



## OPEN

## SUBJECT AREAS:

NEXT-GENERATION  
SEQUENCINGCOMPUTATIONAL BIOLOGY AND  
BIOINFORMATICS

GENOMICS

GENETIC MARKERS

# A practical method to detect SNVs and indels from whole genome and exome sequencing data

Daichi Shigemizu<sup>1\*</sup>, Akihiro Fujimoto<sup>1\*</sup>, Shintaro Akiyama<sup>1</sup>, Tetsuo Abe<sup>1</sup>, Kaoru Nakano<sup>2</sup>, Keith A. Boroevich<sup>1</sup>, Yujiro Yamamoto<sup>2</sup>, Mayuko Furuta<sup>2</sup>, Michiaki Kubo<sup>3</sup>, Hidewaki Nakagawa<sup>2</sup> & Tatsuhiko Tsunoda<sup>1</sup>

Received  
20 February 2013

Accepted  
21 June 2013

Published  
8 July 2013

Correspondence and  
requests for materials  
should be addressed to  
T.T. (tsunoda@src.  
riken.jp)

\* These authors  
contributed equally to  
this work.

<sup>1</sup>Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, <sup>2</sup>Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan, <sup>3</sup>Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

The recent development of massively parallel sequencing technology has allowed the creation of comprehensive catalogs of genetic variation. However, due to the relatively high sequencing error rate for short read sequence data, sophisticated analysis methods are required to obtain high-quality variant calls. Here, we developed a probabilistic multinomial method for the detection of single nucleotide variants (SNVs) as well as short insertions and deletions (indels) in whole genome sequencing (WGS) and whole exome sequencing (WES) data for single sample calling. Evaluation with DNA genotyping arrays revealed a concordance rate of 99.98% for WGS calls and 99.99% for WES calls. Sanger sequencing of the discordant calls determined the false positive and false negative rates for the WGS (0.0068% and 0.17%) and WES (0.0036% and 0.0084%) datasets. Furthermore, short indels were identified with high accuracy (WGS: 94.7%, WES: 97.3%). We believe our method can contribute to the greater understanding of human diseases.

Since genetic variation plays a key role in human disease and, in particular, rare Mendelian disorders, one of the most important goals of genetic studies is to identify genetic variants in individuals<sup>1,2</sup>. Next-generation sequencing (NGS) technology<sup>3–5</sup> has made whole genome sequencing (WGS) possible at an individual level. WGS has revealed numerous single nucleotide variations (SNVs), *de novo* mutations<sup>6</sup> and somatic mutations in cancer genomes<sup>7–10</sup> that had not been previously reported. Whole exome sequencing (WES), which captures and sequences only the coding exons of the genome, is also used to identify genetic variations in the coding regions<sup>11</sup>. WES is more cost-effective for exonic regions than WGS, and can obtain a deeper depth of coverage in the target region. WES has been used to successfully identify causal mutations of Mendelian diseases<sup>2,12</sup> and driver mutations in tumors<sup>13–16</sup>.

The data produced by WGS and WES are composed of numerous short read sequences ranging in length from 50–150 bp<sup>17</sup>. So far, many methods have been developed for short read alignment and variant calling. However, accurate detection of SNVs and indels is still difficult and a critical issue. For example, even for a false positive rate of only 0.1%, three million false positive SNVs ( $3 \times 10^9$  (bp)  $\times$  0.001 = 3,000,000) would be identified in the entire human genome. False negative rate is another critical issue, especially for genetic diagnosis and Mendelian disease studies. For the future of personalized medicine and genetic diagnosis, highly accurate variant calling is still one of the most important problems.

Here, we provide a tool package for our previously reported method that detects SNVs and short insertions and deletions (indels)<sup>18</sup>, and show evaluation and comparison with other available methods for both WGS and WES of new test samples generated on the Illumina DNA-sequencing platform. A high concordance rate with the SNP genotyping arrays was observed for both WGS calls (homozygous concordance rate: 99.99%, heterozygous concordance rate: 99.83%) and WES calls (heterozygous and homozygous concordance rate: 99.99%). Compared to other available methods, our method suppressed both false positive and false negative rates (WGS: 0.0068% and 0.17%, WES: 0.0037% and 0.0084%). Furthermore, we identified short indels with high accuracy (WGS: 94.7%, WES: 97.3%). We believe our method is a useful tool for understanding human diseases through WGS and WES analysis. This program, “Variant Caller with Multinomial probabilistic Model (VCMM)”, is publicly available at <http://emu.src.riken.jp/VCMM/>.



## Results

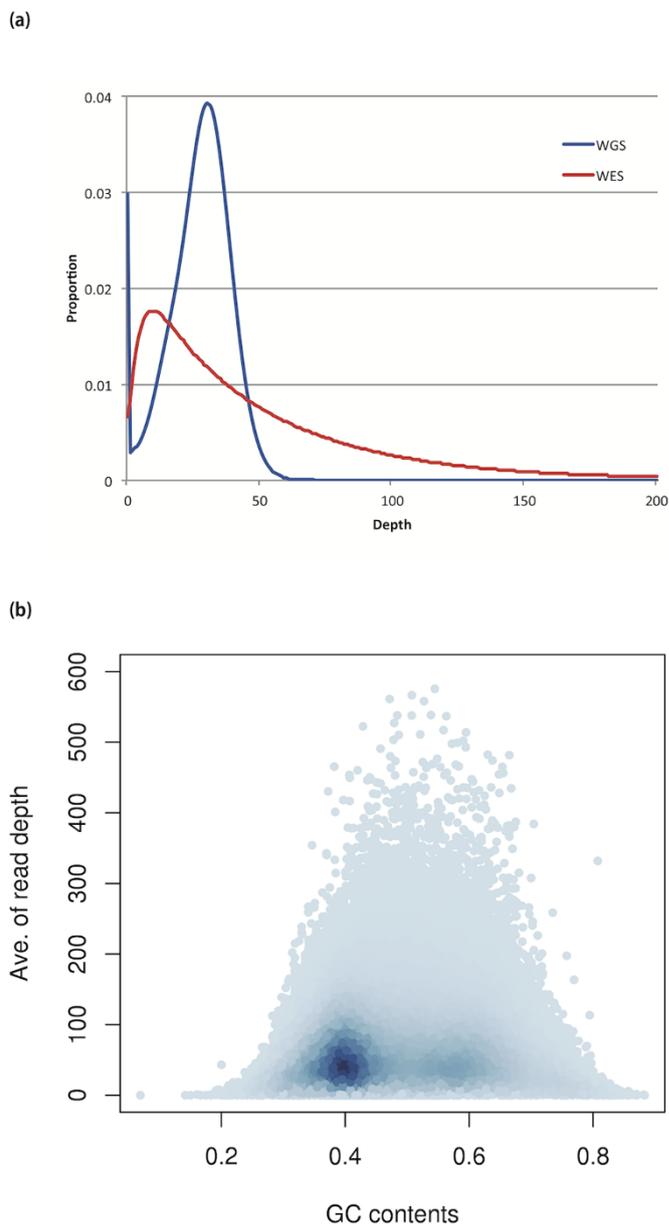
**Sequencing and mapping.** WGS and WES data were generated on the Illumina HiSeq2000 platform with library sizes of 500 bp and 150–200 bp. 101 Gbp and 7 Gbp of short read sequences were obtained for WGS and WES, respectively. Mapping was performed using the short read mapping algorithm BWA<sup>19</sup>; 96.1% of WGS reads and 98.5% of WES reads were mapped to the human reference genome. PCR duplications were defined using SAMtools<sup>20</sup>. The PCR duplication rates were 3.5% and 5.0% in the WGS and WES data, respectively. After PCR duplications removal, the remaining 93.4 Gbp of WGS and 6.5 Gbp of WES contributed sufficient coverage to perform variant calling. A total of 63.7% of the WES reads were mapped to on-target regions.

**Distribution of depth of coverage.** In the WGS data, the average depth was 28.4 and 4.3% of genomic positions were identified to have low read depth (read depth < 5). The average read depth of on-target regions was 70.7 in the WES data, with 4.1% of genomic positions showing low read depth (read depth < 10). We also compared the distribution of the depth of coverage between WGS and WES. The shape of the coverage distributions was quite different (Figure 1a). The distribution for WGS was bell shaped, though approximately 3% of the genome had zero coverage, likely due to repetitive sequence. In contrast, the distribution for WES was wider; the peak of the distribution being lower than the average read depth, and the proportion of positions covered by at least one uniquely mapped read was much higher than that for WGS. This comparison suggests that the skew of the WES distribution is a result of the capturing process (Figure 1b). Improvement of the exon capture method would reduce the proportion of low coverage regions.

**Evaluation of SNV and indel calls.** *SNV calls.* We performed SNV and indel calling using our method, VCMM. The algorithm was originally developed with data from a Japanese male individual (HapMap NA18943)<sup>18</sup>. In this study, we evaluated the accuracy of SNV calling using two independent samples. We compared our results to array-based genotype calls, using Illumina Human OmniExpress BeadChip for WGS and Illumina Human Exome BeadChip for WES. The number of SNPs available for verification was 644,167 for WGS and 193,280 for WES. Since standard NGS analysis aligns reads to a reference sequence and identifies non-reference alleles as variants, we can assume that all non-reported positions were homozygous for the reference allele. We classified variants based on the genotype from the genotyping array (homozygous or heterozygous) and reference allele (homozygous of reference allele and not homozygous of reference allele). We then classified the discordant SNPs as either false positive or false negative candidates (Table 1).

In the WGS analysis, 0.17% ( $n = 1,126$  in OmniExpress BeadChip) of the SNPs were not analyzed due to insufficient depth of coverage (read depth < 5) (Table 1). The total number of false positive and false negative candidates was 98 and 1,083, respectively (Table 1). We performed Sanger sequencing verification for the false positive candidates. Of the 98 false positive candidates, 42 SNPs could not be amplified by PCR. Attempts to amplify these with a lower annealing temperature also failed. Of the amplified 56 false positive candidates, 54 were consistent with our variant calls. The two remaining false positives were located in tandem repeat regions, suggesting that mapping error, rather than sequencing error, was causative. The proportion of false positive and false negative SNPs was 0.0068% ( $44/643,041$ ) and 0.17% ( $1,083/643,041$ ), respectively (Table 1). Note that the false positive rate is a conservative estimation, because PCR unamplified SNPs were also counted as false positives.

In the WES analysis, 1.6% ( $n = 3,083$  in Exome BeadChip) of the SNPs were not analyzed due to insufficient depth of coverage (read depth < 10) (Table 1). The total number of false positive and false negative candidates was 14 and 36, respectively. Sanger sequencing



**Figure 1 | Read depth per nucleotide and GC content.** (a) Distribution of read depth in WGS and WES on-target regions. (b) Distribution of GC content of WES on-target regions.

verification revealed that 7 out of the 14 false positive candidates and 20 out of the 36 false negative candidates were consistent with our variant calls, though 3 out of the 14 false positive and 9 out of the 36 false negative candidates could not be amplified by PCR. The false positive and false negative rates were estimated to be 0.0036% ( $7/190,197$ ) and 0.0084% ( $16/190,197$ ), respectively (Table 1). Note that the reported false positive and the false negative rates are conservative estimations, PCR unamplified SNPs were counted as false positives and false negatives. We also conducted the same examination on additional samples, and equivalent results were obtained (data not shown).

In both the WGS and WES analysis results, the number of SNVs using VCMM was similar to that of previous studies<sup>21</sup> (Table 2). The transition/transversion (Ts/Tv) rates were 2.08 in WGS and 2.39 in WES.

*Indel calls.* It is difficult to detect indels from short read sequence data, but which are likely to be functionally important, particularly



Table 1 | Estimation of the accuracy of VCMM using SNP genotyping platforms

	Genotyping array <sup>†</sup>	WGS or WES <sup>†</sup>	WGS		WES	
			Before Sanger sequencing validation	After Sanger sequencing validation*	Before Sanger sequencing validation	After Sanger sequencing validation*
Not analyzed	-	-	1,126	-	3,083	-
Concordance	No-Ref-Ho	-	137,786	-	2,893	-
	Ref-Ho	-	326,125	-	183,411	-
	Ht	-	177,949	-	3,843	-
	Total	-	641,860	-	190,147	-
False positive	Ref-Ho	No-Ref-Ho	5	2 (2)	0	0
	Ref-Ho	Ht	37	16 (15)	11	6 (3)
	No-Ref-Ho	Ht	31	17 (16)	1	1 (0)
	Ht	Ht (Different genotype)	25	9 (9)	2	0 (0)
	Total		98	44 (42)	14	7 (3)
False negative	Ht	Ho	850	850	23	13 (8)
	No-Ref-Ho	Ref-Ho	233	233	13	3 (1)
	Total		1,083	1,083	36	16 (9)

<sup>†</sup>: No-Ref-Ho; Non reference homozygous genotype, Ref-Ho; Reference homozygous genotype, Ht; Heterozygous genotype.  
<sup>\*</sup>: The numbers in parenthesis represented the number of SNPs that could not be amplified by PCR.

when they cause frameshifts. VCMM identified 763,944 (106,732 in non-repeat region) and 10,999 indels in the WGS and WES data. The proportions of frameshift indels were 45.8% (221/461) in WGS and 51.1% (260/509) in WES. There was no significant difference between these proportions ( $P$ -value = 0.33). Since we could not evaluate all the indels identified, we estimated the concordance rate using PCR and Sanger sequencing verification for a randomly selecting subset of indels. In the WGS analysis, we randomly selected 96 indels from the 461 coding indels detected for validation. Of the 75 successful assays, 71 indels were verified as correct, and 4 were false positives (5.3%). In the WES analysis, we randomly selected 47 indels from 509 coding indels detected for validation. Of the 37 successful assays, 36 indels were verified, and 1 was a false positive (2.7%). In both the WGS and WES analysis results, the number of indels using VCMM was similar to that of previous studies<sup>21</sup> (Table 2). Furthermore, the ratios of indels to SNVs were 0.22 in WGS and 0.14 in WES, similar to that of reported by 1000 Genomes<sup>22</sup> (0.19). Note that we used all indels identified for the calculation in the WGS.

**Comparison with other call methods for SNV.** We compared the VCMM's SNV calling to that of two popular alternative methods, GATK<sup>23</sup> and SAMtools<sup>20</sup> (Figure 2a and c). In the WGS and WES, we observed a large number of common SNVs identified by all three methods, and a similar number of uniquely identified SNVs by VCMM and GATK (Figure 2a and c). The proportion of SNVs identified by the both VCMM and SAMtools was larger in WGS

than WES (Figure 2a and c). We further compared the genotype array concordance rate of our method to that of the two alternative programs. The resulting concordance rates of our analysis were higher than those of GATK and SAMtools for both WGS and WES (Table 3). Our method also achieved acceptable levels for both false positive and false negative rates (false positive rates: 0.015% in WGS and 0.007% in WES, false negative rates: 0.17% in WGS and 0.02% in WES). These results demonstrate our method is an efficient framework for detecting SNVs (Table 3). These comparisons were also performed using three published datasets in HapMap database for WES and another dataset for WGS, and similar results were obtained (see Supplementary Table S1 online).

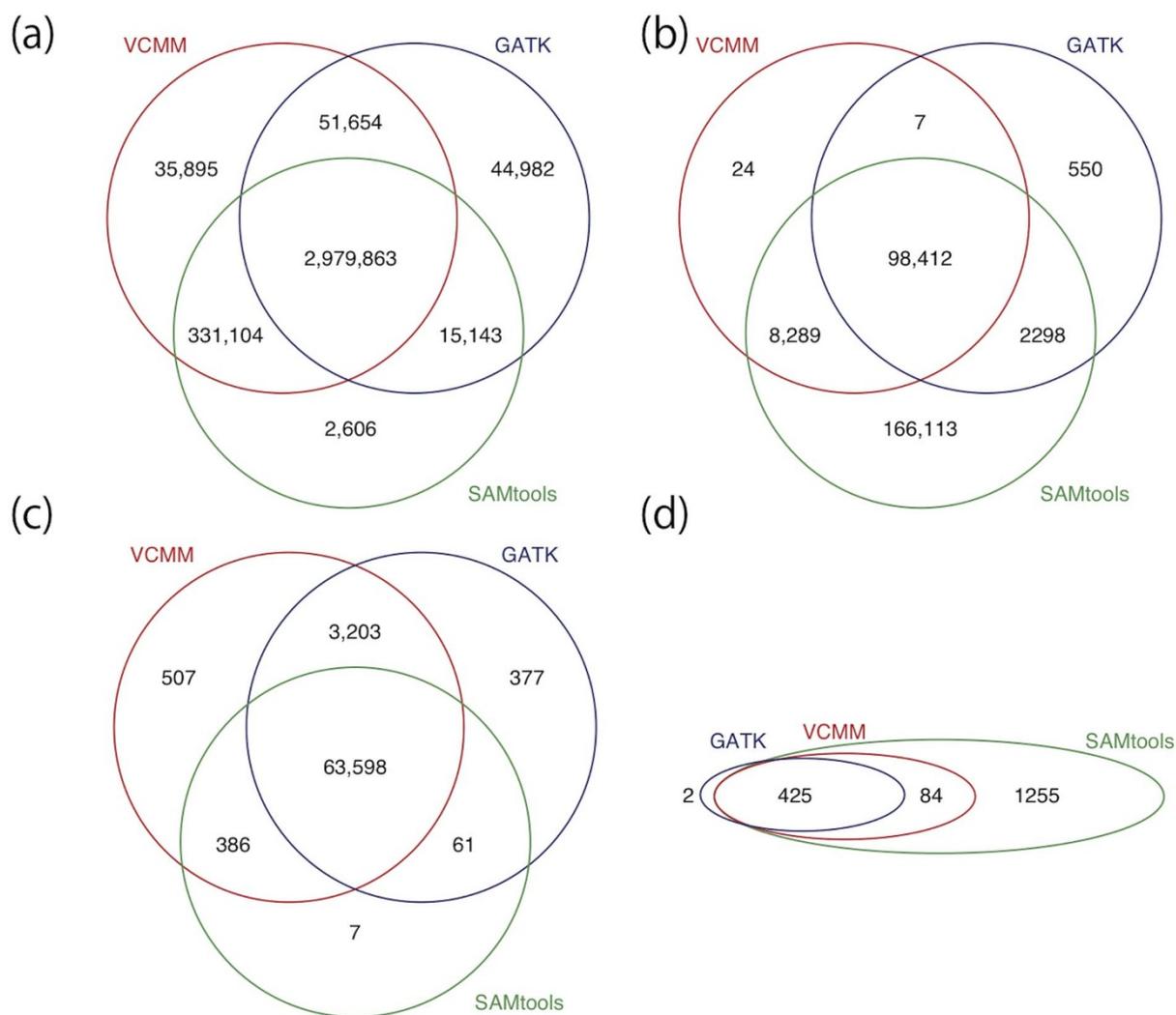
**Comparison with other call methods for indel.** We also compared the indel calling results of each of the three methods (Figure 2b and d). The number of indels identified by GATK was smaller than those by VCMM and SAMtools in both WGS and WES. SAMtools identified larger number of indels than other methods. In the WES, most indels detected by GATK were included these by VCMM (Figure 2d). These results suggest that GATK is more conservative than the other methods. To examine the accuracy of indel calling, we performed Sanger sequencing verification. In the WGS, we sequenced 22 indels in total, 14 were identified by all three methods, four by SAMtools only, two by VCMM and SAMtools, one by GATK only, and one by GATK and SAMtools. For the examined candidates, one indel identified solely by SAMtools, and one by GATK and SAMtools were false positives (see Supplementary Table S2 online). The genotype of the indel identified solely by SAMtools was discordant to that by Sanger sequencing verification. For WES, we verified 24 indels in total, 12 identified by VCMM and SAMtools, 11 by only SAMtools, and one by only GATK. For the examined candidates, seven out of eleven indels identified solely by SAMtools were false positives (see Supplementary Table S2 online). The genotypes of one indel identified by GATK, four by SAMtools and one by VCMM and SAMtools were discordant to those by Sanger sequencing verification.

To compare sensitivity of indel calling, we counted the common indels between those identified by each of the three methods and the Mills Indels<sup>22</sup>: VCMM: 202/509 (40%), GATK: 185/427 (43%), SAMtools: 228/1,764 (13%) in WES; VCMM: 55,810/106,732 (52%), GATK: 54,713/101,267 (54%), SAMtools: 59,356/275,112 (22%) in WGS. The sensitivity of GATK and VCMM was similar. For SAMtools, the number of the common indels was higher than the

Table 2 | Number of identified SNVs and indels

Number	WGS	WES
Total SNVs	3,406,875	79,060
Total indels *	763,944 (106,732)	10,999
Total SNVs in splice sites	105	56
Total SNVs in coding region	20,314	19,861
Missense	9,502	9,360
Nonsense	109	83
Synonymous	10,703	10,418
Total indels in coding region	461	509

\*: In the WGS, the numbers of indels in all region and non-repeat regions are shown.



**Figure 2 | Common indels identified by VCMM, GATK and SAMtools.** (a) SNV in WGS. SNVs in repeat regions and unknown contigs were not used for the comparison. (b) Indel in WGS. Indels in repeat regions and unknown contigs were not used for the comparison. (c) SNV in WES. (d) Coding indel in WES.

other two methods, but the proportion of the common indels was much lower, indicating that the false negative rate was lower while the false positive rate was higher. Appropriate filtering is required for accurate indel calling with SAMtools.

**Computational performance.** VCMM takes a pileup file, as generated by SAMtools, as input. Using one CPU core (2.67 GHz Intel Xeon Processor) on a computational cluster, the variant and indel calling for the largest contig (NT\_032977.9: 2 GB pileup formatted file) took approximately one and half hours for GATK and 3 minutes for pileup file generation by SAMtools and variant calling by VCMM for WES. VCMM is written in the C programming

language. The supporting programs, BWA<sup>19</sup> and SAMtools<sup>20</sup>, are also required.

## Discussion

In the human genome, SNVs and indels are the most abundant type of genetic variation. Accurate identification of SNVs and indels is one of the most important problems in genome analysis. Several variant calling methods and programs have been developed and used for both WGS and WES analysis<sup>20,23</sup>. However, reducing the false positive and false negative rate is still one of the most important challenges in sequencing analysis.

**Table 3 | Comparison of VCMM with other methods using SNP genotyping platforms**

	Chip		Number			Proportion (%)		
			VCMM	GATK	SAMtools	VCMM	GATK	SAMtools
WGS	OmniExpress BeadChip	Concordant	641,860	641,538	639,112	99.816	99.766	99.389
		FN	1,083	1,366	3,832	0.168	0.212	0.595
		FP	98	137	97	0.015	0.021	0.015
WES	Exome BeadChip	Concordant	190,147	190,137	189,825	99.974	99.968	99.804
		FN	36	46	361	0.019	0.024	0.190
		FP	14	14	11	0.007	0.007	0.006



In this study, we introduce a SNV and indel calling method. Our method is based on a multinomial probabilistic model as previously described<sup>18</sup>. Since base quality score (the probability that the called base in the read is the true sequenced base) reflects sequencing error rate as shown in a previous study<sup>24–26</sup>, a multinomial probabilistic model with quality score can be applied to identifying SNVs with high accuracy. Additionally, we applied a strand bias filter (see the Materials and Methods)<sup>27</sup>.

The two false positive SNV calls in the WGS analysis were suspected to be caused by mapping error, indicating that improvement in short read alignment methods should decrease false positives. Most false negative calls were identified in regions of low read coverage, and thus could be corrected by increasing the total coverage depth. This is also evident in the lower false negative rate observed in WES analysis as compared to WGS analysis.

We compared the performance of VCMM to that of existing methods, GATK and SAMtools, for SNV and indel calling. For SNV calling, the results from SAMtools had a larger number of false negatives than VCMM and GATK, suggesting that the variant calling of SAMtools is conservative. Although the false positive and false negative rates of VCMM were lower than these of GATK, the difference was not significant and we consider VCMM and GATK to be comparable in SNV calling. For indel calling, the proportion of commonly identified indels by the three methods was smaller than SNVs, suggesting greater difficulty in indel calling over SNV calling. Sanger sequencing verification revealed that coding indels that were detected by all three methods, or only VCMM and SAMtools, were higher quality than other indels. Indel verification also suggested that indel calls by GATK is more conservative than that of VCMM and SAMtools. Furthermore, although the false positive rate for indels detected only by SAMtools was higher than that of other methods, the false negative rate was lower, suggesting that further filtering is necessary for indel calling.

Although our method showed high concordance with DNA genotyping arrays, most of the SNPs present on the DNA genotyping array are located in uniquely mappable regions. It is unknown whether the observed false positive and false negative rates are applicable to all genomic regions. For variant calling, filtering with Hardy-Weinberg equilibrium, discarding variants in tandem repeat regions and a local realignment around multiple indels should be necessary for accurate variant detection<sup>18,28</sup>.

Our method incorporates a probabilistic error model, base quality filtering, and strand bias. Although our method can improve genotyping accuracy, it is still difficult to identify several types of variants, such as long insertions and deletions, as well as variations in highly repetitive regions. Continued advancement in sequencing technology, such as longer sequence reads and improvements in sequencing accuracy and mapping algorithms can be expected, by which the false positive and false negative rates of our method would further be improved. Additionally, while the target of the current version was deep sequenced single samples, we believe that our likelihood function can be expanded to multi-sample calling by considering population frequency as the prior probability.

## Methods

**DNA sample.** Samples, RK001 (WES) and RK130 (WGS), were obtained with consent and institutional ethics approval from RIKEN. High molecular weight genomic DNA was extracted from a human blood sample. All groups participating in this study approved this work. WES data of three samples were downloaded from NCBI FTP site (<ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/>) and used for the comparison.

**Whole exome and whole genome sequencing.** Exome capture was performed by the Agilent SureSelect Human All Exon V4 according to the manufacturer's instructions. This kit captures genomic DNA by in-solution hybridization with RNA oligonucleotides, enabling specific targeting regions of approximately 51 Mb of the human genome, contained in 185,636 capture regions, as large target regions were composed of several capture regions. The captured DNA was sequenced using the

Illumina HiSeq2000 platform with paired-end reads of 101 bp for insert libraries of 150–200 bp according to the manufacturer's instructions.

For the whole genome sequencing, we prepared 500 bp insert library. The sample was sequenced using the Illumina HiSeq2000 platform with paired-end reads of 101 bp according to the manufacturer's instructions.

**Read mapping.** Read sequences were mapped by the Burrows-Wheeler Aligner (BWA: version 0.6.1)<sup>19</sup> to the human reference genome (GRCh37.p5). Paired reads were mapped by considering not only the mapping distance between the paired reads within average  $\pm 3$  s.d, but also the mapping uniqueness and the orientation. The removal of the possible PCR duplicate reads and the conversion of the mapping results into pileup-formatted files were conducted using SAMtools (v.0.1.8)<sup>20</sup>. Note that the pileup-formatted files were generated by 'pileup' command of SAMtools (v.0.1.8), because several parameters required for SNV and indel calls with VCMM were removed by the 'mpileup' command in SAMtools (v0.1.17 (r973) and later).

**SNV and indel calls.** We performed SNV calling with respect to each nucleotide site using reads with a BWA mapping quality score  $\geq 20$  and a depth of 5 or more in WGS and 10 or more in WES. Both of the major and minor alleles were required to have at least one read with a base quality score  $\geq 30$ , as calculated by the Illumina pipeline software. The SNV calls were distinguished by the ratio of the probabilities that the minor allele at a nucleotide site is an error  $P_{error}$  and a major allele  $P_{allele}$  as described previously<sup>18</sup>. The two probabilities were calculated as follows.

$$P_{error} = \frac{n!}{\prod_i M_i! m_i!} \prod_i (1 - P_{E_i})^{M_i} (P_{E_i})^{m_i}$$

where  $P_{E_i} = 10^{-\frac{Q_i}{10}}$  is the probability that minor allele is an error, and  $M_i$  and  $m_i$  are the number of major and minor alleles with a base quality score  $i$ , and  $n$  is the total read depth ( $n = M_i + m_i$ ).

$$P_{allele} = \frac{n!}{\prod_i M_i! m_i!} \prod_i \left(\frac{1}{2}\right)^{M_i + m_i}$$

If  $\frac{P_{allele}}{P_{error}} > C$ , where  $C$  is the cut-off value, we detected a SNV at that nucleotide site.  $C$  was set into 5,000. We also considered a strand bias. If a minor allele described by over 30 reads in only one orientation, that minor allele was treated as an error. For WGS analysis, we discarded SNVs located within 5 bp of indels and included in short repeat regions (segmental duplications, simple repeat regions detected by tandem repeat finder and microsatellites), and regions where more than three SNVs occurred within 10 bp, as described<sup>18</sup>. This probabilistic model was applied to only SNV calling and not applied to indel calling.

Short indels (<50 bp) were identified on the basis of the gaps within read sequences from BWA. We performed indel calls at nucleotide sites, where more than 0.15 of total read sequences were mapped as indels after removing reads with mapping quality less than 20. The identified indels by the frequency was filtered by the following criteria; (1) average base quality of the base preceding the indel < 10, and (2) SAMtools consensus quality < 60 or SAMtools SNP quality < 60, if three or less reads support the indel. These parameter sets were determined by Sanger sequencing of indels, which was independent from verification in this study. For WGS, we discarded all indels and SNVs in simple repeat regions as defined by tandem repeat finder<sup>29</sup>.

**Accuracy evaluation.** In order to evaluate accuracy of our call method, we compared our SNV calls results with the concordant genotypes from SNP typing platforms: Illumina Human OmniExpress BeadChip and Illumina Human Exome BeadChip. We estimated concordance of genotype calls separately for homozygous and heterozygous SNPs. In total, 644,167 autosomal SNPs on the Illumina Human OmniExpress BeadChip for WGS and 193,280 autosomal SNPs on the Illumina Human Exome BeadChip on target regions were used for the estimation of the concordance rate in WGS and WES. The accuracy of our indel calls was estimated by the randomly selection of 96 indels and 47 indels in WGS and WES, respectively, and subsequent validation using Sanger sequencing.

**SNV calls using other methods.** We performed SNV calls using the GATK (v.1.6.13)<sup>23</sup> and SAMtools (v.0.1.8)<sup>20</sup> with the following parameters; minimum base quality > 30 and minimum mapping quality > 60 for GATK, and consensus quality  $\geq 20$  and root mean square (RMS)  $\geq 25^{30}$  for SAMtools. The commands for SAMtools and GATK were "samtools pileup -s -cf reference.fa bam\_file" and "java -jar GenomeAnalysisTK.jar -I INFO -T UnifiedGenotyper -R reference.fa -mbq 30 --read\_filter MappingQuality --min\_mapping\_quality\_score 60 -I bam\_file", respectively.

For indel calling in SAMtools, we examined candidate indels with the depth  $\geq 5$  for WGS and depth  $\geq 10$  for WES, and indels with a "\*"/\* genotype were excluded. Indels from GATK were identified with the following command; "java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R reference.fa -I bam\_file --out output -glm INDEL".

1. Londin, E. R. *et al.* Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics* **12**, 464 (2011).



2. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30–5 (2010).
3. Rusk, N. & Kiermer, V. Primer: Sequencing—the next generation. *Nat Methods* **5**, 15 (2008).
4. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31–46 (2010).
5. Mardis, E. R. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**, 387–402 (2008).
6. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–9 (2010).
7. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–7 (2010).
8. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–6 (2010).
9. Beroukhi, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
10. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat Genet* **44**, 760–4 (2012).
11. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* **44**, 623–30 (2012).
12. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096–101 (2009).
13. Chang, H. *et al.* Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PLoS One* **6**, e21097 (2011).
14. Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* **43**, 442–6 (2011).
15. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**, 539–42 (2011).
16. Agrawal, N. *et al.* Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**, 1154–7 (2011).
17. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36–46 (2012).
18. Fujimoto, A. *et al.* Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nat Genet* **42**, 931–6 (2010).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
20. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).
21. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745–55 (2011).
22. Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* **21**, 830–9 (2011).
23. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–303 (2010).
24. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175–85 (1998).
25. Brockman, W. *et al.* Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* **18**, 763–70 (2008).
26. Li, M., Nordborg, M. & Li, L. M. Adjust quality scores from alignment and improve sequencing accuracy. *Nucleic Acids Res* **32**, 5183–91 (2004).
27. Challis, D. *et al.* An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* **13**, 8 (2012).
28. Torri, F. *et al.* Next Generation Sequence Analysis and Computational Genomics Using Graphical Pipeline Workflows. *Genes (Basel)* **3**, 545–575 (2012).
29. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–80 (1999).
30. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–6 (2011).

## Acknowledgements

This work was supported partially by the National Project on 'Next-generation Integrated Living Matter Simulation' and Grant-in-Aid for Young Scientists (B) (Number: 23790389) of Ministry of Education, Culture, Sports, Science and Technology.

## Author contributions

D.S. and A.F. analyzed the data and wrote the manuscript; D.S. and A.F. contributed equally as first authors; M.K. performed the experiments of DNA genotyping arrays; K.N. and H.N. performed the experiments of next generation sequencing; K.N., M.F., U.Y. and H.N. performed Sanger sequencing verification; S.A., T.A. and K.A.B. provided the technical assistance; T.T. organized this work and wrote the manuscript. All authors contributed to and approved the final manuscript.

## Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Shigemizu, D. *et al.* A practical method to detect SNVs and indels from whole genome and exome sequencing data. *Sci. Rep.* **3**, 2161; DOI:10.1038/srep02161 (2013).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0>