TECHNICAL NOTE

16GT: a fast and sensitive variant caller using a 16-genotype probabilistic model

Ruibang Luo1,2,*, Michael C. Schatz1,2 and Steven L. Salzberg1,2,3

1Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA, 2Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21218, USA and 3Departments of Biomedical Engineering and Biostatistics, Johns Hopkins University, Baltimore, MD 21218, USA

*Correspondence address: Center for Computational Biology, School of Medicine, Johns Hopkins University, 1900 E. Monument St. Rm 101B, Baltimore, MD 21205. Tel: 667-234-9641; E-mail: rluo5@jhu.edu

Abstract

16GT is a variant caller for Illumina whole-genome and whole-exome sequencing data. It uses a new 16-genotype probabilistic model to unify single nucleotide polymorphism and insertion and deletion calling in a single variant calling algorithm. In benchmark comparisons with 5 other widely used variant callers on a modern 36-core server, 16GT demonstrated improved sensitivity in calling single nucleotide polymorphisms, and it provided comparable sensitivity and accuracy for calling insertions and deletions as compared to the GATK HaplotypeCaller. 16GT is available at https://github.com/aquaskyline/16GT.

Keywords: variant calling; Bayesian model; SNP calling; indel calling

Background

Single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) that occur at a specific genome position are interdependent; i.e., evidence that elevates the probability of 1 variant type should decrease the probability of other possible variant types, and the probability of all possible alleles should sum to 1. However, widely used tools such as GATK’s UnifiedGenotyper [1] and SAMtools [2] use separate models for SNP and indel detection. The model for SNP calling in these 2 tools is nearly identical: both assume all variants are biallelic (i.e., exactly 2 haplotypes are present) and use a probabilistic model allowing for 10 genotypes: AA, AC, AG, AT, CC, CG, CT, GG, GT, TT. For indel calling, the GATK UnifiedGenotyper uses a model from the Dindel’s variant caller [3], while SAMtools’ model is from BAQ [4].

Findings

In order to detect SNPs and indels with a unified approach, we developed a new 16-genotype probabilistic model and its implementation, named 16GT. Building on an idea first introduced in Luo et al. [5], 16GT uses an empirically improved model and is the first publicly available implementation. Using X and Y to denote the indels with the highest (X) and second highest (Y) support, we add 6 new genotypes (AX, CX, GX, TX, XX, and XY) to the traditional 10-genotype probabilistic model. The 6 new
genotypes include: (i) 1 homozygous indel (XX); (ii) 1 reference allele plus 1 heterozygous indel (AX, CX, GX, TX); (iii) 1 heterozygous SNP plus 1 heterozygous indel (AX, CX, GX, TX, reusing the genotypes in ii); and (iv) 2 heterozygous indels (XY). We exclude the 5 possible combinations AY, CY, GY, TY, and YY because X has higher support than Y. By unifying SNP and indel calling in a single variant calling algorithm, 16GT not only runs 4 times faster, but also demonstrates improved sensitivity in calling SNPs and comparable sensitivity in calling indels to the GATK Haplotype Caller.

Posterior probabilities of these 16 genotypes are calculated using a Bayesian model \[ P(F|L) \propto P(L|F)P(F) \], where \( L \) is an assumed genotype. \( F \) refers to the observation of the 6 alleles (A, C, G, T, X, Y) at a given genotype position. \( P(L) \) is the prior probability of the genotype, \( P(F|L) \) is the likelihood of the observed genotype, and \( P(L|F) \) is the posterior probability of the genotype. The resulting genotype \( L_{\text{MAP}} \) is assigned to the genotype with the highest posterior probability. The distance between the highest posterior probability and the second highest posterior probability is used as a quality metric in 16GT, along with some other metrics introduced by GATK (GATK, RRID: SCR_001876) [1].

Calculating the probability of an observation \( F \) given the genotype \( L \)

To test how well an observation fits the expectation of different genotypes, we use a 2-tailed Fisher exact test \( P \) and use the resulting P-value as the goodness of fit. When calculating the likelihood of a homozygous genotype, ideally we expect 100% single allele support from the observation. For example, consider genotype “AA”;

\[
P(F|AA) = P_{\text{hom}}(F_A) \times P_s(F_C, F_G, F_T, F_X, F_Y),
\]

where \( P_s \) is the probability of an erroneous base call.

For a heterozygous genotype, 50% support is expected for each allele in the genotype; e.g., consider “CG”:

\[
P(F|CG) = P_{\text{het}}(F_C, F_G) \times P_s(F_A, F_T, F_X, F_Y),
\]

where

\[
P_{\text{hom}}(F_A) = P \left( F_A \frac{F}{1 - P_{\text{err}}} F \right).
\]

\[
P_{\text{het}}(F_C, F_G) = \prod_{i \in \{C,G\}} P \left( F_i \frac{F}{0.5 - P_{\text{err}}} F \right).
\]

\[
P_s(F_A, F_T, F_X, F_Y) = P \left( F_A + F_T + F_X + F_Y \frac{F}{P_{\text{err}} \times F} \right).
\]

\[
F_s = \sum_{i=1}^{n} f(Q_i, M_i, s) \ s \in \{A, C, G, T, X, Y\},
\]

where \( s \) is the allele type, \( n \) is the number of reads supporting allele \( s \), \( Q_i \) is the base quality, and \( M_i \) is the mapping quality. \( f \) is a function describing how \( s, Q_i, \) and \( M_i \) change the observation:

\[
\begin{align*}
\alpha &= 0 \quad \text{if} \quad M_i = 0 \\
\alpha &= 1 \quad \text{if} \quad M_i \neq 0 \\
\beta &= 0 \quad \text{if} \quad Q_i < 10 \\
\beta &= 1 \quad \text{if} \quad 10 \leq Q_i < 13 \\
\gamma &= 2 \quad \text{if} \quad 13 \leq Q_i < 17 \\
\gamma &= 3 \quad \text{if} \quad 17 \leq Q_i < 20 \\
\gamma &= 4 \quad \text{if} \quad Q_i \geq 20 \\
\epsilon &= 1 \quad \text{if} \quad s \neq \{A, C, G, T\} \\
\epsilon &= 1.375 \quad \text{if} \quad s \in \{X, Y\}
\end{align*}
\]

The possible reasons for an observation that does not match the reference genome are (i) a true variant; (ii) an error generated in library construction; (iii) a base calling error; (iv) a mapping error; and (v) an error in the reference genome. Reasons (iii) and (iv) are explicitly captured in our model. For reasons (ii) and (v), we include 2 error probabilities, \( P_t \) for SNP error and \( P_i \) for indel error. We define \( P_{\text{err}} = P_t + P_i \), where \( P_t \) and \( P_i \) are set to 0.01 and 0.005, respectively. These 2 values were set empirically based on the observation that SNP errors are more common than indel errors in library construction and in the reference genome.

In addition, most short read aligners use a dynamic programming algorithm to enable gapped alignment, using a scoring scheme that usually penalizes gap opening and extension more than mismatch. Consequently, authentic gaps that occur at an end of a read are more likely to be substituted by a set of false SNPs or alternatively to get trimmed or clipped. Thus, we applied a coefficient \( \gamma \) to weight indel observations more than SNPs in order to increase the sensitivity on indels.

Calculating the probability of the genotype \( L \)

Given (i) a known rate of single nucleotide differences between 2 unrelated haplotypes; (ii) a known rate of single indel differences between 2 unrelated haplotypes; and (iii) a known Transitions to Transversions ratio (Ti/Tv), the 16GT model’s prior probabilities are calculated as shown in Table 1.

Given (i) a known rate \( \theta \) of single nucleotide differences between 2 unrelated haplotypes; (ii) a known rate \( \omega \) of single indel differences between 2 unrelated haplotypes; and (iii) a known Ti/Tv \( \epsilon \), transition is expected to occur more frequently than transversion under selective pressure. The default known rates for human genome are \( \theta = 0.001, \omega = 0.0001, \epsilon = 2.1 \), where \( \epsilon \) is set to the value for human and needs to be changed for other species.

Results

We benchmarked 16GT with GATK UnifiedGenotyper, GATK HaplotypeCaller (GATK, RRID: SCR_001876) [1], Freebayes (FreeBayes, RRID: SCR_00761) [6], Fermikit [7], ISAAC (Isaac, RRID: SCR_012772) [8], and VarScan2 [9] using a set of very high-confidence variants developed by the Genome in a Bottle project for genome NA12878 (Coriell Cat# GM12878, RRID: CVCL_7526; version 2.19) (Additional File 1: Supplementary Note) [10]. The results are shown in Table 2 and as receiver operating characteristic curves in Supplementary Fig. S1.

For SNPs, 16GT produced the most true positive calls and the fewest false negative calls; i.e., it has the highest sensitiv-
Table 1: P(L), Genotype prior probabilities for a reference allele “A”.

<table>
<thead>
<tr>
<th>L</th>
<th>Zygosity</th>
<th>Number of SNPs</th>
<th>Number of indels</th>
<th>Number of transversions</th>
<th>Prior probability P(L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Hom.</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>GG</td>
<td>Hom.</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>θ/2 * ε * ε</td>
</tr>
<tr>
<td>CC, TT</td>
<td>Hom.</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>θ/2</td>
</tr>
<tr>
<td>AG</td>
<td>Het.</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>θ * ε</td>
</tr>
<tr>
<td>AC, AT</td>
<td>Het.</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>θ</td>
</tr>
<tr>
<td>CG, GT</td>
<td>Het.</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>θ * θ/2 * ε</td>
</tr>
<tr>
<td>CT</td>
<td>Het.</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>θ * θ/2</td>
</tr>
<tr>
<td>AX</td>
<td>Het.</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>ω</td>
</tr>
<tr>
<td>GX</td>
<td>Het.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>ω * θ/2 * ε</td>
</tr>
<tr>
<td>CX, TX</td>
<td>Het.</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>ω * θ/2</td>
</tr>
<tr>
<td>XX</td>
<td>Hom.</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>ω/2</td>
</tr>
<tr>
<td>XY</td>
<td>Het.</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>ω * ω/2</td>
</tr>
</tbody>
</table>

Hom.: homozygous; Het.: heterozygous.

Table 2: Benchmark comparisons between 16GT and 5 other variant callers on a dataset from the Genome in a Bottle project consisting of 787M read pairs (53-fold) from genome NA12878.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Time (minutes w/36 cores)</th>
<th>FP</th>
<th>TP</th>
<th>Total</th>
<th>dbSNP</th>
<th>dbSNP</th>
<th>TP in Omni 2.5</th>
<th>FN</th>
<th>FP</th>
<th>TP</th>
<th>Total</th>
<th>dbSNP</th>
<th>dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>16GT</td>
<td>121</td>
<td>2663179</td>
<td>5346</td>
<td>4220</td>
<td>79%</td>
<td>20/20</td>
<td>918</td>
<td>167549</td>
<td>1462</td>
<td>944</td>
<td>65%</td>
<td>3180</td>
<td></td>
</tr>
<tr>
<td>UG</td>
<td>29</td>
<td>2655608</td>
<td>1639</td>
<td>563</td>
<td>34%</td>
<td>15/15</td>
<td>8489</td>
<td>163839</td>
<td>624</td>
<td>546</td>
<td>88%</td>
<td>6890</td>
<td></td>
</tr>
<tr>
<td>HC</td>
<td>539</td>
<td>2653684</td>
<td>419</td>
<td>143</td>
<td>34%</td>
<td>4/4</td>
<td>10413</td>
<td>168444</td>
<td>1232</td>
<td>726</td>
<td>59%</td>
<td>2285</td>
<td></td>
</tr>
<tr>
<td>Freebayes</td>
<td>52</td>
<td>2655513</td>
<td>724</td>
<td>353</td>
<td>49%</td>
<td>11/14</td>
<td>8584</td>
<td>162505</td>
<td>559</td>
<td>0</td>
<td>0%</td>
<td>2224</td>
<td></td>
</tr>
<tr>
<td>Fermikit</td>
<td>45</td>
<td>2567672</td>
<td>2036</td>
<td>509</td>
<td>25%</td>
<td>9/9</td>
<td>96425</td>
<td>161916</td>
<td>1996</td>
<td>1076</td>
<td>54%</td>
<td>8813</td>
<td></td>
</tr>
<tr>
<td>ISAAC</td>
<td>63</td>
<td>2654438</td>
<td>1115</td>
<td>586</td>
<td>53%</td>
<td>15/15</td>
<td>4659</td>
<td>158642</td>
<td>1239</td>
<td>710</td>
<td>57%</td>
<td>12087</td>
<td></td>
</tr>
<tr>
<td>VarScan2</td>
<td>136</td>
<td>2658358</td>
<td>1680</td>
<td>718</td>
<td>43%</td>
<td>10/10</td>
<td>5739</td>
<td>158906</td>
<td>574</td>
<td>481</td>
<td>84%</td>
<td>11823</td>
<td></td>
</tr>
</tbody>
</table>

FP: false positive; FN: false negative; HC: GATK HaplotypeCaller; UG: GATK UnifiedGenotyper.

Conclusions

16GT is the firstly publicly available implementation using a 16-genotype probabilistic model for variant calling. Compared with local assembly based variant callers, 16GT provides better sensitivity in SNP calling and comparable sensitivity in indel calling. In the current implementation, 16GT can only be applied to germline variant detection. In the future, we will enhance 16GT to support multi-sample variant calling and GVCF output and to support somatic variant detection and extend the model to support variant calling in species with more than 2 haplotypes.

Additional files

Additional File 1.docx

Abbreviations

indel: insertions and deletions; SNP: single nucleotide polymorphism; Ti/Tv: Transitions to Transversions.
Acknowledgements

We thank United Electronics Co. Limited for providing code samples for the bam2snapshot function.

Funding

This work was supported by the US National Institutes of Health under grants R01-HL129239 and R01-HG006677.

Availability of source code and requirements

Project name: 16GT
Project homepage: https://github.com/aquaskyline/16GT
Archived version: https://github.com/aquaskyline/16GT/releases/tag/1.0
Operating system: Platform independent
Programming language: C++ and Perl
Other requirements: See GitHub page
License: GPLv3
Any restrictions to use by non-academics: None

Availability of supporting data and materials

Snapshots of the code and data are available in the GigaScience repository, GigaDB [12], and are also available via the CodeOcean reproducibility platform [13].

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

R.L., M.C.S., and S.L.S. conceived the study. R.L. developed and implemented the 16GT algorithm and benchmarked 16GT with other variant callers. R.L., M.C.S., and S.L.S. wrote the paper. All authors have read and approved the final version of the manuscript.

References