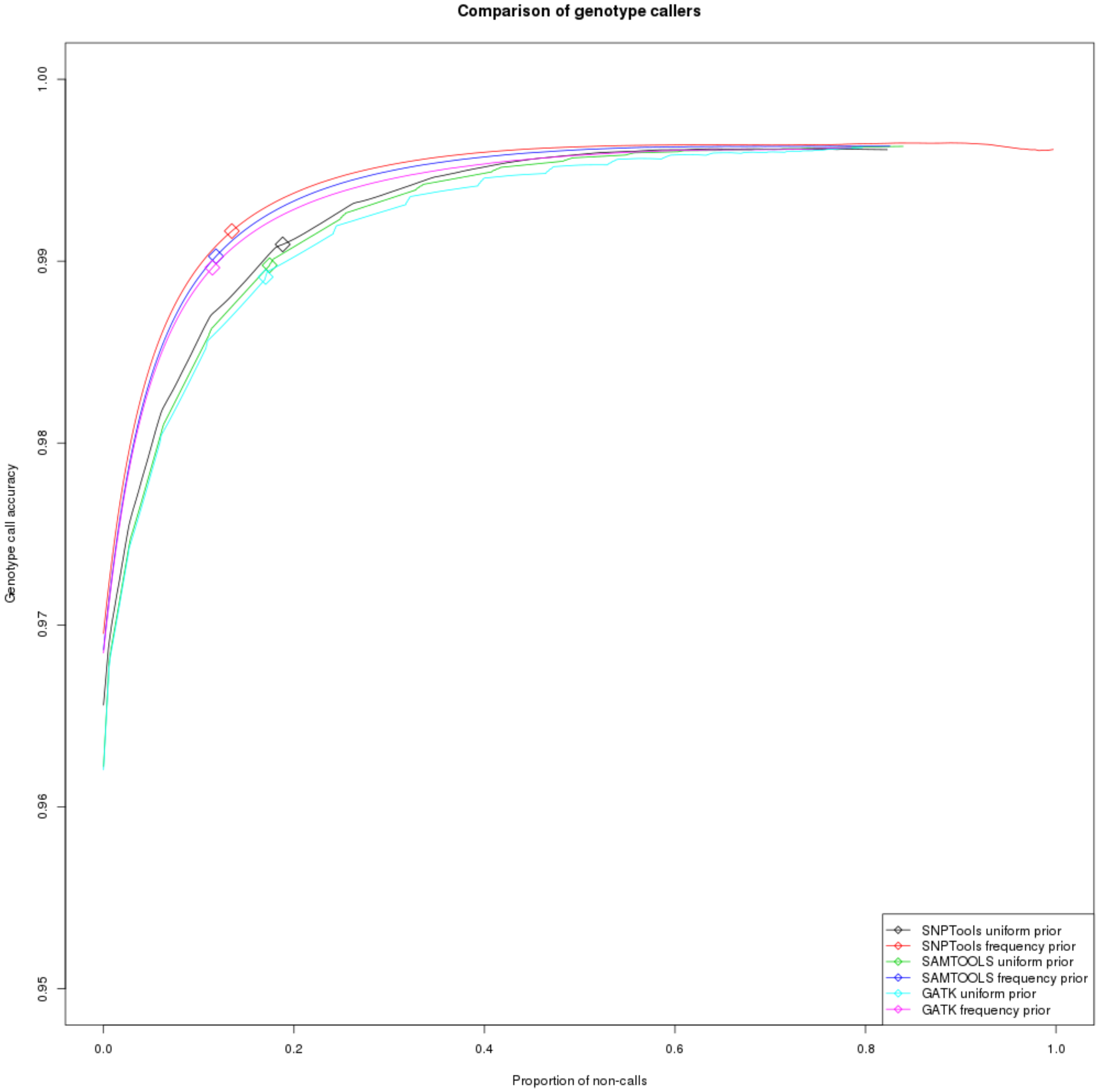

Figure 2: A comparison of genotype calls made using genotype likelihoods from SNPTools, SAMTOOLS and GATK.

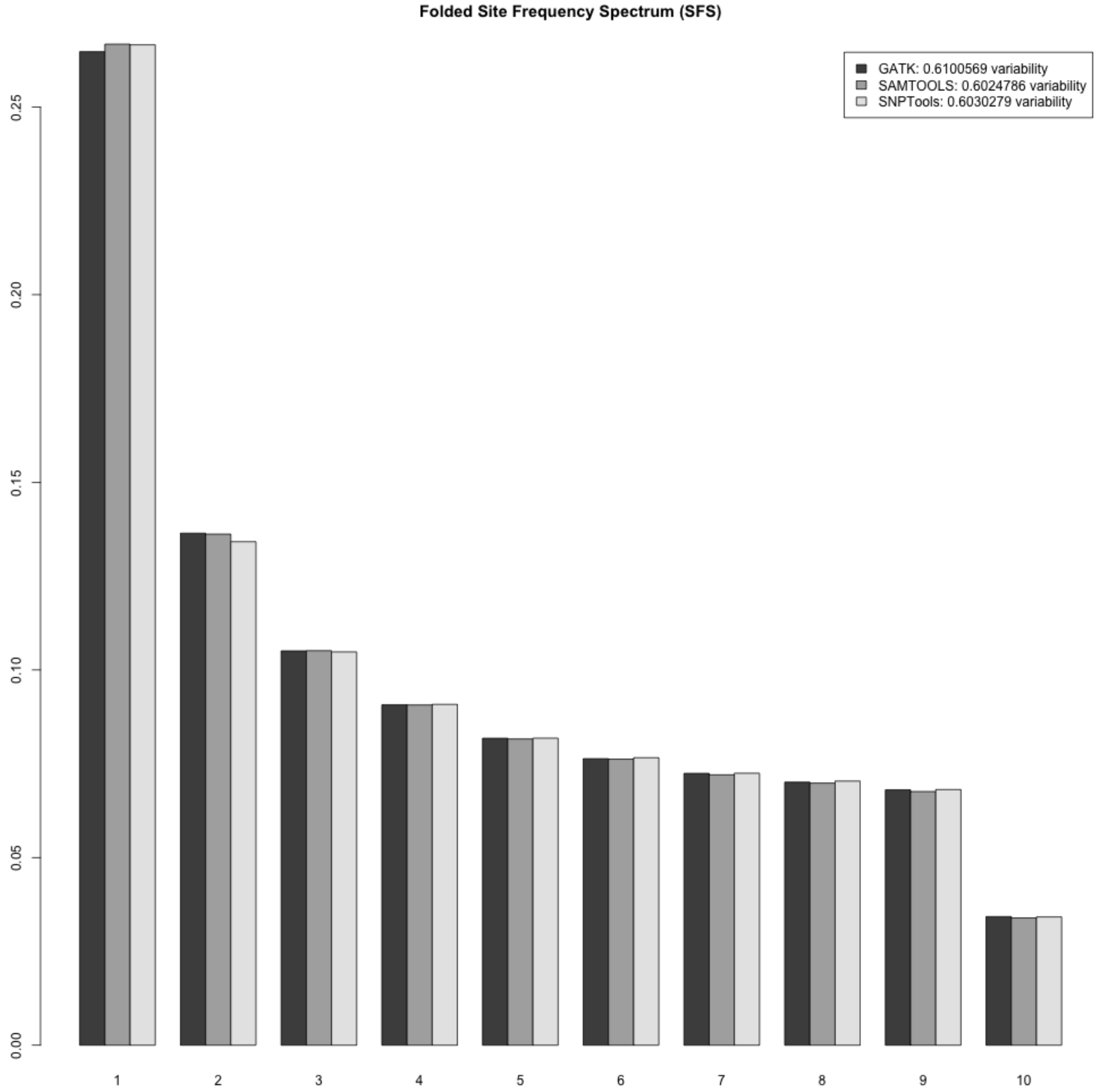

Figure 3: Folded Site Frequency Spectrum (SFS) for SNPTools, GATK and SAMTOOLS