UC San Diego
SCHOOL OF MEDICINE

Detection and genotyping of short indels using sequence data from multiple samples

Alastair M. Kilpatrick and Vikas Bansal
Department of Pediatrics, University of California San Diego, 9500 Gilman Drive, 92093 La Jolla CA, USA

Introduction

Short insertions and deletions (indels) are the second most common type of variation in the human genome. Recent studies estimate more than 1 million indels in the NA12878 genome [1]. Despite advances in high-throughput sequencing and computational methods for variant calling from DNA sequence data, accurate detection of indels remains a challenge. Some of the reasons for this difficulty include over-representation of short indels in regions of low sequence complexity [2], variability in indel error rates across different platforms as well as the lack of good error models for indels.

Methods

Figure 1 shows a simplified flow diagram of the indel calling method. For each variant, we find the maximum likelihood estimates of the context-specific error rates (*E*) and the population allele frequencies (*P*) using an approach based on the Expectation-Maximization (EM) algorithm. Given read counts for each individual, EM iteratively calculates the expected posterior probability of a given genotype for each individual (E-step):

$$p(G|R,\theta^{(t)})$$

These probabilities are then used to find the maximum likelihood estimates of the model parameters (M-step):

$$\theta^{(t+1)}=\underset{\theta}{argmax}\sum_{G} p(G|R,\theta^{(t)})\ln p(G,R|\theta)$$

Predicted variants are output in the VCF format. The EM algorithm is implemented as a module within the CRISP package for variant detection [3].

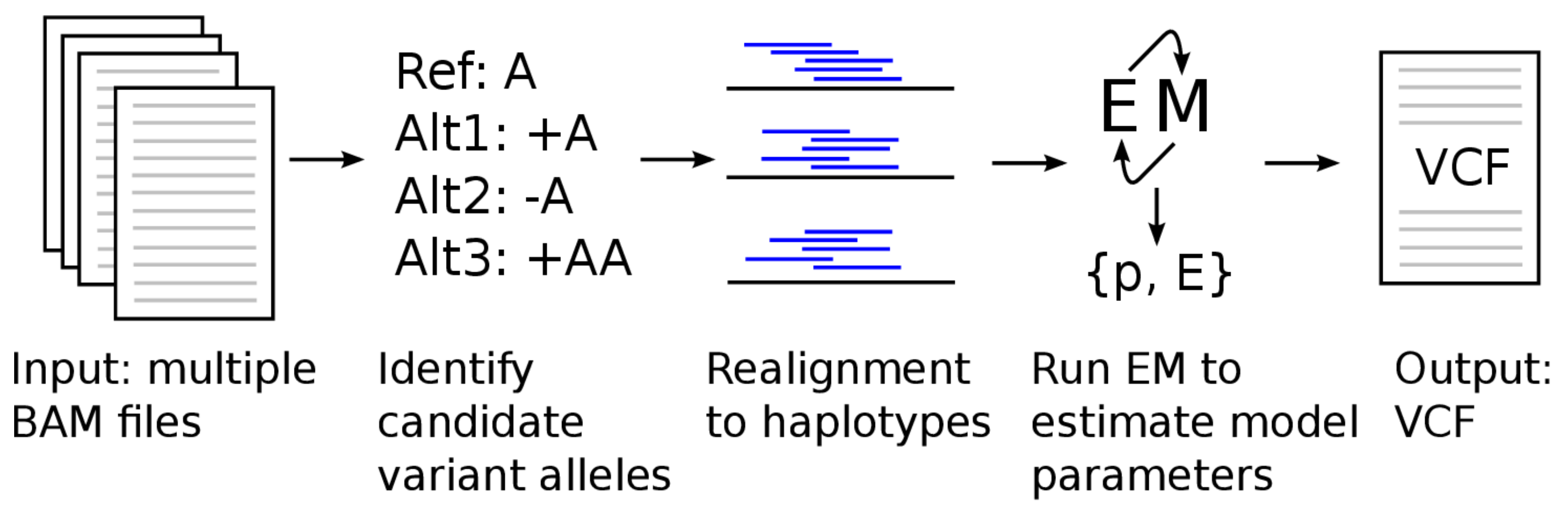
Ref: A
Alt1: +A
Alt2: -A
Alt3: +AA
E M
{p, E}
VCF
Input: multiple BAM files
Identify candidate variant alleles
Realignment to haplotypes
Run EM to estimate model parameters
Output: VCF

Figure 1. Flow diagram of the indel calling method.

Results

We assessed the performance of the EM algorithm against the population-based indel calling method SOAP-popIndel using simulated read count data. EM was found to return significantly fewer false positive results, particularly for simulated indels with high error rates (**Figure 2**).

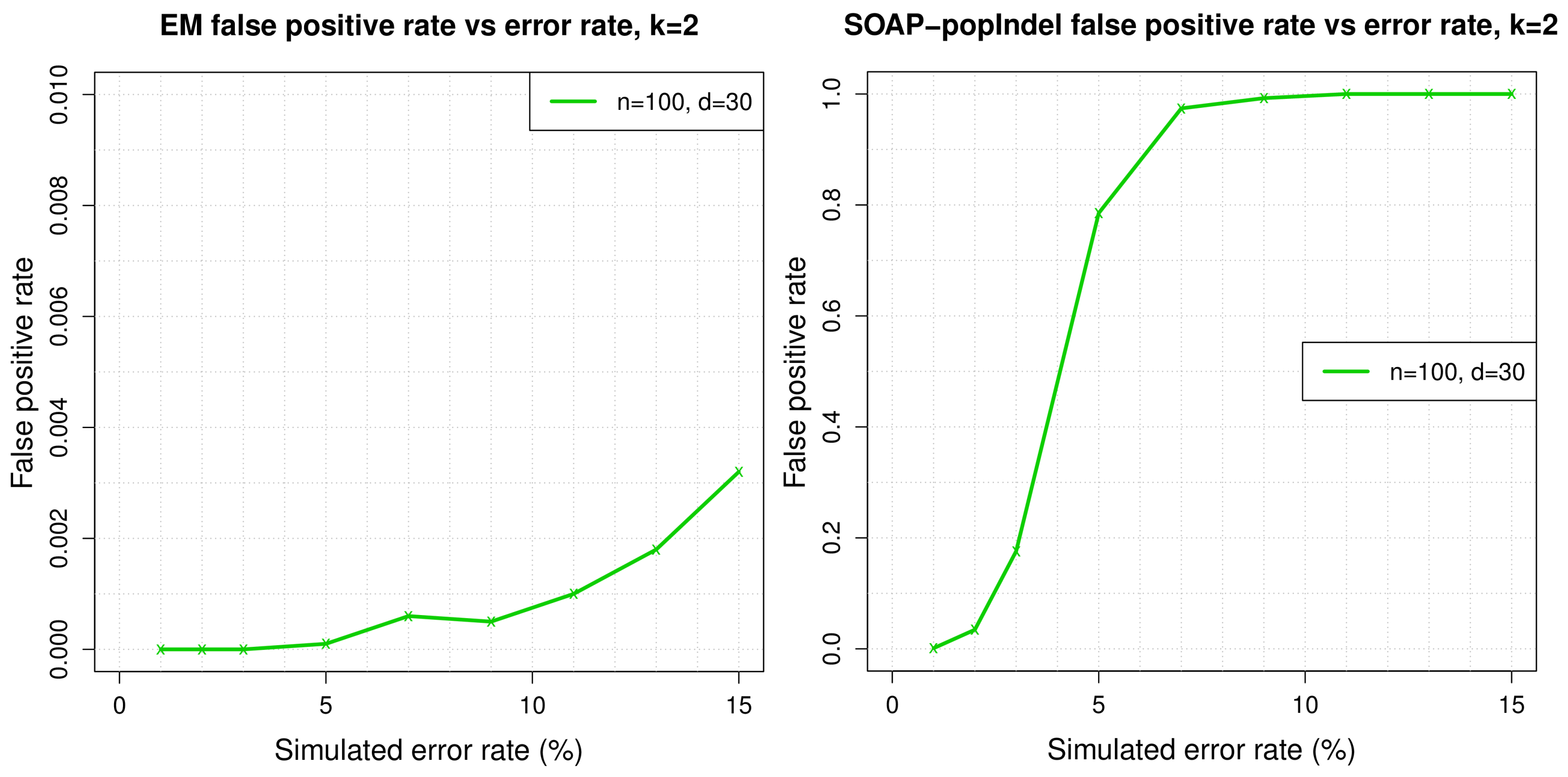

Figure 2. False positive prediction rate for EM (left) and SOAP-popIndel (right) on simulated read count data with 1 variant allele, 100 samples and read depth of 30x, as a function of simulated indel error rate.

We further assessed the performance of the EM algorithm using population sequence data from the 1000 Genomes Project Phase 3 dataset. We analyzed sequence data from 99 individuals of European (CEU) ancestry using EM, FreeBayes, Platypus and GATK HaplotypeCaller.

Concordance between indel calling methods is low: only 32.8% of predicted indels are predicted by all methods (**Figure 3**). EM and Platypus both call a lower number of unique indels, suggesting fewer false positive calls.

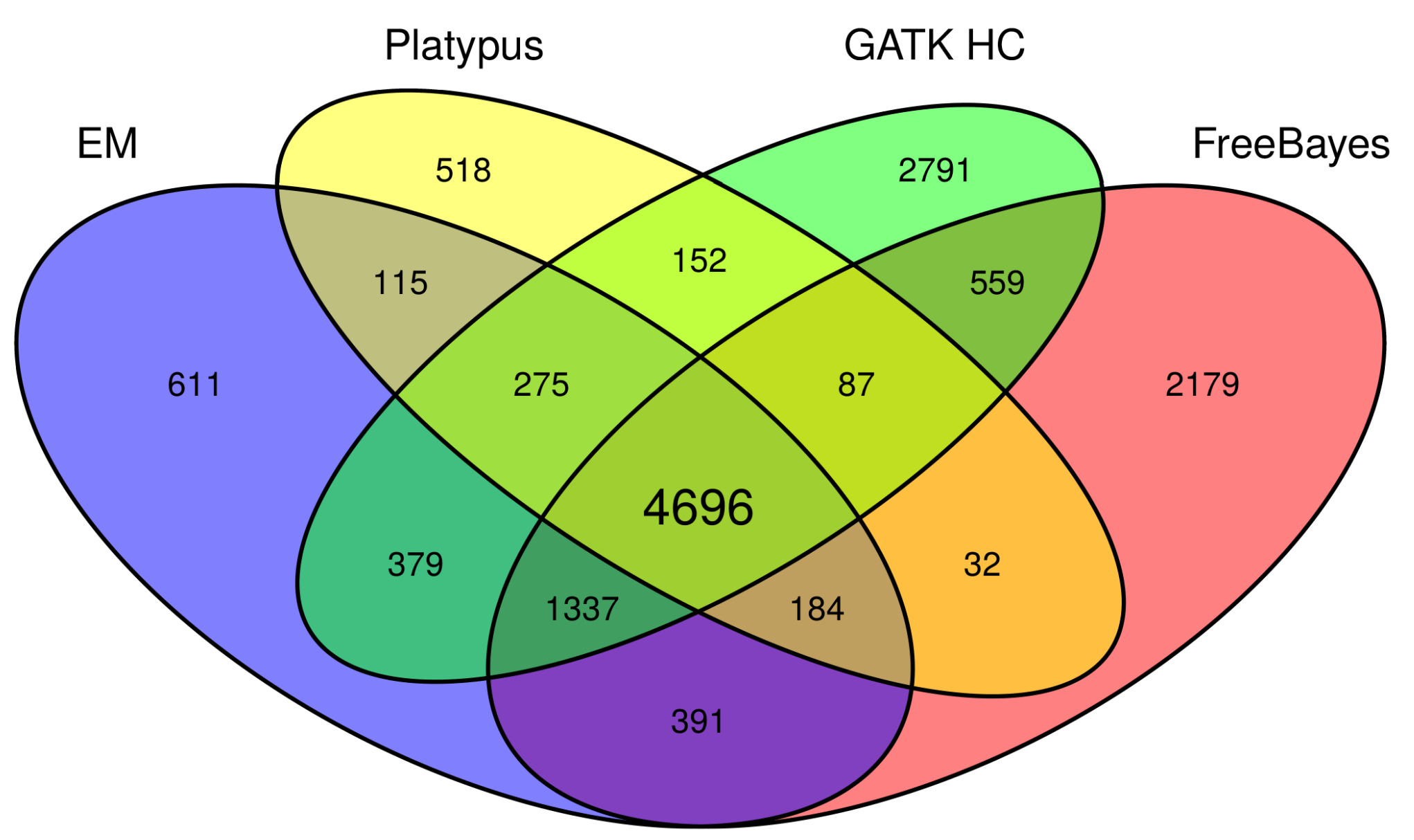

Figure 3. Venn diagram illustrating numbers of indels called in the CEU population by each tested method.

We also evaluated the sensitivity and false discovery rate for each of the methods using the gold standard Illumina Platinum Genomes call set for CEU individual NA12878 [4] (**Table 1**). EM is found to improve both sensitivity and false discovery rate over the other tested methods.

	Variants called	Sensitivity	FDR
EM	2,142	**0.7878**	**0.2063**
FreeBayes	2,353	0.7040	0.2954
Platypus	1,760	0.5516	0.2619
GATK HC	3,135	0.7040	0.4711

Table 1. Numbers of indels called in NA12878, sensitivity and false discovery rate for the four tested methods compared with Illumina Platinum Genomes 'platinum' confidence calls (2,355).

We analyzed the length of coding indels in NA12878 (**Figure 4**): a bias in favor of indels with length 3*n* was found, consistent with previous studies of coding indels and indicating low numbers of false positive predictions.

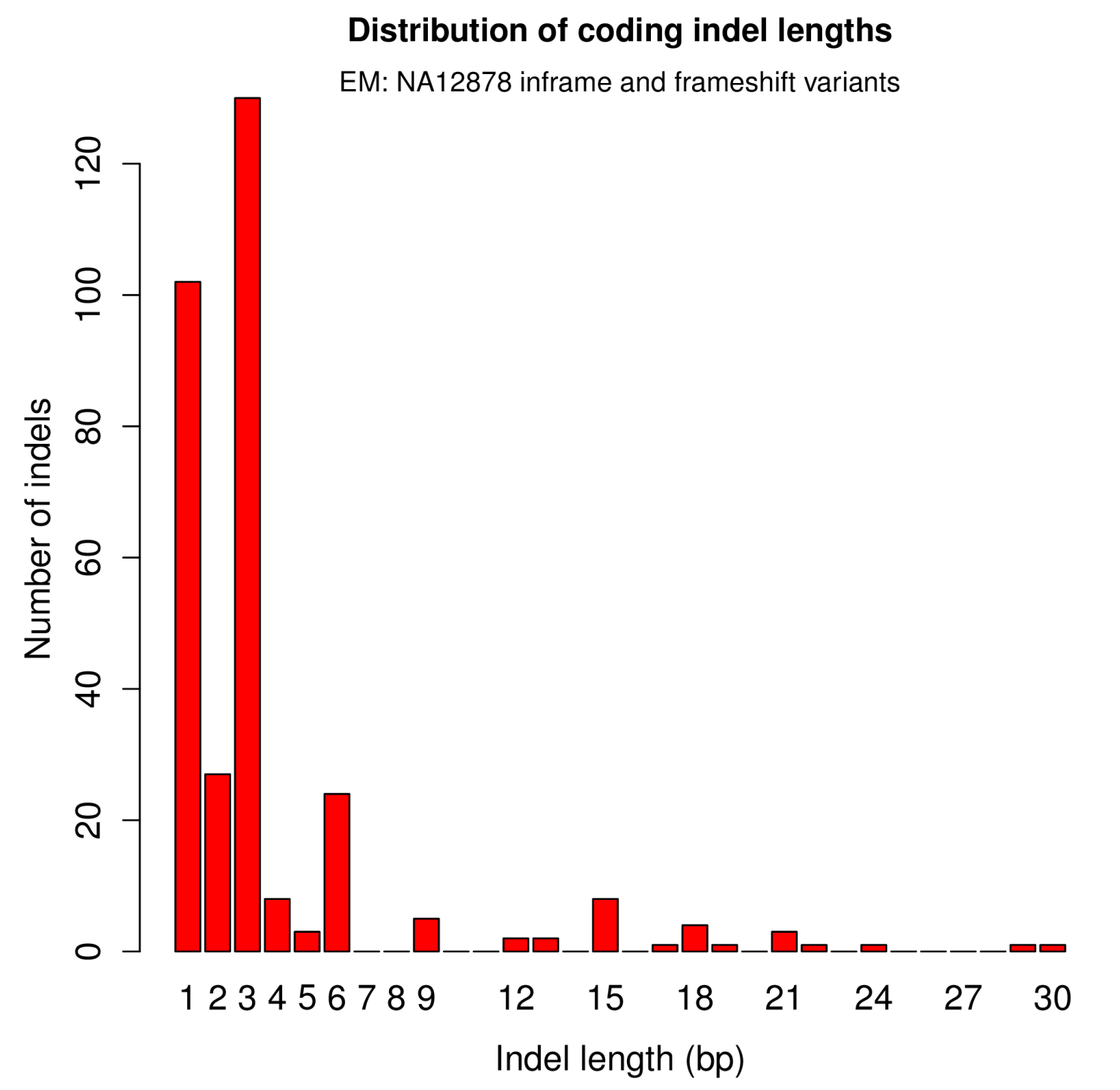

Figure 4. Distribution of coding indel lengths called by EM in NA12878.

Modeling context-specific error rates is shown to be particularly important for indel detection in regions of low sequence complexity. An estimated 19.4% of indels in the CEU population occur in regions with homopolymer, or di- or trinucleotide repeats of at least 10bp, the majority in homopolymer tracts. In NA12878, 34 Platinum Genomes indels were only called by EM; almost all were in homopolymer or di- or trinucleotide repeat regions.

Conclusion

Our EM algorithm is shown to be more accurate than other popular indel calling methods, in terms of sensitivity and false discovery rate in tests using Illumina Platinum Genomes data. Our method also shows a significant reduction in false positive results compared to the alternative population-based method SOAP-popIndel.

References

[1] Y. Jiang, A.L. Turinsky and M. Brudno. "The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection". *Nucleic Acids Research*, 43(15):7217-7228, 2015

[2] S.B. Montgomery, *et al.* "The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes". *Genome Research*, 23:749-761, 2013

[3] V. Bansal. "A statistical method for the detection of variants from next-generation resequencing of DNA pools". *Bioinformatics*, 26(12):i318-i324, 2011

[4] M.A. Eberle, *et al.* "A reference dataset of 5.4 million human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". *BioRxiv*, 2016 (doi:10.1101/055541)