Title

MultiGeMS: detection of SNVs from multiple samples using model selection on high-throughput sequencing data

Journal

Bioinformatics, 32(10)

ISSN

1367-4803

Authors

Murillo, Gabriel H
You, Na
Su, Xiaoquan
et al.

Publication Date

2016-05-15

DOI

10.1093/bioinformatics/btv753

Peer reviewed

Bioinformatics, 32(10), 2016, 1486–1492
doi: 10.1093/bioinformatics/btv753
Advance Access Publication Date: 18 January 2016
Original Paper

OXFORD

Genetics and population analysis

MultiGeMS: detection of SNVs from multiple samples using model selection on high-throughput sequencing data

Gabriel H. Murillo[1], Na You[2], Xiaoquan Su[3], Wei Cui[1], Muredach P. Reilly[4], Mingyao Li[5], Kang Ning[6] and Xinping Cui[1,7,*]

[1]Department of Statistics, University of California, Riverside, CA 92521, USA, [2]Department of Statistical Science, School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, Guangdong 510275, China, [3]Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, Shandong 266101, China, [4]Cardiovascular Institute, [5]Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA, [6]Key Laboratory of Molecular Biophysics of the Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China and [7]Center for Plant Cell Biology, Institute for Integrative Genome Biology, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on 16 July 2015; revised on 19 December 2015; accepted on 21 December 2015

Abstract

Motivation: Single nucleotide variant (SNV) detection procedures are being utilized as never before to analyze the recent abundance of high-throughput DNA sequencing data, both on single and multiple sample datasets. Building on previously published work with the single sample SNV caller genotype model selection (GeMS), a multiple sample version of GeMS (MultiGeMS) is introduced. Unlike other popular multiple sample SNV callers, the MultiGeMS statistical model accounts for enzymatic substitution sequencing errors. It also addresses the multiple testing problem endemic to multiple sample SNV calling and utilizes high performance computing (HPC) techniques.

Results: A simulation study demonstrates that MultiGeMS ranks highest in precision among a selection of popular multiple sample SNV callers, while showing exceptional recall in calling common SNVs. Further, both simulation studies and real data analyses indicate that MultiGeMS is robust to low-quality data. We also demonstrate that accounting for enzymatic substitution sequencing errors not only improves SNV call precision at low mapping quality regions, but also improves recall at reference allele-dominated sites with high mapping quality.

Availability and implementation: The MultiGeMS package can be downloaded from https://github.com/cui-lab/multigems.

Contact: xinping.cui@ucr.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, a precipitous fall in the costs of DNA sequencing has led to a similarly precipitous rise in the amount of available DNA sequencing data. Using such data, projects such as the 1000 Genomes Project (Consortium, 2012) and the International HapMap Project (The International HapMap 3 Consortium, 2010) have done much to catalog human genetic variation. Such projects have put an emphasis on sequencing multiple samples, as it has been demonstrated that

given a fixed sequencing budget, the total number of population variants identified can often be increased by decreasing the coverage and increasing the number of samples sequenced (Le and Durbin, 2011). This is, in fact, a motivation for the 1000 Genomes Project low-coverage pilot project (1000 Genomes Project Consortium *et al.*, 2010).

As single nucleotide variants (SNVs) are the most common type of sequence variation, many data analysis tools have been created to detect SNVs from DNA sequencing data. However, the prevalent single sample SNV detection procedure has proved insufficient with the availability of data from multiple samples.

There are generally two possibilities for SNV detection using single sample SNV callers on multiple sample data. First, it is possible to pool the multiple sample data, and run single sample SNV callers on all the data as if it were one sample. This can, however, reduce the number of legitimate SNV calls, as rare SNVs are often dismissed as errors. Also, pooled data genotype calls may not be indicative of the true sample genotypes. For example, if half of the samples are homozygous reference and the other half are homozygous variant, then the pooled genotype call will likely be heterozygous, even when none of the samples are such.

Second, single sample SNV callers can be run on all the samples independently and the output SNV call lists can be combined. This procedure may be useful if the analysis objective is to find variants unique to individual samples. However, for the most part, this procedure will greatly inflate the false positive rate.

A common solution to these issues is the development of multiple sample SNV callers. These procedures utilize data from multiple samples to identify both common and rare SNVs, as well as SNVs which may have limited support across multiple samples. The amount of SNVs called would generally be between that of the two possibilities mentioned above. That is, the goal of multiple sample SNV callers is to have a good balance of precision and recall with respect to all the samples analyzed.

There are, however, concerns with currently available multiple sample SNV callers. A primary concern is how to utilize all important and available information to accurately call SNVs from multiple samples. Other concerns include the resulting multiple testing problem that arises, computational performance and robustness to low-quality sequencing and alignment data.

All of these concerns are addressed with the multiple sample genotype model selection (MultiGeMS) SNV caller, which will here be introduced. MultiGeMS builds on the previously published single sample SNV caller GeMS (You *et al.*, 2012), and estimates sample genotypes and genotype probabilities for possible SNV sites. Unlike other popular multiple sample SNV callers, the MultiGeMS statistical model accounts for enzymatic substitution sequencing errors. Also, in consideration of the multiple testing problem associated with SNV calling, SNVs are called using a local false discovery rate (lFDR) estimator. Further, MultiGeMS utilizes high performance computing (HPC) techniques for computational efficiency and is robust to low-quality sequencing and alignment data.

2 Methods

The MultiGeMS procedure is motivated by a consideration of the sequencing by synthesis procedure used by the Illumina HTS platform. Before the well-known stages of base-calling, alignment and SNV calling, there are many steps leading up to sequencing by synthesis. These steps begin with the acquisition and fragmentation of the DNA samples to be sequenced, and then include various procedures involving enzymatic binding of complimentary bases. These procedures include fragment end repair, PCR amplification, bridge amplification and the pre-base-calling sequencing by synthesis. Each time bases are enzymatically bound together, a base substitution error is possible. For example, the non-complimentary bases A and G may bind. Likewise, C and T may bind in error. When this happens, a base sequenced by the sequencing by synthesis process may be in conflict with the alleles of the underlying genotype, leading to downstream problems in base-calling, alignment and SNV calling. While recent advances in Illumina platform sequencing and variant calling pipelines have mitigated many types of sequencing errors, the MultiGeMS procedure, as a multiple sample SNV caller accounting for such enzymatic substitution sequencing errors, represents a substantial contribution to variant calling (see Supplementary Materials Section 1).

Table 1. MultiGeMS notation

Notation	Explanation
$i \in \{1, \ldots, s\}$	sample index
$\{X_{ij}\}$	n_i alleles consisting of reference alleles (R) and 'most prevalent non-reference' alleles (N); other alleles discarded
$\{G_i\}$	unobserved genotypes $g \in \{RR, RN, NN\}$, $P(G_i = g) = p_g$
Y_{ij}	unobserved 'sequencing' allele associated with X_{ij}
q_k^g	$P(Y_{ij} = k \mid G_i = g, \mu_{RN}, \mu_{NR})$ where $k \in \{R, N\}$
μ_{RN}, μ_{NR}	'sequencing' allele (Y_{ij}) mutation rates simultaneously estimated over (0,0.5] from all samples, where $\mu_{RN} = P(Y_{ij} = N \mid G_i = RR)$ and $\mu_{NR} = P(Y_{ij} = R \mid G_i = NN)$
w_{ij}	'weight' of X_{ij}, based on base-calling and alignment scores
D_{ij}	$\{X_{ij}, w_{ij}\}$
D_i	$\{D_{ij} : j \in \{1, \ldots, n_i\}\}$, the observed allele data
$\boldsymbol{\theta}$	set of 5 parameters: $\{p_g\}$, μ_{RN}, μ_{NR}
$e_{i,g}$	$P(G_i = g \mid D_i, \boldsymbol{\theta})$

Given the assumption that each site is independent, only one site is considered at a time, and hence a site index variable is not necessary.

The MultiGeMS procedure requires a FASTA genome reference file and corresponding Illumina HTS platform alignment data in the form of PILEUP files for each sample. Hence, sites on the reference file with a corresponding allele pileup are associated with their reference base R, observed alleles aligned to the site, and associated base-calling and alignment quality scores. MultiGeMS assumes alignment data from a diploid organism, and thus, in theory, a maximum of only two alleles should be observed in the allele pileup at a site of any sample. MultiGeMS further assumes that one of these alleles is R, while the other is the site's most prevalent non-reference allele N, which is the observed non-reference allele mode taken after pooling the site's sample allele pileups (justification of this assumption and commentary on triallelic sites is provided in Section 2 of the Supplementary Materials). At each site, only R and N are considered. All other alleles in a pileup are discarded Table 1 summarizes the notation used by MultiGeMS as expressed below.

Assuming s samples, let $X_{ij} \in \{R, N\}$ denote the j^{th} ($j \in \{1, 2, \ldots, n_i\}$) observed allele from sample i at a particular site on the reference genome. Given a sample i, all the X_{ij} obtained from the aligned HTS reads at a given genomic site are independent. Each site is also assumed to be independent and all the samples are assumed to be independent. Further, let $Y_{ij} \in \{R, N\}$ be defined as the unobserved 'sequencing' allele associated with X_{ij}, that is, the base that is sequenced in the sequencing by synthesis procedure. Finally, let G_i be the unobserved genotype associated with sample i,

Table 2. q_k^g, given for $k \in \{R, N\}$ and $g \in \{RR, RN, NN\}$, as expressed as a function of $\{\mu_{RN}, \mu_{NR}\}$

	$k = R$	$k = N$
$g = RR$	$1 - \mu_{RN}$	μ_{RN}
$g = RN$	$\frac{1}{2}(1 - \mu_{RN}) + \frac{1}{2}\mu_{NR}$	$\frac{1}{2}\mu_{RN} + \frac{1}{2}(1 - \mu_{NR})$
$g = NN$	μ_{NR}	$1 - \mu_{NR}$

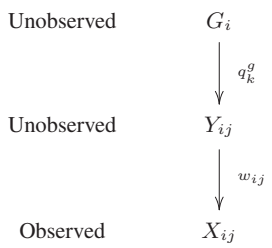

Fig. 1. The relationship between G_i, Y_{ij} and X_{ij}. The genotype G_i associated with sample i at a particular site is unobserved. As is probabilistically modeled by q_k^g, the unobserved 'sequencing' allele, Y_{ij}, depends on G_i. Finally, the observed allele, X_{ij}, depends on Y_{ij} and w_{ij}

which, in the MultiGeMS framework, can be either RR, RN or NN. $P(G_i = g)$, the probability of genotype g, is notated by p_g.

Revisiting the enzymatic base substitution errors described above, let μ_{RN} be the probability that $Y_{ij} = N$, assuming the genotype $G_i = RR$. That is, because of enzymatic base substitution errors, the 'sequencing' allele may not be representative of the genotype. Likewise, let $\mu_{NR} = P(Y_{ij} = R|G_i = NN)$. The MultiGeMS procedure estimates these 'sequencing' allele mutation rates. The probability expressions of $Y_{ij}|G_i$ are derived in Section 3 of the Supplementary Materials and are listed in Table 2, where

$$q_k^g = P(Y_{ij} = k|G_i = g, \mu_{RN}, \mu_{NR}). \quad (1)$$

MultiGeMS also utilizes the Phred-scaled base-calling (B_{ij}) and alignment (M_{ij}) quality values for each observed allele (X_{ij}). Specifically, the minimum of an aligned allele's P(Correct Base-Call) and P(Correct Alignment) is taken as the accuracy or weight of that aligned allele. Given the Phred quality scoring scheme, this weight is given as follows.

$$\begin{aligned} w_{ij} &= \min\{P(\text{Correct Base-Call}), P(\text{Correct Alignment})\} \\ &= 1 - 10^{-0.1\min\{B_{ij}, M_{ij}\}} \end{aligned} \quad (2)$$

Collectively, the observed allele data, $\{X_{ij}, w_{ij}\}$, is notated D_{ij}.

Given a correct base-call and alignment, we can assume that $X_{ij} = Y_{ij}$. We, therefore, propose the following probability distribution for $X_{ij}|Y_{ij}$.

$$\begin{aligned} P(X_{ij} = Y_{ij}|Y_{ij}) &= w_{ij} \\ P(X_{ij} \neq Y_{ij}|Y_{ij}) &= 1 - w_{ij} \end{aligned} \quad (3)$$

Figure 1 demonstrates the relationship between G_i, Y_{ij} and X_{ij}.

The conditional likelihood of the data given the genotype is given by the following, where $I(\cdot)$ represents the indicator function.

$$\begin{aligned} &P(D_i|G_i = g, \mu_{RN}, \mu_{NR}) \\ &= \prod_{j=1}^{n_i} P(D_{ij}|G_i = g, \mu_{RN}, \mu_{NR}) \\ &= \prod_{j=1}^{n_i} \sum_{k \in \{R,N\}} [P(X_{ij}, w_{ij}|Y_{ij} = k)]P(Y_{ij} = k|G_i = g, \mu_{RN}, \mu_{NR}) \\ &= \prod_{j=1}^{n_i} \sum_{k \in \{R,N\}} [w_{ij}^{I(X_{ij}=k)}(1 - w_{ij})^{I(X_{ij}\neq k)}]q_k^g \end{aligned} \quad (4)$$

Further, assuming that the samples are independent, the complete log-likelihood with parameter $\theta = \{p_{RR}, p_{RN}, p_{NN}, \mu_{RN}, \mu_{NR}\}$ and unobserved $\boldsymbol{G} = (G_1, \ldots, G_s)$ is as follows.

$$\begin{aligned} &l(\boldsymbol{\theta}|\boldsymbol{D}, \boldsymbol{G}) \\ &= \log P(\boldsymbol{D}|\boldsymbol{G}, \boldsymbol{\theta})\mathrm{P}(\boldsymbol{G}|\boldsymbol{\theta}) \\ &= \log \prod_{i=1}^{s} \prod_{g} [P(D_i|G_i = g, \mu_{RN}, \mu_{NR})P(G_i = g|p_g)]^{I(G_i=g)} \\ &= \sum_{i=1}^{s} \sum_{g} I(G_i = g)\{\log P(D_i|G_i = g, \mu_{RN}, \mu_{NR}) + \log p_g\} \end{aligned} \quad (5)$$

The MultiGeMS procedure begins by identifying sites that could possibly be a SNV. Then, using the above complete log-likelihood, it runs an EM algorithm estimation procedure (see Supplementary Materials Section 3) on said sites to accurately genotype both these sites and the multiple samples at these sites. This multiple sample genotyping process, or multi-sample genotype model selection, involves calculating the probabilities for the genotypes RR, RN and NN, and then selecting the genotype with the largest associated probability. Further, using the local false discovery rate (lFDR) result in Muralidharan *et al.* (2012), MultiGeMS then addresses the multiple testing problem endemic to SNV detection by calling SNVs if $\widehat{\text{lFDR}} \leq 0.1$ (see Supplementary Materials Section 4). $\widehat{\text{lFDR}}$ is given below, where c_i are the coverage weights, $e_{i,RR} = P(G_i = RR|D_i, \boldsymbol{\theta})$ and the asterisk $(*)$ indicates final EM iteration values.

$$\widehat{\text{lFDR}} = \exp\left[\frac{\sum_i c_i \log(e_{i,RR}^*)}{\sum_i c_i}\right] \quad (6)$$

$\widehat{\text{lFDR}}$ is a coverage-weighted geometric mean of $\{e_{i,RR}^*\}$, where each $e_{i,RR}^*$ is dependent on the estimated 'sequencing' allele mutation rates $(\hat{\mu}_{RN}, \hat{\mu}_{NR})$, the observed alleles $(\{X_{ij}\})$ and the quality weights $(\{w_{ij}\})$. The effect of estimating the 'sequencing' allele mutation rates is to increase SNV calling power at sites dominated by R alleles (see Section 3.2). In particular, assuming high base-calling and mapping quality data, a non-zero estimate of $\mu_{NR} = P(Y_{ij} = R|G_i = NN)$ allows for the possibility of an enzymatic base substitution error at bases observed as R, increasing the likelihood of a SNV call determination. Alternatively at such sites, when μ_{NR} is estimated to be near 0, MultiGeMS is more conservative in its SNV calling, determining only those sites with extensive evidence of a variant.

3 Results

3.1 Simulation study

To validate the MultiGeMS procedure, we conducted a simulation of 10 samples of HTS data from the roughly 2.5 Mbp *Thermoanaerobacter sp. X514* reference. Each sample was simulated at an average of 50× coverage using DWGSIM (http://sourceforge.net/projects/dnaa/). The data were simulated with three levels of SNVs: 'population' (the set of simulated variants which are present in the reads of all samples), 'group' (the set of simulated variants which are present in the reads of a certain subset of samples) and 'individual' (the set of simulated variants which are present in

the reads of only a single sample) SNVs. All samples were simulated with the population SNVs at a rate of 0.625 per 1000 bp. The group SNVs were simulated at a rate of 0.3125 per 1000 bp, with half the samples simulated with group 'A' SNVs and the other half simulated with group 'B' SNVs. Finally, each sample was simulated with individual sample SNVs at a rate of 0.0625 per 1000 bp. Hence, the overall SNV rate in every sample was 1 per 1000 bp and the total number of simulated population, group and individual sample SNVs were kept roughly the same: 1536, 1536 and 1541, respectively. Also, it is noteworthy that the DWGSIM option `-Q 30` was utilized to simulate a large standard deviation in the base-calling quality scores, which can lead to downstream alignment and variant calling difficulties. The full details of the simulation study are given in Table 3 of the Supplementary Materials.

MultiGeMS and four popular multiple sample SNV detection procedures were run on the BWA (Li and Durbin, 2009) aligned simulation data: FreeBayes (Garrison and Marth, 2012), GATK (DePristo *et al.*, 2011; McKenna *et al.*, 2010), SAMtools (Li *et al.*, 2009) and VarScan (Koboldt *et al.*, 2012). For comparison, the two methods of calling SNVs on multiple samples using a single sample SNV caller, as described in Section 1, are demonstrated using single sample GeMS (You *et al.*, 2012). 'GeMS' represents the results from running GeMS on each sample and then combining the resulting SNV call lists. 'GeMS-pooled' represents the results from pooling all the sample alignment files into one alignment file using BamTools (Barnett *et al.*, 2011), and then running GeMS on this file. FreeBayes, GATK, SAMtools and VarScan are also capable of single sample SNV detection, hence these SNV callers can also be implemented with the aforementioned pooling methods. The overall results, presented with each SNV caller's recall, precision and *F-score* (the harmonic mean of precision and recall), are given in Table 3. The 'CPU Time (min)' and 'Max Memory (MB)' columns of Table 3 will be discussed in Section 3.3. While FreeBayes, GATK, SAMtools and VarScan are able to detect variants other than SNVs, for example insertions and deletions, the release version of MultiGeMS as described herein is limited to SNV detection. Hence only the SNV call output of all the tested procedures were compared, without regard to whether these procedures also, by default, detected other types of variants (which may have impacted needed computational resources and the SNV call results themselves). See Section 4 for further details on the detection of indels and other variant types.

As shown in Table 3, the SNV callers were conservative in general, other than GATK and GeMS. GATK was relatively aggressive, favoring recall over precision and even calling over 900 more sites than the total simulated SNV count. Likewise, the GeMS results demonstrate the issues occuring when single sample SNV call lists are combined. Despite having the greatest overall recall, GeMS has the second lowest *F-score* because of its poor precision. Hence, because of the resulting inflated false positive rate, the GeMS method is not recommended, unless perhaps, when searching for rare SNVs.

The best balance of precision and recall, as measured by the *F-score*, can be seen in the performance of FreeBayes, with MultiGeMS as a close second. Each of the other SNV calling results, such as those from GATK or SAMtools, show a greater relative weakness in either precision or recall, but not both. Thus, because of the large standard deviation of the simulated base-calling quality scores, we can see that FreeBayes and MultiGeMS are robust to low-quality sequencing data. Of special note is the perfect precision of MultiGeMS. The only other SNV caller with a precision of 1 is VarScan, but it also has the lowest recall and *F-score* values.

The recall results, in Table 4, help us to understand more about the temperament of each SNV caller with respect to the simulated population, group and individual SNVs. For example, all the SNV callers, other than SAMtools, demonstrate better recall for the population SNVs compared to the individual SNVs. This behavior is expected as the population SNV sites would have a higher variant signal than individual SNV sites. In general, a relatively large number of variant genotype samples would provide a higher variant signal, and hence, such a site is more likely to be called a SNV (naturally, multiple sample SNV caller VCF output indicates the genotype calls for each sample). Also as expected is the extremely low individual SNV recall of GeMS-pooled. Since the alignment data from the 10 samples were pooled into one file, any variant simulated in just one of the samples would have very low variant signal. Hence, the GeMS-pooled method is not recommended for multiple sample data, unless perhaps, when searching for common SNVs. In contrast to the GeMS-pooled results, the low group and individual SNV recall results of VarScan does not result from the data being pooled, as the output VCF file contains columns for each of the 10 samples.

Table 4. Simulation study recall results, by level of SNV

	Count	Population	Group	Individual
GeMS	7843	0.9818	0.9831	0.9676
GATK	5574	0.9792	0.9798	0.9676
FreeBayes	4476	0.9596	0.9661	0.9513
MultiGeMS	4141	0.9772	0.9805	0.7398
SAMtools	3619	0.4805	0.9785	0.8027
GeMS-pooled	3166	0.9818	0.9831	0.0019
VarScan	2013	0.9245	0.3730	0.0143

The results are sorted in descending order of the SNV count.

Table 3. Simulation study results

	Count	Recall	Precision	*F*-score	CPU Time (min)	Max Memory (MB)
FreeBayes	4476	0.9596	0.9877	0.9735	80.3	204
MultiGeMS	4141	0.8988	1.0000	0.9467	17.5	24
GATK	5574	0.9761	0.8068	0.8834	11.6	1211
SAMtools	3619	0.7541	0.9599	0.8446	29.5	39
GeMS-pooled	3166	0.6544	0.9523	0.7758	10.8	416
GeMS	7843	0.9774	0.5741	0.7234	16.7	417
VarScan	2013	0.4369	1.0000	0.6082	45.6	993

The results are sorted in descending order of *F-score*, which is the harmonic mean of the SNV caller precision and recall. Total CPU time in minutes and maximum procedure memory in megabytes (MB) utilized by the SNV callers running on a single thread for the simulation study analysis are provided in the two far right columns, respectively.

Of note is the exceptional population and superior group recall values of MultiGeMS compared to the other multiple sample SNV callers. However, the reason for FreeBayes' superior overall *F-score* in Table 3, is the lower individual SNV recall of MultiGeMS. When greater individual SNV recall is required, a larger lFDR threshold can be used, though this is not generally recommended (see Supplementary Materials Section 4). The results of various lFDR threshold values can be seen in Table 5. Though it is notable that the MultiGeMS *F-score* at the lFDR threshold of 0.4 is slightly greater than that of FreeBayes, we acknowledge that the results of any SNV caller can be filtered in various ways to suit the user's objectives. Table 4 of the Supplementary Materials provides the MultiGeMS simulation study recall results by lFDR threshold and level of SNV. There we see that increasing the lFDR threshold greatly increases the individual SNV recall results, without much affecting the recall results of the population or group SNVs.

As we can see, MultiGeMS demonstrates perfect precision along with robustness to low-quality sequencing data. It also boasts exceptional recall of simulated population and group SNVs. However, care is needed when using MultiGeMS on multiple sample data known to harbor individual or rare SNVs.

3.2 Human data analysis

In addition to simulated data, the MultiGeMS procedure was also validated using data from 10 human samples. The 10 samples are NA12878 and NA12877, and 8 of their offspring: NA12879, NA12880, NA12882, NA12883, NA12885, NA12886, NA12888 and NA12893. These data were retrieved from 'WGS sequencing BAMs for the entire pedigree produced by Illumina as part of their Platinum project', additionally a 'gold standard variant dataset' is available from the Genome in a Bottle Consortium (https://sites.stanford.edu/abms/content/availability-phase-consistent-gold-standard-variant-set-na12878-and-rtgtools-software). As many SNV callers, MultiGeMS treats all samples, such as the 10 samples from the aforementioned family, independently. However, work has been done to model family relationships in variant calling, for example, see Li *et al.* (2012). The full details of this human data analysis are given in Table 6 of the Supplementary Materials. Also, Section 5 of the Supplementary Materials considers the addition of an unrelated sample to this human data analysis.

The aforementioned SNV callers were run on the data isolated to chromosomes 18-22. As shown in Table 6, MultiGeMS displays good precision and recall, but is not top-ranked in terms of these metrics or *F-score*. Also, as in Section 3.1, we see the superiority of MultiGeMS to GeMS and GeMS-pooled.

As indicated in Section 2, an accounting of enzymatic substitution sequencing errors provided motivation for the MultiGeMS SNV caller. It is evident that bases with such errors may be high in base-calling quality, and the corresponding read mapping quality will also be high if only a small number of mismatches occurs. For samples at a given genomic site where R is the dominant allele, the estimated μ_{NR} is often larger than the estimated μ_{RN}, suggesting that the R allele is more likely to be perceived as an enzymatic base substitution error than the N allele. This would increase the likelihood for an RN or NN genotype call and therefore a SNV call. If μ_{RN} and μ_{NR} are assumed to be zero, as is the case with many other SNV callers, a SNV call will be most likely missed at such sites.

On the other hand, at a site where N is the dominant allele, the estimated μ_{NR} is often smaller than the estimated μ_{RN}, suggesting that the R allele is less likely to be perceived as an enzymatic base substitution error than the N allele. This would increase the likelihood for an RN genotype call and therefore lead to a SNV call. If μ_{RN} and μ_{NR} are assumed to be zero, as is the case with many other SNV callers, the likelihood of the NN genotype call will be high which will still lead to a SNV call. Therefore, as shown in Table 7, we see the increased SNV calls at R allele-dominated sites and only a slight increase in SNV calls at N allele-dominated sites.

Sometimes bases with enzymatic substitution sequencing errors may be high in base-calling quality but low in read mapping quality if many mismatches occur. Although mapping-quality-based SNV detection methods will most likely not make a SNV call if the mapping quality score is low, MultiGeMS takes extra caution in making SNV calls in this situation. For the cases such as $RR \to N \to N$ ($G_i \to Y_{ij} \to X_{ij}$, as in Fig. 1) with low mapping quality scores, if μ_{RN} is not very close to 0, the low w_{ij} and high estimated μ_{RN} will inhibit a MultiGeMS SNV call. This possibly explains the marginal better performance of MultiGeMS in low-mapping quality regions as shown below.

Consider the union of the sites called by the multiple sample SNV callers and those sites validated as SNVs on chromosome 22 (the analyzed chromosome with the lowest average read mapping quality).We can divide these sites by average mapping quality quartiles and compute the *F-score* for each SNV caller. Let us call the set of such sites with average mapping quality between the minimum and the first quartile as 'quarter 1', the set of such sites with average mapping quality between the first quartile and the median as 'quarter 2', and so on. As shown in Table 8, MultiGeMS is the top-ranked SNV caller, in terms of *F-score*, in quarter 1. That is, MultiGeMS has the advantage at low mapping quality sites.

Table 5. MultiGeMS simulation study results

	Count	Recall	Precision	*F*-score
0.1	4141	0.8988	1.0000	0.9467
0.2	4308	0.9347	0.9995	0.9660
0.3	4438	0.9538	0.9901	0.9716
0.4	4510	0.9635	0.9843	0.9738
0.5	4638	0.9698	0.9633	0.9666

The results are sorted by the lFDR threshold values in the first column.

Table 6. Human chromosomes 18–22 data analysis results

	Count	Recall	Precision	*F*-score
SAMtools	522 597	0.9858	0.8544	0.9154
VarScan	519 860	0.9766	0.8509	0.9094
MultiGeMS	543 457	0.9874	0.8229	0.8977
FreeBayes	552 472	0.9690	0.7944	0.8730
GeMS-pooled	499 267	0.8390	0.7612	0.7982
GeMS	715 224	0.9858	0.6243	0.7645
GATK	799 770	0.9954	0.5637	0.7198

The results are sorted in descending order of *F-score*.

Table 7. SNV call counts for two versions of MultiGeMS, the regular version (second column) and the $\boldsymbol{\mu \equiv 0}$ version (third column), by sites of various *R* allele proportion values

R Proportion	Count	Count ($\boldsymbol{\mu \equiv 0}$)
$[0.75, 1]$	136 993	192
$[0.5, 0.75)$	173 011	48758
$[0.25, 0.5)$	103 626	99466
$[0, 0.25)$	129 827	128688

Table 8. Human chromosome 22 data analysis *F-score* results divided by mapping quality quarters

	Quarter 1	Quarter 2	Quarter 3	Quarter 4
MultiGeMS	0.5127 (9319)	0.7747 (10 137)	0.9550 (23 976)	0.9772 (27 753)
SAMtools	0.4971 (9656)	0.8099 (9163)	0.9613 (23 610)	0.9784 (27 696)
VarScan	0.4929 (9254)	0.7912 (9426)	0.9589 (23 335)	0.9782 (27 390)
FreeBayes	0.4073 (13 284)	0.6614 (11 874)	0.9494 (23 141)	0.9777 (27 165)
GeMS-pooled	0.3906 (12 046)	0.6766 (11 173)	0.9230 (22 209)	0.9432 (25 856)
GeMS	0.2944 (22 607)	0.4803 (20 800)	0.9266 (25 490)	0.9764 (27 818)
GATK	0.2569 (26 829)	0.3872 (27 563)	0.8830 (27 921)	0.9717 (28 111)

The *F-score* and SNV count (in parentheses) are provided for each SNV caller and mapping quality quarter. The results are sorted in descending order of the 'Quarter 1' *F-score* values.

In considering the performance of the SNV callers with rare SNVs, HTS alignment and validated variant data from one unrelated human sample can be analyzed along with the 10 related human samples as described above. The unrelated sample, identified as 1245 (data available upon request), was sequenced as part of a clinical trial and is not distinguished with any unusual genetic or data characteristics. The results for chromosomes 18–22 are provided in Table 9. Compared with Table 6, we see that other than the GeMS-pooled precision, the SNV caller performance metrics have all decreased. This indicates that the unrelated sample data is noisy relative to the data of the 10 related samples. It is of interest to note, however, that the relative ranking of MultiGeMS, with respect to the *F-score*, has increased from the third to the first position. See Section 5 of the Supplementary Materials for more details.

3.3 Computational performance

Both the CPU time in minutes and the memory usage in megabytes (MB) were recorded for the tested SNV callers for both the simulation study in Section 3.1 and the human chromosome 22 analysis in Section 3.2. The computational resources used during the data preprocessing necessitated by GeMS, GeMS-pooled, MultiGeMS and VarScan were also recorded.

For the simulation study analysis, the specifications of the computer used are CPU: AMD Opteron^{TM} Processor 6376 2.3 GHz and RAM: 512 GB. Total CPU time and maximum procedure memory utilized by the SNV callers running on a single thread are provided in Table 3, whereas other computational performance details are given in Table 5 of the Supplementary Materials. It is noteworthy that the MultiGeMS procedure has the lowest memory requirement of all the tested SNV callers. Also, MultiGeMS outpaces all the multiple sample SNV callers other than GATK. Though, while GATK requires 5.9 fewer minutes than MultiGeMS, its memory requirement is roughly 50 times greater than MultiGeMS.

For the human chromosome 22 analysis, the specifications of the computer used are CPU: AMD Opteron^{TM} Processor 6136 2.4 GHz and RAM: 16 GB. The single thread results are listed in Table 10. Also, since MultiGeMS and GATK have multiple thread options, the results of these SNV callers running on four threads are listed in Table 7 of the Supplementary Materials.

A disparity can be seen in the RAM utilized in GATK and VarScan compared to the other multiple sample SNV callers. Since GATK and VarScan are based on the Java programming language, they tend to utilize more of the available memory than the SNV callers written in C (SAMtools) or C++ (MultiGeMS, FreeBayes).

Given single thread processing, Table 10 identifies MultiGeMS as the most memory efficient SNV caller. It is also the second fastest, in terms of total CPU time. As seen in Table 7 of the Supplementary Materials, running on multiple threads can further improve the MultiGeMS SNV calling time. This, however, comes at a cost of using more memory. In particular, though GATK on 4 threads can complete the SNV calling in about 3 h, 11 GB of RAM are utilized. The performance of MultiGeMS on four threads seems more reasonable, especially for computers with limited RAM, requiring less than 3.5 h and much less RAM.

Table 9. Human chromosomes 18–22 data analysis results with an additional unrelated sample

	Count	Recall	Precision	*F*-score
MultiGeMS	615 349	0.8628	0.7364	0.7946
VarScan	636 082	0.8608	0.7108	0.7787
SAMtools	654 780	0.8684	0.6966	0.7731
FreeBayes	666 385	0.8508	0.6706	0.7500
GeMS-pooled	410 884	0.5994	0.7662	0.6726
GATK	946 836	0.8802	0.4883	0.6282
GeMS	918 512	0.8818	0.5042	0.6416

The results are sorted in descending order of *F-score*.

Given these results, it is expected that modern high-performance computers will have no difficulties in performing the SNV calling procedures described in Section 3. For computers with limited resources, multiple sample SNV callers based on C/C++, especially MultiGeMS and FreeBayes, will be especially efficient.

4 Discussion

Since each genomic site is considered independently in the MultiGeMS likelihood, analysis of each site can be parallelized for simultaneous computation. Therefore, we have utilized parallel computing based on the multi-core CPU framework (OpenMP technology, GNU C/C++) for increased procedure performance.

The MultiGeMS software package currently supports the input of the SAMtools pileup alignment format (http://samtools.sourceforge.net/pileup.shtml) and will soon support SAM/BAM alignment files (Li *et al.*, 2009). For more information on filtering SAM/BAM files before converting to SAMtools pileup files, please see the PDF document entitled 'Filtering Alignment Files' (https://github.com/cui-lab/multigems).

The MultiGeMS algorithm can be expanded to call other types of variants, such as insertions and deletions, by extending the usage of the *N* variable in the existing framework. This area of investigation would need to consider the impact of utilizing the base-call quality scores when calling indels or other variant types, as compared to SNVs. For users currently desiring an indel calling analysis, procedures such as FreeBayes, GATK, SAMtools, VarScan or DINDEL (Albers *et al.*, 2011) are available and can be added to an analysis pipeline. Likewise, procedures are available for the detection of other types of variation, such as copy number variation.

Table 10. Computational resources used by SNV callers running on a single thread

	FreeBayes	MultiGeMS	GeMS	GeMS-pooled	GATK	VarScan	SAMtools
Pre-processing CPU Time (min)	NA	54	54	283	NA	631	NA
Procedure CPU Time (min)	464	508	572	380	792	650	1493
Total CPU Time (min)	464	562	626	663	792	1281	1493
Average Pre-processing Memory (MB)	NA	50	50	564	NA	57	NA
Maximum Pre-processing Memory (MB)	NA	51	51	1100	NA	57	NA
Average Procedure Memory (MB)	73	22	2775	2709	8050	1852	90
Maximum Procedure Memory (MB)	91	27	5600	5700	9100	2000	91

These figures are from the human chromosome 22 analysis in Section 3.2. The results are sorted in ascending order of total CPU time in minutes.

Acknowledgements

The authors are grateful to the referees and associate editor. They also thank the Institute for Integrative Genome Biology for providing the bioinformatics cluster.

Funding

This work was supported by the National Science Foundation [ATD-1222718 to G.M., W.C., X.C.]; the University of California, Riverside [AES-CE RSAP A01869 to G.M., W.C., X.C.]; R01HG113147 [to MP. R., M.L.]; the National Science Foundation of China [NSFC 11301554 to N.Y., NSFC 61103167, 31271410, 31401076]; the Ministry of Science and Technology [high-tech (863) grant 2012AA023107, 2014AA021502]; the Sino-German Research Center [GZ878]; and the Huazhong University of Science and Technology [start-up fund to X.S., K.N.].

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium, Abecasis,G.R., Altshuler,D. *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Albers,C.A. *et al.* (2011) Dindel: Accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973.

Barnett,D. *et al.* (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, btr174–bt1692.

Consortium,T.G.P. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint* arXiv:1207.3907 [q-bio.GN].

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Le,S.Q. and Durbin,R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.*, **21**, 952–960.

Li,B. *et al.* (2012) A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet.*, **8**, e1002944 +.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

McKenna,A. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

Muralidharan,O. *et al.* (2012) A cross-sample statistical model for SNP detection in short-read sequencing data. *Nucleic Acids Res.*, **40**, e5.

The International HapMap 3 Consortium. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

You,N. *et al.* (2012) SNP calling using genotype model selection on high-throughput sequencing data. *Bioinformatics*, **28**, 643–650.