BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 12 2014, pages 1707–1713
doi:10.1093/bioinformatics/btu067

Genetics and population analysis

Advance Access publication February 19, 2014

Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals

Anthony Youzhi Cheng[1], Yik-Ying Teo[1,2,3,4,5] and Rick Twee-Hee Ong[1,*]

[1]Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117597, [2]Life Sciences Institute, National University of Singapore, Singapore 117456, [3]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, [4]NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456 and [5]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Whole-genome sequencing (WGS) is now routinely used for the detection and identification of genetic variants, particularly single nucleotide polymorphisms (SNPs) in humans, and this has provided valuable new insights into human diversity, population histories and genetic association studies of traits and diseases. However, this relies on accurate detection and genotyping calling of the polymorphisms present in the samples sequenced. To minimize cost, the majority of current WGS studies, including the 1000 Genomes Project (1 KGP) have adopted low coverage sequencing of large number of samples, where such designs have inadvertently influenced the development of variant calling methods on WGS data. Assessment of variant accuracy are usually performed on the same set of low coverage individuals or a smaller number of deeply sequenced individuals. It is thus unclear how these variant calling methods would fare for a dataset of ~100 samples from a population not part of the 1 KGP that have been sequenced at various coverage depths.

Results: Using down-sampling of the sequencing reads obtained from the Singapore Sequencing Malay Project (SSMP), and a set of SNP calls from the same individuals genotyped on the Illumina Omni1-Quad array, we assessed the sensitivity of SNP detection, accuracy of genotype calls made and variant accuracy for six commonly used variant calling methods of GATK, SAMtools, Consensus Assessment of Sequence and Variation (CASAVA), VarScan, glfTools and SOAPsnp. The results indicate that at 5× coverage depth, the multi-sample callers of GATK and SAMtools yield the best accuracy particularly if the study samples are called together with a large number of individuals such as those from 1000 Genomes Project. If study samples are sequenced at a high coverage depth such as 30×, CASAVA has the highest variant accuracy as compared with the other variant callers assessed.

Availability and implementation:

Contact: twee_hee_ong@nuhs.edu.sg

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

Received on August 27, 2013; revised on December 28, 2013; accepted on January 28, 2014

*To whom correspondence should be addressed.

1 INTRODUCTION

Fueled by rapidly decreasing costs and advancements in high-throughput sequencing technologies, it is now possible to sequence one complete human genome at a fraction of the time and cost of the initial Human Genome Project. These developments have thus transformed the field of human genomics studies and have contributed significantly to our knowledge and understanding of human diversity (The 1000 Genomes Project Consortium, 2010, 2012) and population histories (Green *et al.*, 2010; Reich *et al.*, 2010; Schuster *et al.*, 2010), as well as aided the identification of causal variants for human diseases (Ng *et al.*, 2010) and traits (Morrison *et al.*, 2013). The successful application of whole-genome sequencing (WGS) in these diverse fields of studies, however, relies on the accurate detection and genotype calling of the polymorphisms present in the individuals studied. Compared with the traditional Sanger sequencing technology, WGS by next-generation sequencing technologies have much shorter reads and higher per-base sequencing error rates, which confounds true polymorphisms from sequencing errors (Bentley *et al.*, 2008).

Intuitively, having high sequencing coverage data where each base in the genome is sequenced multiple times will allow the true polymorphisms to be better distinguished from the sequencing errors. As such, single-sample variant calling methods will usually perform well with high coverage depth, but fare poorly when applied to individuals sequenced at low coverage depths (Liet *et al.*, 2009b). However, deep WGS on a large number of individuals is currently not feasible economically.

Based on the results from a simulation study conducted (Le and Durbin, 2011), it was found that for a given fixed amount of sequencing performed, sequencing a large number of individuals at low coverage depths of 4–6× provide higher power in detecting rare single nucleotide polymorphisms (SNPs) than compared with sequencing a smaller number of individuals at high coverage depth. The former design is what has been adopted by the 1 KGP. This, thus, resulted in the development of different variant calling methods that seek to generate accurate genotype calls at variant sites identified across multiple individuals that have been sequenced at low coverage depths. However, the assessment of these variant calling methods are performed mainly on simulated data and on a small number of individuals who have been sequenced to high coverage depths of at least 30×.

In this work, we thus seek to investigate the following issues: given a population-based sample of individuals such as the Southeast Asian Malays that are not present in 1 KGP and are deep whole-genome sequenced (Wong *et al.*, 2013), which of the variant calling methods (i) GATK (McKenna *et al.,* 2010); (ii) SAMtools (Li *et al.*, 2009a); (iii) Consensus Assessment of Sequence and Variation (CASAVA); (iv) VarScan (Koboldt *et al.*, 2009); (v) glfTools (http://www.sph.umich.edu/csg/abecasis/glfTools); or (vi) SOAPsnp (Li *et al.*, 2009b) will have the best performance in (i) variant discovery or detection; (ii) genotype calling and (iii) overall variant accuracy defined as a product of the two above metrics across sequencing coverage depths from 5 to 30×.

2 MATERIALS AND METHODS

2.1 Samples and generation of sequencing and genotyping data

In the Singapore Sequencing Malay Project (SSMP), 100 unrelated individuals (50 male and 50 female) of Southeast Asian Malay ancestry residing in Singapore were pseudo-randomly selected from the multi-ethnic cohort of the Singapore Population Health Study. Genomic DNA for all 100 individuals was extracted and delivered to the Illumina facility at Hayward, CA, USA, for WGS on the Illumina HiSeq 2000, with a target sequencing coverage of 30× and having paired-end reads of 2 × 100 bp. The sequenced reads were then mapped to the human reference genome (NCBI Build 37) using the Efficient Large-Scale Alignment of Nucleotide Databases v.2e software, a component of Illumina proprietary sequencing-analysis software CASAVA v.1.8, where the alignment results are stored per sample in a single BAM file. Four samples were subsequently excluded owing to population structure outlier or anomalous distribution of insert sizes. To ensure consistency in the analysis with different variant calling software, all of the remaining 96 sample BAM files were re-headed with the UCSC hg19 reference after checking for length and sequence inconsistencies. Reads that are not properly aligned, mapped or paired, or deemed to be PCR duplicates or those that did not pass the chastity filter in Illumina CASAVA pipeline were removed from the BAM files. Approximately, 87.71 and 94.25% of the callable bases (excluding 'N' regions) in chromosomes 1 and 20, respectively, were still covered by the sequence reads.

Each of these 100 individuals was also genotyped on the Illumina HumanOmni1-Quad (Omni1-Quad) array for internal QC measures by the Illumina sequencing facility. Genotype calls were determined using the Illumina Genome Studio (manifest version D) for 1 138 747 markers, and the marker positions were later updated to NCBI Build 37 coordinates. Markers that met any of the following criteria were then filtered: (i) invalid chromosomal position or strand; (ii) duplicated chromosomal positions; (iii) high rates of missingness (call rate <95%) and (iv) Hardy–Weinberg equilibrium $P<10^{-4}$. This resulted in a 'qc-cleaned' set of 994 302 autosomal markers. In this evaluation, we only considered the QC SNPs (excluding indels) on chromosomes 1 and 20 from the Omni1-Quad array with no missing genotype calls in any of the 96 samples, thus yielding a set of 79 029 + 26 805 = 105 834 loci (Table 1) and 105 834 × 96 = 10 160 064 genotypes, which will be used as the 'truth-dataset' for comparison.

Table 1. Number of loci on chromosomes 1 and 20 from Omni1-Quad genotyped on 96 SSMP samples as 'truth-dataset' for comparison

Allele frequency spectrum	Number of loci on chromosome 1	Number of loci on chromosome 20
Monomorphic reference	14 635	6202
Monomorphic alternate	1524	329
Rare (MAF <1%)	2452	825
Low (1% ≤ MAF < 5%)	7967	2558
Common (MAF ≥ 5%)	52 451	16 891

2.2 Down-sampling of BAM files

Statistical down-sampling was performed with PICARD on the 30× coverage BAM files where the original reads are scaled with a supplied probability parameter and in the process retaining both reads in a read pair. We thus down-sampled the 30× BAM files to 5, 10 and 20× with the following probabilities of 1/6, 1/3 and 2/3, respectively.

2.3 Variant calling methods

There are basically two different types of variant calling methods: (i) single-sample and (ii) multi-sample. Some methods such as CASAVA are strictly single-sample callers, while others such as GATK and SAMtools can be used as either single-sample or multi-sample calling. We therefore applied GATK, SAMtools, CASAVA, VarScan, glfSingle and SOAPsnp on each of the 96 individuals separately (as single-sample caller where the method will be prefixed with a S- if it can also be used as a multi-sample caller); GATK, SAMtools, VarScan and glfMultiples on all 96 individuals together (multi-sample calling). In addition, we also assessed the performance of both GATK and SAMtools as multi-sample callers at 5 and 30× sequencing coverage by calling all 96 Malay individuals together with the 1092 low coverage whole-genome sequenced samples from Phase 1 of the 1000 Genomes Project (prefix with 1kgp-). For variants on chromosome 1, we only assessed the performance of the following callers: GATK, SAMtools (both single and multi-sample calling) and CASAVA.

Different preprocessing steps were performed for the reads on chromosomes 1 and 20. In chromosome 1, reads were preprocessed according to the standard procedures recommended for each variant calling methods, while the same preprocessing steps (GATK IndelRealigner, PICARD's MarkDuplicates, GATK BaseRecalibrator) were applied on the chromosome 20 reads for all methods except CASAVA and SOAPsnp, which have their preprocessing steps built into the methods.

For CASAVA, we applied the assembleIndels and callSmallVariants modules in the CASAVA analysis suite and applied a Q20 variant quality score threshold to obtain the variants detected and genotype calls.

In SAMtools, BAQ was enabled in mpileup, with reads with mapping quality scores >90 used to calculate the likelihoods at all sites before bcftools was applied to call the variants and genotypes. SNPs were filtered using the varFilter module in vcfutils, with the following quality criteria: (i) variant quality >3; (ii) maximum read depth = mean read depth + 3 standard deviation of read depth; (iii) minimum read depth of 3; and (iv) SNP must not be within 10 bp of a gap.

GATK UnifiedGenotyper was first used to determine the variant calls before filtering was performed with the VQSR (VariantRecalibrator) module at 99% truth sensitivity level.

The default parameters were used for variant calling in VarScan (consenus sequence with *mpileup2cns*), glfTools and SOAPsnp before variants were filtered with variant quality of Q20 in the single-sample mode and Q60 in the multi-sample modes (Liu *et al.*, 2013).

2.4 Variant detection

To determine which of the variant calling methods perform the best in terms of variant discovery or detection across the different sequencing

coverage depths, we used two evaluation metrics. The first metric is the sensitivity of each variant calling method to correctly detect the polymorphic sites as identified on the Omni1-Quad array:

$$Sensitivity = \frac{Num\_correct\_polymorphic\_loci}{Num\_polymorphic\_loci\_array}$$

where

Num_correct_polymorphic_loci is the number of polymorphic loci that has been identified by both the variant calling method and Omni1-Quad array where different alleles are present at the site for 96 SSMP samples.

Num_polymorphic_loci_array is the total number of polymorphic loci as identified on the Omni1-Quad array for 96 SSMP samples (sum of rare, low and common frequency loci, which is a constant of 62 870 and 20 274 for chromosomes 1 and 20, respectively).

In addition, we also assessed a false-positive measure for each variant calling method by determining how many of the monomorphic reference loci on the Omni1-Quad array, i.e. all 96 SSMP samples, are identified to carry the homozygous reference genotypes at that loci, and yet are erroneously determined to be a variant loci by the variant calling method:

$$False\_Positive = \frac{Num\_incorrect\_monomorphic\_loci}{Num\_monomorphic\_ref\_loci\_array}$$

where

Num_incorrect_monomorphic_loci is the number of loci determined to be a variant site by the variant calling method, but all 96 SSMP samples carry the homozygous reference genotypes as identified from the Omni1-Quad array.

Num_monomorphic_ref_loci_array is the total number of loci identified in Omni1-Quad array where all 96 SSMP samples carry the homozygous reference genotypes (constant of 14 635 and 6202 for chromosomes 1 and 20, respectively).

2.5 Genotype calls concordance

To determine which is the 'best' genotype caller, we assessed the concordance of the genotypes made by each of the variant calling methods to those genotypes identified on the Omni1-Quad array for each of the 96 SSMP samples at the variant loci across the various sequencing coverage depths considered:

$$Variant\_Loci\_Concordance = \frac{Num\_correct\_variant\_loci\_genotypes}{Num\_variant\_loci\_genotypes\_called}$$

where

Num_correct_variant_loci_genotypes is the number of genotypes whose calls made by the variant calling method are concordant with that obtained in the Omni1-Quad array at all variant loci.

Num_variant_loci_genotypes_called is the total number of genotypes called by each variant calling method at all variant loci (where at least 1 of the 96 SSMP samples carry a variant allele according to Omni1-Quad).

In this concordance measure, we are only measuring the proportion of correct genotype calls made by each variant calling method, which is thus unaffected by the sensitivity of the different methods. In addition, the homozygous reference genotypes as identified from Omni1-Quad array are not counted toward the concordance measure.

2.6 Variant accuracy

We then define an overall measure of variant accuracy where both sensitivity and genotype concordance are considered to determine the 'best' variant calling method:

$$Variant\_Accuracy = \frac{Num\_correct\_variant\_loci\_genotypes}{Num\_variant\_loci\_genotypes}$$

where

Num_correct_variant_loci_genotypes is the number of genotypes whose calls are made by the variant calling method and are concordant with that obtained in the Omni1-Quad array at all variant loci. *Num_variant_loci_genotypes* is the total number of genotypes at all variant loci that is a constant of 6 181 824 and 1 977 888 for chromosomes 1 and 20, respectively.

3 RESULTS

3.1 Assessing variant detection

It is clear that there exists a monotonically increasing relationship between sequencing coverage depth and the sensitivity to correctly determine the polymorphic sites (raw and filtered) on the Omni1-Quad array for all variant calling methods assessed (Fig. 1A, Supplementary Figs S1 and S5A and Supplementary Table S1). For most of the methods, there is only a marginal increase in the number of additional polymorphic sites correctly detected beyond 20× for both raw and filtered sites (Supplementary Table S1). It is somewhat surprising that at 30× coverage, GATK is unable to detect ~0.5, ~1.2 and ~1%, ~1.3% of the filtered common frequency variant loci as compared with SAMtools and CASAVA for chromosomes 1 and 20, respectively. We believe this could likely be attributed to GATK's VQSR filtering, as the sensitivity achieved before filtering is comparable among the methods (Supplementary Tables S2 and S5). Categorizing by minor allele frequencies, we find that the increase in sequencing coverage depth primarily identifies additional rare (Fig. 1D and Supplementary Figs S4 and S5D) and low frequency (Fig. 1C and Supplementary Figs S3 and S5C) sites, while the number of common frequency variant loci detected (Fig. 1B and Supplementary Figs S2 and S5B) remains almost constant for most variant calling methods across the various sequencing depths.

The single-sample variant calling methods (S-GATK, S-SAMtools, S-VarScan and glfSingle) were found to have better sensitivity in correctly detecting the variant sites than if used as multi-sample methods across all different minor allele frequency spectrums considered, though their differences decrease with increasing sequencing depths (Fig. 1A, Supplementary Fig. S5A and Supplementary Tables S1–S5). However, for each sequencing coverage depth considered, the differences between the single-sample and multi-sample modes for each variant method increases after filtering is applied. At 5× coverage, the differences in sensitivity between S-GATK and GATK are 0.4 and 1.6% for raw and filtered, respectively. This trend can also be observed in SAMtools, VarScan and glfTools, though the differences are not as large as GATK. Therefore, the effects of filtering contribute significantly to the differences observed between the methods for sensitivity of filtered loci detected particularly for GATK at 5× coverage. For most variant calling

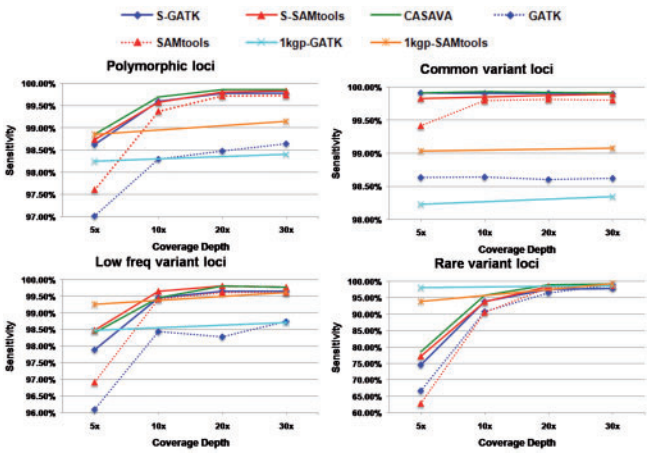

Fig. 1. Comparison of the sensitivity of GATK and SAMtools (single-sample mode, multi with 96 SSMP samples and multi combining 96 SSMP samples with 1092 1 KGP Phase 1 samples), and CASAVA across various sequencing coverage depths for filtered sites of (**A**) all polymorphic loci; (**B**) common frequency variant loci; (**C**) low frequency variant loci; (**D**) rare frequency variant loci across 96 SSMP samples identified on chromosome 20 of the Omni1-Quad array. Sensitivity values for 1kgp-GATK and 1kgp-SAMtools at 10 and 20× are interpolated between 5 and 30×

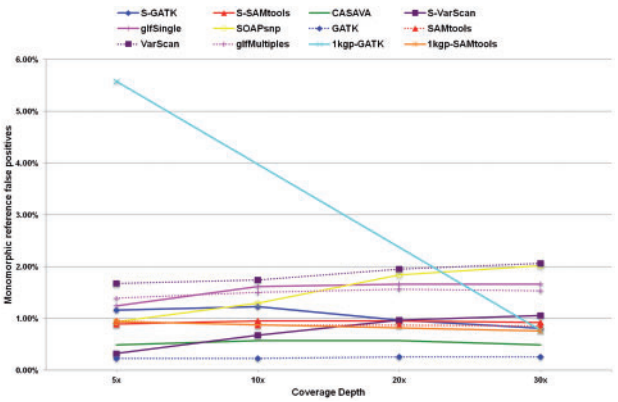

Fig. 2. Comparison of the false positives of variant calling methods in erroneously determining a reference site to be a variant locus across various sequencing coverage depths for the filtered loci on chromosome 20 of the Omni1-Quad array. False-positive values for 1kgp-GATK and 1kgp-SAMtools at 10 and 20× are interpolated between 5 and 30×

methods, the differences in sensitivity between the rare and filtered sites at 5× coverage depth can be primarily attributed to the rare variants determination, where, for example, in SOAPsnp, the differences are 1.3, 24.3 and 48.6% for common, low and rare frequency variant loci identification, respectively (Supplementary Tables S2–S4). However, in GATK, the differences are almost similar across the frequency spectrum, and thus the effects of filtering in GATK are not biased toward any frequency spectrum.

The observation that single-sample variant calling methods have higher sensitivity than their multi-sample mode at low sequencing coverage depths, particularly at 5× seem counter-intuitive. However, if the 96 SSMP samples are called together with the low coverage 1092 samples from the 1000 Genomes Project, the sensitivity achieved by these multi-sample callers (1kgp-GATK and 1kgp-SAMtools) are comparable with the single-sample variant callers across all frequency spectrums assessed, and are in fact much higher with ~20% improved sensitivity for the rare frequency variants at 5× coverage (Fig. 1D and Supplementary Table S4). At 5× coverage, sensitivity for 1kgp-GATK and 1kgp-SAMtools is the highest among the callers compared for all polymorphic raw variants, while filtering does slightly lower the sensitivity for common variants, but for low and rare variants, the sensitivity is still much higher than the other callers. For high sequencing coverage depths of 30×, almost all of the variant callers achieve >99% sensitivity.

When we compared the false-positive measures between the different variant calling methods in erroneously determining a reference site to be a variant locus (Fig. 2 and Supplementary Tables S6 and S7), we observed that by increasing the sequencing coverage depths, the false positives generally decrease for most methods, with GATK having the lowest value at all depths considered. At 5× coverage, the false positives for 1 kgp-GATK is, however, much higher than that of GATK, while values for 1kgp-SAMtools and SAMtools are comparable. However, the differences in false positives between raw and filtered variants are significantly higher for 1kgp-GATK, GATK and SOAPsnp than the other methods, suggesting that the variant filtering for GATK and SOAPsnp while reducing the false positives significantly also substantially reduces its sensitivity.

3.2 Impact of filtering on variant detection

Variant filters are an inherent, but important, procedure in the variant calling methods, as it serves to reduce the number of false positives while optimally retaining the true variants. The filtering procedures range from simple thresholds on quality scores to sophisticated machine learning methods that separate true variants from sequencing artifacts. We thus attempted to study the effectiveness of the filtering procedure used by the different variant calling methods by measuring the ratio of decrease in sensitivity against the decrease in discordance after filtering (Table 2). An effective filter should thus have low values, preferably below 1, as it means the filter applied has removed substantially more false positives than true variants.

Generally, for each variant calling method, as the sequencing coverage depth increases, the rates decrease, which implies greater confidence and ability of each method in distinguishing between true variant loci and sequencing artifacts. At 5× coverage depth, filters applied in SAMtools and CASAVA are unable to differentiate between true variants from false positives, and thus a significant portion of true variant loci are filtered together with the discordant calls. Interestingly for GATK, the ratios are all below one, thus suggesting that across 5–30× coverage depths, the sophisticated filtering process of VQSR had removed more discordant or artifacts than true variant calls. Thus, at 5× coverage depth, VQSR applied in GATK was most efficient at differentiating between true variants and artifacts, while at 30× coverage, a simple Q20 filter on the quality threshold was sufficient.

Table 2. Effectiveness of filters used in each variant calling methods across various sequencing coverage depths

Chromosome	Method	5×	10×	20×	30×
Chromosome 1	GATK	0.69	0.64	0.66	0.61
	SAMtools	5.46	4.03	2.73	2.01
	CASAVA	1.69	0.60	0.36	0.23
Chromosome 20	GATK	0.67	0.66	0.69	0.61
	SAMtools	3.54	0.54	0.84	1.22
	CASAVA	0.94	0.29	0.05	0.04
	SOAPsnp	3.54	0.72	0.11	0.03
	glfMultiples	4.20	3.56	0.59	0.25

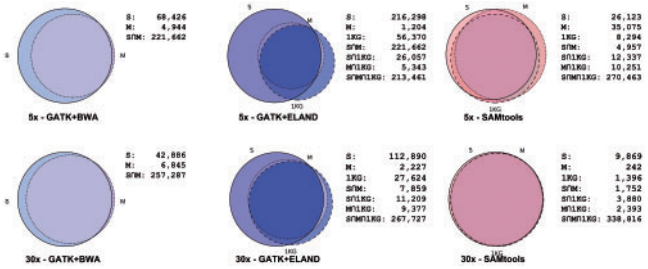

Fig. 3. Venn diagrams showing the total number of variant loci detected on chromosome 20 at 5 and 30× sequencing coverage for GATK with BWA mapping; GATK and SAMtools with ELAND mapping for single, 96 SSMP multi-sample and 96 SSMP with 1092 1 KGP multi-sample variant calling configurations

3.3 Impact of alignment on variant calling

Although the sensitivity of S-GATK and GATK is almost similar, the total number of variant sites detected by both methods differs substantially such that almost half of the sites identified in S-GATK are unique (Fig. 3 and Supplementary Fig. S6). Assuming the widely accepted genome-wide Ts/Tv ratio of 2.0–2.4 is also applicable for chromosome 20, we find that almost all of known Ts/Tv ratios at all sequencing coverage depths for almost all variant callers are within the range. However, at 5 and 30×, the Ts/Tv ratios for novel sites differ significantly for S-GATK, while the values for GATK, S-SAMtools and SAMtools are within the range (Supplementary Table S8). To investigate further, we used a different alignment software BWA-MEM with default parameters (Li *et al.*, 2013), while applying the same processing steps for variant calling at 5 and 30× for both single- and multi-sample modes. The number of variant loci identified in S-GATK at both 5 and 30× reduced significantly and showed a much greater overlap between GATK in multi-sample mode (Fig. 3) and in addition the novel Ts/Tv ratios are now much closer or within the acceptable range (Supplementary Table S8). Therefore, good read alignments are fundamental in the variant calling process (Ruffalo *et al.*, 2011), and the combinations of aligners and variant callers are similarly essential in both sensitivity and specificity of the variant calling process.

3.4 Assessing genotype concordance

From Figure 4, we observe that all methods are able to obtain high genotype concordance rates at the variant loci across the various sequencing coverage depths, where at the lowest coverage of 5×, the minimum and maximum concordance are at 78.21 and 98.96% achieved by VarScan and SOAPsnp, respectively. At 30× sequencing coverage, the minimum and maximum genotype concordance are achieved by VarScan and GATK at 99.79 and 99.98%. In general, all variant calling methods are able to produce the correct genotype calls once a variant site is correctly determined, and the higher sequencing coverage depths do result in better genotype calls and with only marginal increase beyond 20× coverage depth. Comparatively, GATK has the highest genotype concordance rates among all methods across the various sequencing coverage depths (Fig. 4 and Supplementary Fig. S7).

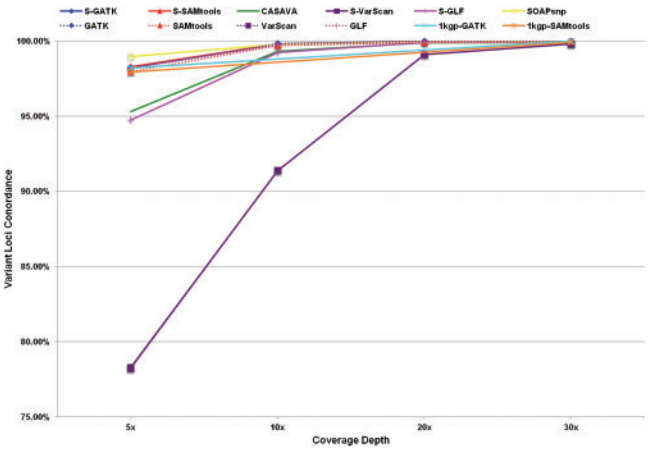

Fig. 4. Comparison of the variant genotype concordance of variant calling methods across various sequencing coverage depths for variants (filtered sites) on chromosome 20 of the Omni1-Quad array. Genotype concordance values for 1kgp-GATK and 1kgp-SAMtools at 10 and 20× are interpolated between 5 and 30×

3.5 Assessing the variant accuracy

Aggregating variant sensitivity and genotype concordance into a single variant accuracy measure, we observe similar trends (Fig. 5 and Supplementary Fig. S8) where variant accuracy increases as sequencing coverage depths increase, and after 20× coverage, the additional gains in accuracy are minimal for all methods considered. At 5× coverage depth, the multi-sample methods of GATK, SAMtools and glfMultiples outperform the rest of the variant callers, with 1kgp-GATK and 1kgp-SAMtools having the highest variant accuracy. However, at 20× and beyond, there does not seem to have additional advantages in pooling of information across the samples for most methods, where at 30× coverage depth, the most accurate variant callers are glfSingle, S-GLF and CASAVA. Although glfSingle seems to be the best variant caller method, as it has among the methods, one of the highest accuracy across the sequencing coverage depths, it also unfortunately has a much higher false discovery rate than compared with the other methods, where it is ~1% higher than CASAVA at both 5 and 30×, while the false discovery rate for CASAVA remains at low constant across the coverage depths.

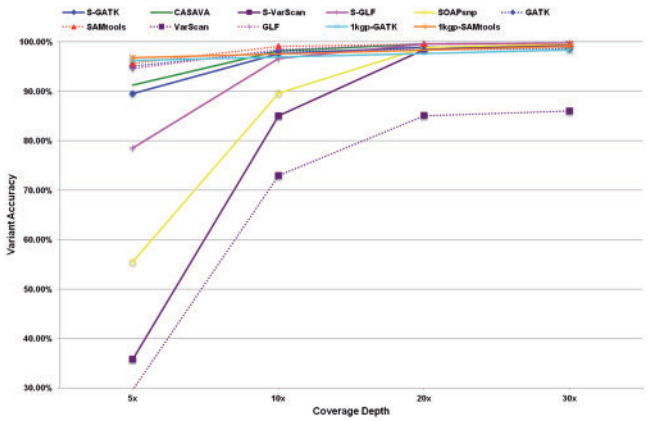

Fig. 5. Comparison of the accuracy of variant calling methods across various sequencing coverage depths for variants (filtered) on chromosome 20 of the Omni1-Quad array. Accuracy values for 1kgp-GATK and 1kgp-SAMtools at 10 and 20× are interpolated between 5 and 30×

4 DISCUSSION

In this article, we compared the sensitivity of variant site detection, genotype calls concordance and an overall measure of variant accuracy across six commonly used variant calling methods of (i) GATK; (ii) SAMtools; (iii) CASAVA; (iv) VarScan; (v) glfTools; and (vi) SOAPsnp at various sequencing coverage depths of 5–30×. First, our findings indicate that increasing the sequencing coverage depth will ultimately improve both the sensitivity and accuracy of variant genotype calls for all variant calling methods. However, it is probably still not economically feasible to perform deep whole-genome sequencing for a large number of individuals for various studies of human traits and diseases. For individuals who are sequenced at low coverage depths averaging 4–6× such as in 1 KGP, the multi-sample method of variant calling such as GATK and SAMtools which aggregate information across multiple samples in correctly detecting variant sites with low false discovery rate and generate accurate genotype calls should be used. An even better approach would be to include the large number of samples in 1 KGP, together with the sequenced individuals in the study for multi-sample calling. If the sequencing depth per sample can be increased to at least 30×, then the single-sample method of CASAVA, which is the default variant caller in Illumina's pipeline, seems much better at distinguishing between sequencing errors from true polymorphisms present in the samples than the multi-sample methods. We also caution that stricter filters, rather than the default recommended, should be applied to reduce the number of false positives if GATK and/or SAMtools are used in the single-sample mode of variant calling. If possible, multiple samples should be called together using GATK and/or SAMtools to take advantage of the strengths of these methods. In addition for GATK, it is also probably advisable to use BWA for reads mapping.

We have relied on the accuracy of the genotypes obtained from Omni1-Quad array to assess the performance of the different variant calling methods. This is obviously imperfect, as genotype calling algorithms for microarrays tend to be unreliable at SNPs with low minor allele counts, even if Illumina's proprietary SNP genotype-calling software GenCall for the Illumina microarrays has been reported to exhibit low error rates (Shah *et al.*, 2012). However, in our comparison, any errors in the Omni1-Quad genotypes will contribute the same impact to the comparison of the different variant callers. Not all polymorphisms present on chromosomes 1 and 20 for these 96 SSMP individuals are known and compared. However as non-overlapping sets of variant calls are made by the different variant calling methods, relying on the accuracy of Omni1-Quad array genotypes would be the best and unbiased option. A small number of these SSMP individuals have also been genotyped on the Illumina Omni2.5M array, and common SNPs present on both arrays were checked for concordance. We did not find any significant differences in genotype calls made between the two platforms that have altered the trends observed in our conclusions.

In our assessment of variant calling methods, we have not considered variant calling methods that are designed to take advantage of multi-sample linkage disequilibrium (LD) such as GATK + BEAGLE (Browning and Yu, 2009). The primary aim of these LD-based methods is to generate accurate genotype calls by grouping individuals who are carrying similar haplotypes, which we believe will only benefit specific populations with shorter LD ranges, and might not be generally applicable.

With more WGS-based studies being conducted, we believe there is still need for further active research into accurate variant detection and genotype calling from massive sequence reads from next-generation sequencing platforms, including for insertions and deletions (indels) and structural variants in addition to SNPs. Publicly available resources such as the SSMP database and other population-level deep whole-genome sequencing will be valuable both for the design and evaluation of variant calling methods, as such data will present a more accurate reflection of the variants that are present within each sample.

ACKNOWLEDGEMENTS

This project acknowledges the support of the Saw Swee Hock School of Public Health, the Yong Loo Lin School of Medicine, the National University Health System, the Life Science Institute and the Office of Deputy President (Research and Technology) from the National University of Singapore.

Funding: A.Y.C, R.T.H.O and Y.Y.T acknowledge support from the National Research Foundation Singapore (NRF-RF-2010-05).

Conflict of Interest: none declared.

REFERENCES

Bentley,D.R. *et al.* (2008) Accurate whole-human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Browning,B.L. and Yu,Z. (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.*, **85**, 847–861.

Green,R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Koboldt,D.C. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.

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,H. *et al.* (2009a) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Li,R.Q. *et al.* (2009b) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.
Li,H. *et al.* (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997.
Liu,X. *et al.* (2013) Variant callers for next-generation sequencing data: a comparison study. *PLoS One*, **8**, e75619.
McKenna,A. *et al.* (2012) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
Morrison,A.C. *et al.* (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat. Genet.*, **45**, 899–901.
Ng,S.B. *et al.* (2010) Exome sequencing identifies MLL2 mutations as a cause of kabuki syndrome. *Nat. Genet.*, **42**, 790–793.
Reich,D. *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Sibera. *Nature*, **468**, 1053–1060.
Ruffalo,M. *et al.* (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, **27**, 2790–2796.
Schuster,S.C. *et al.* (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature*, **463**, 943–947.
The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
Shah,T.S. *et al.* (2012) optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics*, **28**, 1598–1603.
Wong,L.P. *et al.* (2013) Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.*, **92**, 1–15.