PLOS ONE

RESEARCH ARTICLE

Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens

Jing Liu, Qingmiao Shen, Haigang Bao*

National Engineering Laboratory for Animal Breeding, Beijing Key Laboratory for Animal Genetic Improvement, College of Animal Science and Technology, China Agricultural University, Beijing, China

* zjbhg@126.com

Citation: Liu J, Shen Q, Bao H (2022) Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. PLoS ONE 17(1): e0262574. https://doi.org/10.1371/journal.pone.0262574

Editor: Shu-Biao Wu, University of New England, AUSTRALIA

Received: March 25, 2021

Accepted: December 29, 2021

Published: January 31, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0262574

Data Availability Statement: The DNA sequencing and genotyping data for this study can be downloaded from the China National GeneBank (Accession numbers: CNP0001419 and CNP0001435).

Abstract

Single nucleotide polymorphisms (SNPs) are widely used in genome-wide association studies and population genetics analyses. Next-generation sequencing (NGS) has become convenient, and many SNP-calling pipelines have been developed for human NGS data. We took advantage of a gap knowledge in selecting the appropriated SNP calling pipeline to handle with high-throughput NGS data. To fill this gap, we studied and compared seven SNP calling pipelines, which include 16GT, genome analysis toolkit (GATK), Bcftools-single (Bcftools single sample mode), Bcftools-multiple (Bcftools multiple sample mode), VarScan2-single (VarScan2 single sample mode), VarScan2-multiple (VarScan2 multiple sample mode) and Freebayes pipelines, using 96 NGS data with the different depth gradients of approximately 5X, 10X, 20X, 30X, 40X, and 50X coverage from 16 Rhode Island Red chickens. The sixteen chickens were also genotyped with a 50K SNP array, and the sensitivity and specificity of each pipeline were assessed by comparison to the results of SNP arrays. For each pipeline, except Freebayes, the number of detected SNPs increased as the input read depth increased. In comparison with other pipelines, 16GT, followed by Bcftools-multiple, obtained the most SNPs when the input coverage exceeded 10X, and Bcftools-multiple obtained the most when the input was 5X and 10X. The sensitivity and specificity of each pipeline increased with increasing input. Bcftools-multiple had the highest sensitivity numerically when the input ranged from 5X to 30X, and 16GT showed the highest sensitivity when the input was 40X and 50X. Bcftools-multiple also had the highest specificity, followed by GATK, at almost all input levels. For most calling pipelines, there were no obvious changes in SNP numbers, sensitivities or specificities beyond 20X. In conclusion, (1) if only SNPs were detected, the sequencing depth did not need to exceed 20X; (2) the Bcftools-multiple may be the best choice for detecting SNPs from chicken NGS data, but for a single sample or sequencing depth greater than 20X, 16GT was recommended. Our findings provide a reference for researchers to select suitable pipelines to obtain SNPs from the NGS data of chickens or nonhuman animals.

Funding: This study was supported by the Modern Agricultural Industry Technology System of China [grant number CARS-40]. The funder did not play any role in the design of the study, collection, analysis, interpretation of data or writing the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: Bcftools-multiple, Bcftools multiple sample mode; Bcftools-single, Bcftools single sample mode; FP, the false genotype with false positive SNPs; GATK, genome analysis toolkit; GE, the false genotype with true positive SNPs; MG, the missing genotypes from sequencing data at the positive array sites; NGS, next-generation sequencing; SNP, single nucleotide polymorphisms; Ti/Tv, transition/transversion ratio; TP, the true genotype with true positive SNPs; VarScan2-multiple, VarScan2 multiple sample mode; VarScan2-single, VarScan2 single sample mode.

Introduction

In the last decade, next-generation sequencing (NGS) has been extensively used in human, livestock and plant research [1–5]. An increasing number of single nucleotide polymorphisms (SNPs) have been detected in NGS datasets using various calling pipelines [6–8]. SNPs might occur at nonspecific positions in the genome and have been widely used in genome-wide association studies and population genetics analyses [9]. Many SNPs related to complex diseases or traits in humans or animals have been discovered by whole-genome sequencing and whole-exome sequencing [10]. Some SNPs have been shown to be causal mutations of some traits or diseases [11,12].

Many variant calling pipelines have been developed to detect SNPs from NGS data; however, each pipeline has its own advantages and disadvantages [13]. The genome analysis toolkit (GATK, https://software.broadinstitute.org/gatk/) [14] and Bcftools (https://samtools.github.io/bcftools/bcftools.html) [15] may be the most widely used SNP calling pipelines to date. A brief characteristic summary of several calling tools is listed in Table 1 and described as follows. GATK was originally used to analyze human genome and exome sequencing data, and now it may be regarded as the industry standard for identifying SNPs in germline DNA and RNA NGS data [14]. The toolkit contains a wide variety of tools with a primary focus on variant discovery and genotyping. Bcftools is a high-speed program for calling variants. It can manipulate variant calls in compressed/uncompressed VCF and BCF files [15]. VarScan2 (http://varscan.sourceforge.net/using-varscan.html) is the first tool used for the detection of somatic mutations and copy number alterations in exome data from tumor-normal pairs [16]. The VarScan2 algorithm reads the SAMtools pileup or mpileup output of tumor and normal samples simultaneously, performs pairwise comparisons of base calls, and normalizes sequencing depths at each position [17]. Freebayes (https://github.com/ekg/freebayes) is a Bayesian genetic variant caller designed to find SNPs, indels, multinucleotide polymorphisms, and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment [18]. Freebayes uses short-read alignments for any number of individuals from a population and uses a reference genome to determine the most likely combination of genotypes at each position in the population [18]. 16GT (https://github.com/aquaskyline/16GT) is the first publicly available caller that uses a 16-genotype probabilistic model to unify SNPs and indel calling in a single algorithm [19]. Compared with the traditional 10-genotype probabilistic model, 16GT added 6 new genotypes. Compared to GATK with HaplotypeCaller, 16GT not only runs 4 times faster but also improves sensitivity in calling SNPs by unifying SNPs and indel calling in a single algorithm of variant calling. Recently, Chiara et al. also provided a consensus variant calling system, CoVaCS (https://bioinformatics.cineca.it/covacs), for the analysis of human genome resequencing studies [20].

Table 1. A brief summary of different tools.

caller	Bcftools	16GT	Freebayes	VarScan2	GATK
Code	C	Perl	C++	Java	Java
Model	HMM & MAQ	16-genotype probabilistic	Bayesian	heuristic algorithm	Bayesian
Sampling	Single & multiple	Single	Single	Single & multiple	Single & multiple
Variants	SNPs & indels	SNPs & indels	SNPs & indels&MNPs	SNPs & indels	SNPs & indels
Features	Sorting, indexing, etc.	easy to use, timesaving	straightforward	meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance	Realignment, per base recalibration, VQSR
Reference	Danecek et al., 2017 [15]	Luo et al., 2017 [19]	Garrison and Marth, 2012 [18]	Koboldt et al., 2012 [16]	Mckenna et al., 2010 [14]

https://doi.org/10.1371/journal.pone.0262574.t001

Using simulation and real NGS data of humans, many studies have shown that different tools have their own advantages and disadvantages [6,8,12,21]. Different variant callers may produce different results, so ensemble methods of variant calling algorithms or analytic pipelines can improve variant accuracy [22,23]. However, a single pipeline, such as the pipelines of BWA-MEM and GATK-HaplotypeCaller, can be run similarly to the pipeline ensemble method [23]. GATK may be the most popular pipeline for detecting SNPs from human high-throughput data sets [24], and it has also been widely used in chicken NGS data in recent studies [25–27]. Compared with known human variant information resources, the corresponding resources of chickens are quite few, which may affect the detection results if we use GATK to detect SNPs from chicken data. Ni et al. [7] compared variants detected with GATK (UnifiedGenotyper and hard filtering), Freebayes, and SAMtools using chicken NGS data with an average coverage of 7.6 X and found that all three pipelines, particularly GATK and SAMtools, perform well in general. In the present study, we used NGS data from 16 Rhode Island Red chickens to evaluate seven SNP calling pipelines, including 16GT, GATK, Bcftools-single (Bcftools single sample mode), Bcftools-multiple (Bcftools multiple sample mode), VarScan2--single (VarScan2 single sample mode), VarScan2-multiple (VarScan2 multiple sample mode), and Freebayes, in terms of the number of detected SNPs, sensitivity, and specificity. We aim to select a high-performance SNP calling pipeline for chicken NGS data studies.

Materials and methods

Ethics statement

All experimental procedures and animals used were approved by the Ethics Review Committee for Laboratory Animal Welfare and Animal Experiment of China Agricultural University (Approval number: AW70101202-1-1).

Animals and DNA samples

The animal experimental process complied with the regulations and guidelines of the Experimental Animal Welfare and Animal Experiment Ethics Review Committee of China Agricultural University. A total of 16 chickens at 18 weeks of age randomly selected from the Rhode Island Red population, and blood samples were collected from each chicken's wing vein using 2 mL injectors. After blood was collected, we put the 16 chickens back to the population and keep them with other individuals reared in the Experimental Chicken Farm of China Agricultural University. Our subsequent research did not work with animals. Genomic DNA of blood was extracted using the TIANamp Genomic DNA Kit (Cat. #DP304-02, TIANGEN) according to the protocol supplied. After checking and qualification, each DNA sample was divided into two parts, one part for next-generation sequencing (paired-end sequencing, 150 bp, 50X, Illumina HiSeq™ 4000, Beijing Novogene Bioinformatics Technology Co., Ltd) and the other for SNP array analyses (50K, KPS CAULayer Breeding Chip v1, Beijing Compass Biotechnology Co., Ltd, S1 Table).

NGS data sets and SNP calling pipelines

Cleaned reads were obtained by Trimmomatic (version 0.39; S1 Word) from raw sequencing data. After quality control, the cleaned data of each of the 16 samples were split into 10 parts evenly and reorganized to form 6 subsets of various sequencing depth gradients of approximately 5X, 10X, 20X, 30X, 40X, and 50X coverage according to Bentley et al. [28]. Thus, we finally had 16 samples × 6 gradients = 96 data points. Bowtie 2 [29] was chosen as the common aligner with the chicken genome reference (Gallus_gallus-5.0) for all SNP calling pipelines in

the present study. We conducted alignment with Bowtie 2, converted the SAM files to BAM files, and then processed the same BAM files with seven SNP calling pipelines, including 16GT, GATK, Bcftools-single, Bcftools-multiple, VarScan2-single, VarScan2-multiple and Freebayes. All results of this study depended on programs' defaults in each pipeline. Details of processing with all these pipelines are described in S1 Word.

Analysis of the sensitivity and specificity of SNP-calling pipelines

We compared the SNP array genotypes with the genotypes of SNP loci in the array detected by sequencing pipelines. In order to assess the sensitivity, and specificity of the pipelines with input read depth gradients of 5X-50X coverage, SNP loci in the array that were also detected from sequencing data for each individual were divided into 4 categories (Table 2) referring to Liu et al. [6] as follows: (1) sequencing SNPs with matched array genotypes (the true genotype with true positive SNPs (TP)); (2) false genotypes from sequencing data at the matched positive array sites (the false genotype with true positive SNPs (GE)); (3) false genotypes from sequencing data with negative array genotypes (the false genotype with false positive SNPs (FP)); and (4) the missing genotypes from sequencing data at the positive array sites (MG). Four metrics, including the SNP number, sensitivity, specificity and transition/transversion ratio (Ti/Tv), were used to assess the performance of each SNP calling pipeline. The SNP number indicates the number of detected SNPs in each sample at any input read depth. The sensitivity of each pipeline was calculated as (TP + GE)/(TP + GE + MG), and the specificity was calculated as TP/(TP + FP + GE). The Ti/Tv ratios were calculated using VCFtools (Version 0.1.17) [30].

Statistical analysis

Means and standard errors were calculated for the SNP number, sensitivity and specificity of each pipeline at each input level. Mean differences were tested by the Duncan test of SPSS 19.0 (SPSS Inc., Chicago, IL), and the statistical significance level was set at $P < 0.05$.

Results

The NGS data sets and alignment

Approximately 3.5 billion paired-end cleaned data reads were obtained with an average coverage of approximately 50X for each sequenced Rhode Island Red chicken (S2 Table). The cleaned data set of each sample was split into 10 parts evenly and reorganized, and we obtained a total of 96 data sets. Each sample had 6 data sets with different coverages of approximately 5X, 10X, 20X, 30X, 40X and 50X (S3 Table). Paired-end cleaned reads were aligned against the chicken reference genome (Gallus_gallus-5.0) using Bowtie 2 (version 2.2.9). A summary of cleaned data alignments is displayed in S3 Table. The alignment rate of the cleaned data of each sample was between 90.91% and 95.21% (S3 Table).

Table 2. Descriptions of genotype categories.

Genotype categories		Genotype from SNP array		
		00	01	11
Genotype from sequencing data	01	FP	TP, MG	GE
	11	FP	GE	TP, MG

*Notes: TP means sequencing SNPs with matched array genotypes (The true genotype with true positive SNPs); GE means false genotypes from sequencing data at the matched positive array sites (The false genotype with true positive SNPs); FP means false genotypes from sequencing data with negative array genotypes (the false genotype with false positive SNPs) and MG means the missing genotypes from sequencing data at the positive array sites.

https://doi.org/10.1371/journal.pone.0262574.t002

Comparisons of the numbers of SNPs detected by different SNP calling pipelines

The numbers of SNPs detected with different input read depths are shown in Fig 1 and S4 Table. From Fig 1B, we could see that an increasing number of SNPs were detected with increasing input read depths by each variant caller except Freebayes. When the sequencing depth was less than 20X, the number of SNPs found by any caller increased rapidly with increasing sequencing depth, while when the sequencing depth was greater than 20X, the speed of increase slowed down obviously, and Freebayes even reached the maximum at 20X (Fig 1B). In comparison with other callers, 16GT obtained the most abundant SNPs at almost all input read depths (except 5X) in the present study; VarScan2-single and VarScan2-multiple obtained the same SNP numbers at all input read depths, and both called out the fewest SNPs at low sequencing depths (< 20X), while Freebayes called the fewest SNPs at high sequencing depths (> = 20X), and GATK and Bcftools-single performed moderately (Fig 1A). From Fig 1A, we could also see that Bcftools-multiple obtained the most abundant SNPs at 5X and 10X input levels, and at high input depths (> = 20X), Bcftools-multiple also obtained higher SNP numbers in comparison with any other pipeline except 16GT.

Comparisons of the sensitivity and specificity among the seven SNP calling pipelines

To assess the sensitivity, and specificity of each pipeline with different input read depths, a 50K chicken SNP array (KPS CAULayer Breeding Chip v1, Beijing Compass Biotechnology Co.,

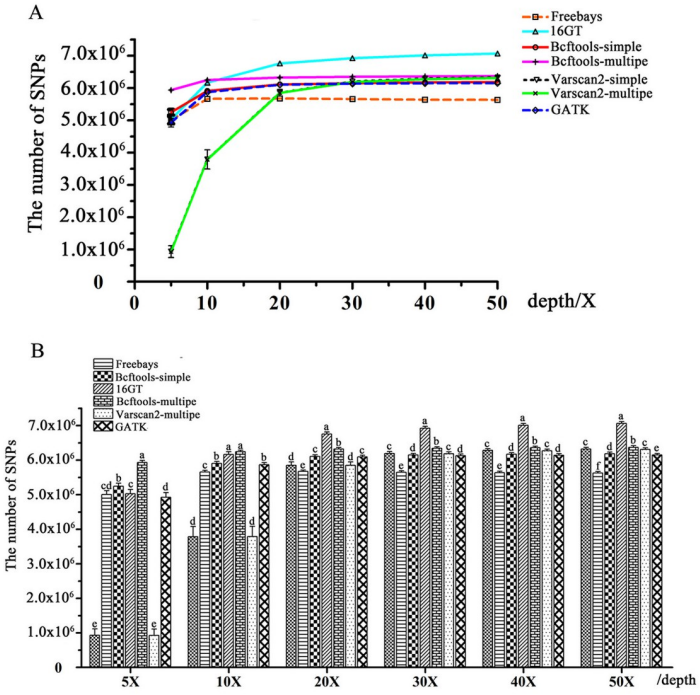

Fig 1. Comparisons of the total number of SNPs called out by seven different SNP calling pipelines. A: Comparisons of the number of SNPs called out by different calling pipelines at each input read depth level. For each input level, the same letters indicate that the difference is not significant ($P > 0.05$), and the different letters indicate significant differences ($P < = 0.05$). B: The tendency of the number of SNPs called out by each pipeline with increasing input level.

https://doi.org/10.1371/journal.pone.0262574.g001

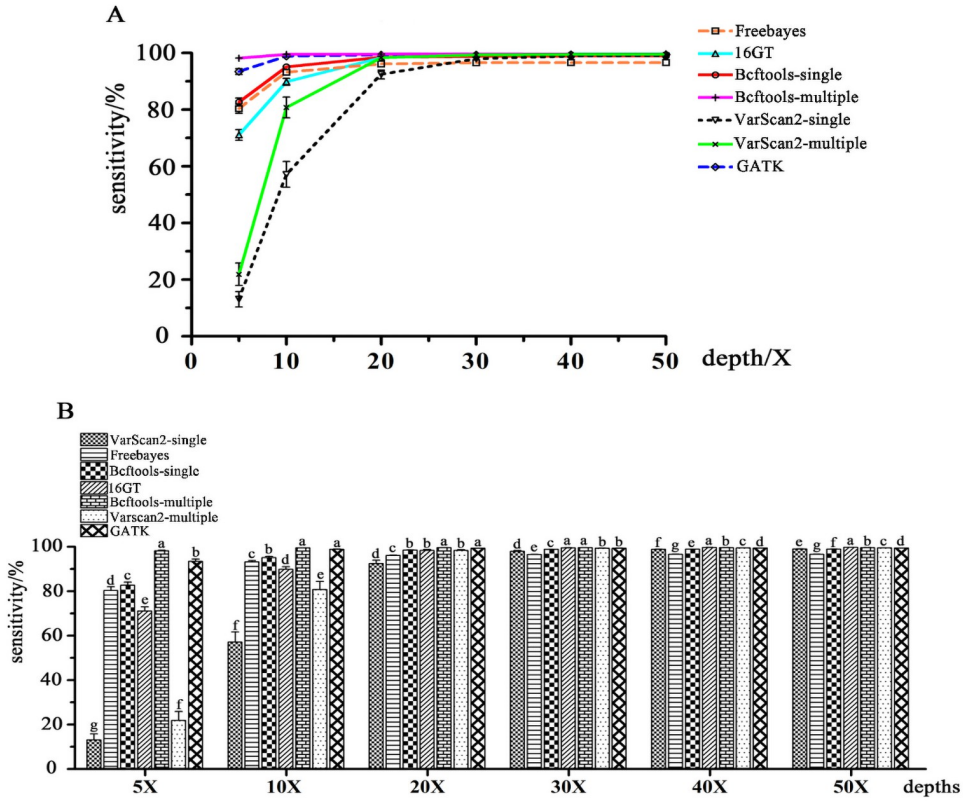

Fig 2. The sensitivities of seven SNP calling pipelines. A: The sensitivity tendencies of each SNP calling pipeline with the input level increasing; B: Comparisons of the sensitivities of different calling pipelines at each input read depth level. For each input level, the same letters indicate that the difference is not significant ($P > 0.05$), and different letters indicate significant differences ($P < = 0.05$).

https://doi.org/10.1371/journal.pone.0262574.g002

Ltd, Beijing, China) with a total of 43,681 SNP sites (S1 Table) was used to genotype individuals. We compared the SNP array genotypes with the genotypes of SNP loci in the array detected by sequencing pipelines, and the array results were regarded as a standard to evaluate the specificity and sensitivity of each calling pipeline. The array results showed an average call rate of 99.20% (S5 Table).

The sensitivity of each pipeline is displayed in Figs 2 and 4 and S6 Table. As shown in Fig 2, the sensitivity of various pipelines tended to rapidly increase at lower input read depths and then slightly increase at higher input read depths with increasing sequencing depth. In comparison with any other pipeline in the present study, 16GT had higher sensitivity when input read depths were equal to or greater than 20X, and Freebayes showed its sensitivity moderately at lower sequencing depths (< = 20X) but the lowest from 30X to 50X. The two VarScan2 pipelines displayed the lowest sensitivity but increased rapidly at the low input read depths and then tended to stabilize. In Fig 2, Bcftools-multiple showed the best sensitivity from 5X to 30X input depths and was then exceeded by 16GT. GATK and Bcftools-multiple both showed the best sensitivity at 10X and 20X input depths, as shown in Fig 2B.

The differences in specificity among the seven pipelines were similar to the differences in sensitivity among them. Fig 3 and S7 Table show the specificities of the seven SNP calling pipelines at different input depths for SNP calling. From Fig 3, we observed that the specificity of each pipeline increased as the input read depth increased. In comparison with any other calling pipeline in the present study, Bcftools-multiple had higher specificity with any input read

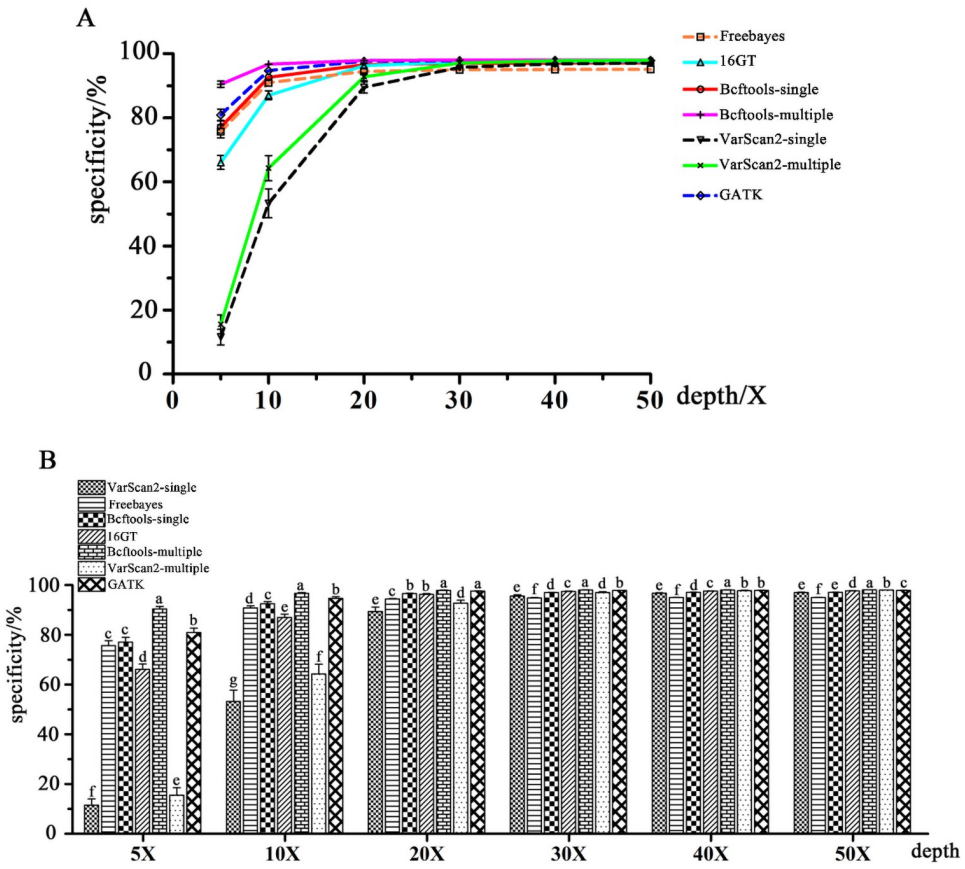

Fig 3. The specificities of seven SNP calling pipelines. A: The specificity tendencies of each SNP calling pipeline with the input level increasing; B: Comparisons of the specificities of different calling pipelines at each input read depth level. The same letter indicates that the difference is not significant ($P > 0.05$), and different letters indicate significant differences ($P < = 0.05$).

https://doi.org/10.1371/journal.pone.0262574.g003

depth in the present study (Fig 3B). 16GT showed moderate specificity at any read depth. Compared with other pipelines, the two VarScan2 pipelines displayed the lowest specificity, but it increased rapidly at the low input read depths (< = 20X), while Freebayes showed the lowest specificity at the high input read depths (> = 30X). GATK had better specificity than any other pipeline at 5X to 40X input read depths except Bcftools-multiple in the present study.

Two-dimensional scatter plots with the specificities and sensitivities of seven SNP calling pipelines in different input read depths are displayed in Fig 4. From Fig 4, we can see that Bcftools-multiple may be the best pipeline in most cases considering both sensitivity and specificity.

Effects of single and multiple modes on the sensitivity and specificity of Bcftools and VarScan2 Pipelines

Bcftools and VarScan2 can process files one by one (Bcftools-single and VarScan2-single pipelines) or multiple files once a time (Bcftools-multiple and VarScan2-multiple pipelines). From Fig 5, we could see that the sensitivity and specificity of calling procedures increased with increasing input read depth whether in a one-by-one way or multiple files a time. Bcftools-multiple and VarScan2-multiple had higher sensitivity and specificity than Bcftools-single and VarScan2-single, respectively (Fig 5; S6 and S7 Tables). Especially at low input read depths,

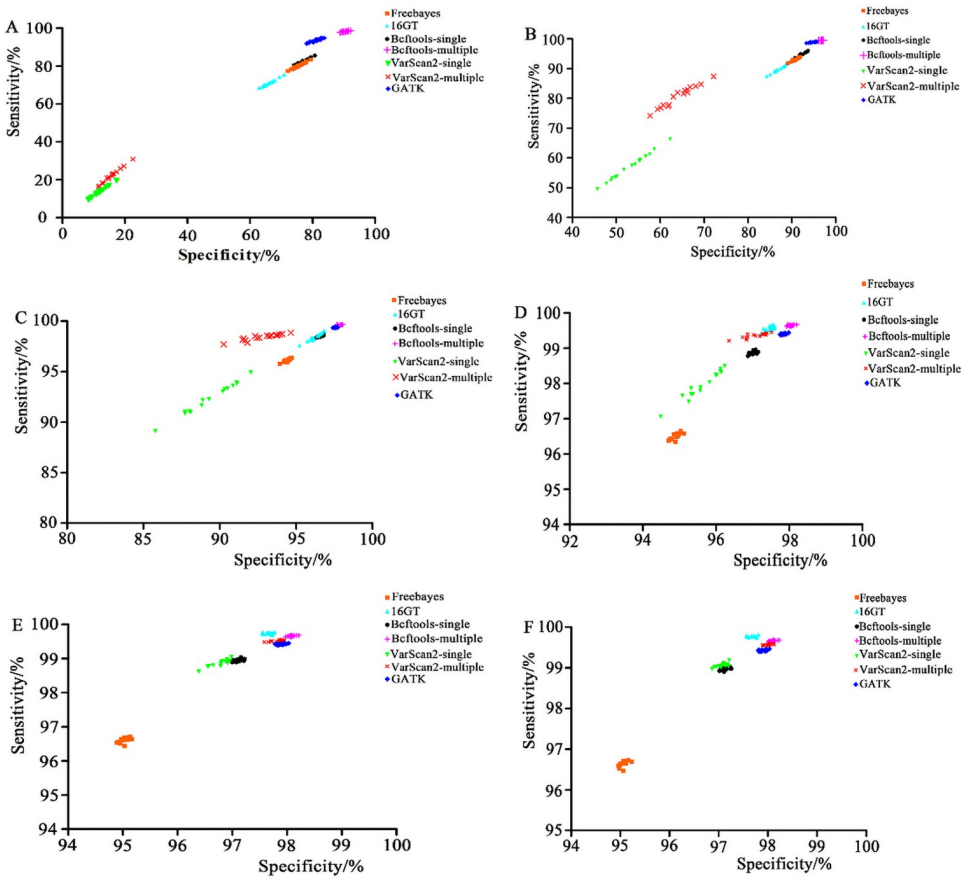

Fig 4. Two-dimensional scatter plots with specificities and sensitivities of each pipeline at different input read depths. A, The input read depth is 5X; B, 10X; C, 20X; D, 30X; E, 40X; and F, 50X.

https://doi.org/10.1371/journal.pone.0262574.g004

Bcftools-multiple considerably improved the specificity and sensitivity of the detection in comparison with Bcftools-single. For example, under the condition of a 5X input read depth, the specificity increased from 0.771 to 0.905, and the sensitivity increased from 0.827 to 0.982. VarScan2-multiple also improved the performance but not Bcftools-multiple (Fig 5).

Comparisons of the Ti/Tv ratios of each predictor with different input read depths

The Ti/Tv ratios of each predictor with different input read depths are shown in Fig 6 and S8 Table. From Fig 6, we can see that all Ti/Tv values are between 2.04 and 2.44. No significant ($P < = 0.05$) differences in the ratios were observed among the pipelines with the same input read depths, and among different coverages using the same pipelines in this study. The absolute value of the deviation between the Ti/Tv ratios of the maximum and minimum values in each pipeline did not exceed 0.2, and the absolute deviations of the Ti/Tv ratios of the maximum and minimum values of different pipelines with the same input read depths were less than 0.4 (Fig 6 and S8 Table).

Discussion

SNPs are widely used in functional gene mapping and population genetics [9,31,32]. As the cost of high-throughput sequencing declined, detecting SNPs from NGS data became

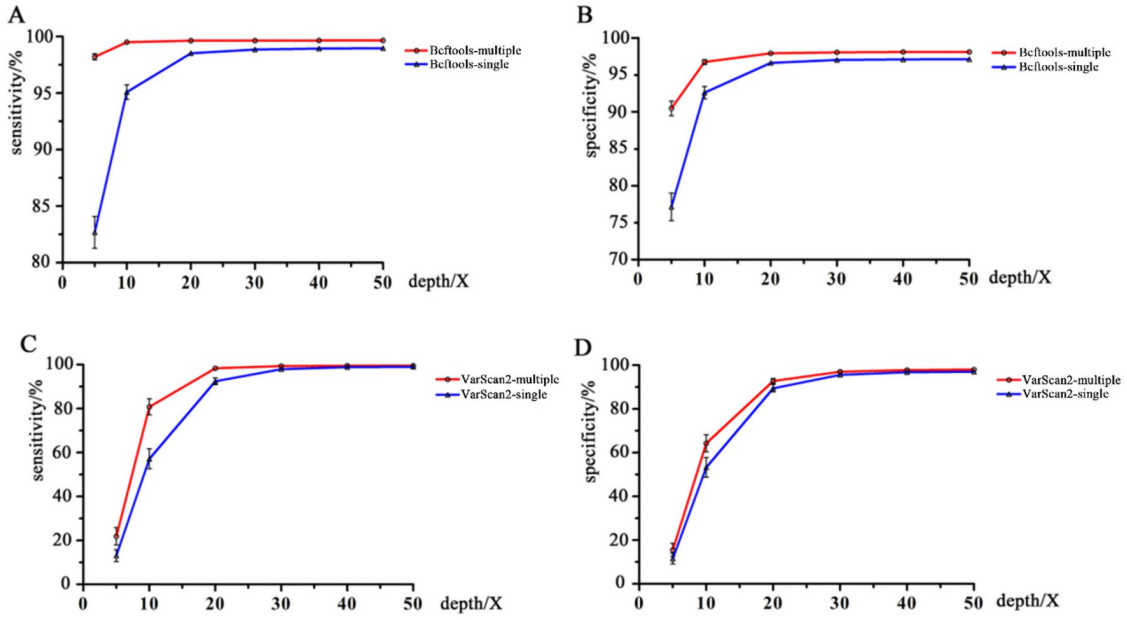

Fig 5. Comparisons of the sensitivity and specificity of Bcftools and VarScan2 with different sample modes. A: Comparisons of the sensitivity between Bcftools-single and Bcftools-multiple; B: Comparisons of the specificity between Bcftools-single and Bcftools-multiple; C: Comparisons of the sensitivity between VarScan2-single and VarScan2-multiple; and D: Comparisons of the specificity between VarScan2-single and VarScan2-multiple.

https://doi.org/10.1371/journal.pone.0262574.g005

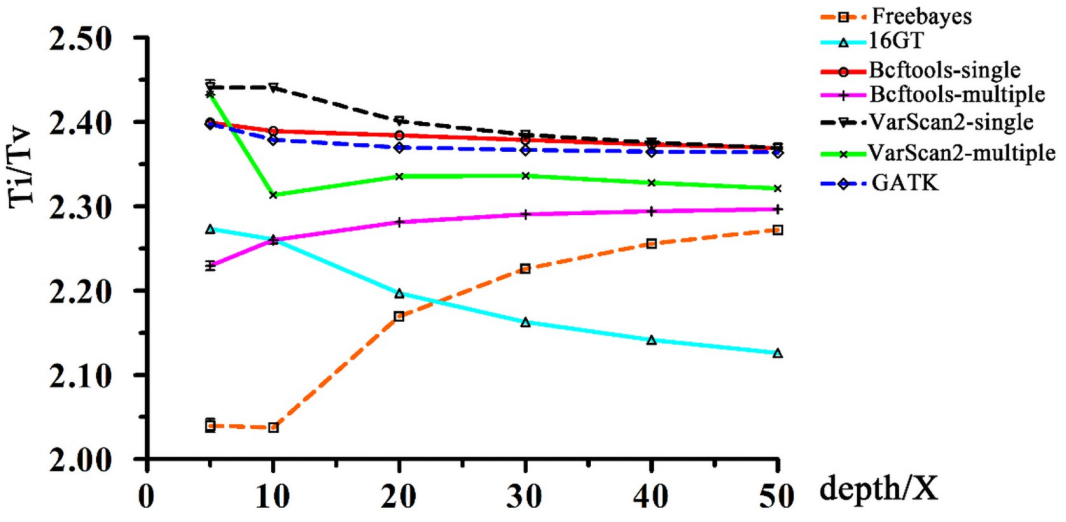

Fig 6. The transition/transversion ratios of each predictor with different input read depths.

https://doi.org/10.1371/journal.pone.0262574.g006

increasingly common. Generally, NGS data are initially aligned to a reference genome and then subjected to variant calling. Bowtie 2 was chosen to map short reads in the present study since it has a high speed, sensitivity, and accuracy and was particularly good at aligning reads to relatively large genomes (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) [29]. Many previous studies have reported the capabilities of several available SNP

calling pipelines from NGS data, which were often applied to human data or simulated data [33–36]. GATK is often regarded as the most effective procedure to detect variants from NGS data using resources of known variations, truth sets and other metadata (https://software.broadinstitute.org/gatk/best-practices/about). However, we have fewer known variation resources in poultry than in humans or mice, which may lead to the reduced accuracy of GATK. Ni et al. [7] thought that GATK, SAMtools and Freebayes were all good for processing high-throughput chicken data, but we found that the research in the article used low sequencing depth data, tested relatively few pipelines, and lacked detailed implementation procedures. Thereby, further research was needed. In the present study, we compared the seven SNP calling procedures using 96 NGS datasets with different input read depths of 5X-50X coverage of Rhode Island Red chickens. Luo et al. [19] found that 16GT not only ran fast but also showed the highest sensitivity and specificity in calling SNPs among all tools (GATK UnifedGenotyper, GATK HaplotypeCaller, Freebayes, Fermikit, ISAAC, and VarScan2). In our study, we also found that 16GT was more sensitive than any other pipeline at input read depths ranging from 30X to 50X (Figs 2 and 4), but the specificity of 16GT was moderate (Figs 3 and 4). Freebayes was easy to operate and could be run in one step [18]. However, Freebayes may not be a good pipeline to call SNPs from the short read data sets of the 16 Rhode Island Red chickens due to its unremarkable performances in SNP calling (Figs 1–4). GATK is a popular toolkit and is widely used in many studies [6,37–41]. In our study, the GATK performance was not bad, but at whatever input depth, Bcftools-multiple, and sometimes 16GT, always showed better detection performances than GATK (Figs 1–4). Therefore, we did not recommend GATK for detecting SNPs from chicken NGS data.

A large number of SNPs were detected out by next-generation sequencing, however, we could not evaluate the accuracy of all SNP loci. In order to evaluate the sensitivity and specificity of each SNP calling pipeline, we compared the SNP array genotypes with the genotypes of SNP loci in the array detected by sequencing pipelines with different input read depths, and regarded the array genotyping as the reference data set which were distributed evenly throughout the whole chicken genome. In the present study, 16 chickens were genotyped with the 50K SNP array, and the result was regarded as a standard to evaluate the specificity and sensitivity of each SNP calling pipeline. Since the reference data only consisted of a subset of all SNPs in the genome, the estimated specificity and sensitivity here might differ from the actual values.

The Ti/Tv ratio is also an index used to evaluate the accuracy of SNP calling [40]. A high Ti/Tv ratio (> 2.0) often indicates a high-accuracy SNP set, whereas a low value (~ 0.5) implies low-quality SNP calling [42]. In our study, although each pipeline has a higher or lower value of the Ti/Tv ratio in each different input read depth, all the Ti/Tv ratios fall in the range of 2.04–2.44 (Fig 6, S8 Table), which can be considered as high accurate [42]. Moreover, the Ti/Tv ratio of each pipeline except 16GT approach slowly to around 2.3 with the increase of input read depth (Fig 6, S8 Table), and we speculate that the Ti/Tv = 2.3 could be a genome-wide approximation of chicken in this study.

Conclusions

In conclusion, (1) if only SNPs were detected, the sequencing depth did not need to exceed 20X since there were no obvious changes in the number of SNPs, sensitivity or specificity beyond 20X. (2) Bcftools-multiple may be the best choice to detect SNPs from chicken NGS data, but for a single sample or a sequencing depth greater than 20X, 16GT was also recommended. Our findings provide a reference for researchers to select suitable pipelines to obtain SNPs from the NGS data of chicken or nonhuman animals.

Supporting information

S1 Table. The genotyped results of the Illumina 50 K SNP Beadchip.
(XLS)

S2 Table. The sequencing results of 16 Rhode Island Red chickens.
(XLSX)

S3 Table. The coverage and alignment rate of each sample.
(XLSX)

S4 Table. The total number of SNPs called out by each pipeline in different input depths.
(XLSX)

S5 Table. The call rate results of array.
(XLSX)

S6 Table. The sensitivity of each pipeline in different input depths.
(XLSX)

S7 Table. The specificity of each pipeline in different input depths.
(XLSX)

S8 Table. The Ti/Tv ratios of 7 pipelines.
(XLSX)

S1 Word. SNP calling pipelines for chicken NGS sets.
(DOCX)

Acknowledgments

We wish to thank Wenpeng Han for his help in the experimental methods and polishing of this manuscript during our study.

Author Contributions

Conceptualization: Haigang Bao.

Data curation: Jing Liu, Qingmiao Shen, Haigang Bao.

Formal analysis: Jing Liu, Haigang Bao.

Funding acquisition: Haigang Bao.

Investigation: Jing Liu, Haigang Bao.

Methodology: Jing Liu, Qingmiao Shen.

Software: Jing Liu, Haigang Bao.

Supervision: Haigang Bao.

Validation: Jing Liu, Haigang Bao.

Visualization: Jing Liu, Haigang Bao.

Writing – original draft: Jing Liu, Qingmiao Shen.

Writing – review & editing: Jing Liu, Haigang Bao.

References

1. Wang BB, Zhang YB, Zhang F, Lin HB, Wang XM, Wan N, et al. On the origin of Tibetans and their genetic basis in adapting high-altitude environments. PloS One. 2011; 6 (2): e17002. https://doi.org/10.1371/journal.pone.0017002 PMID: 21386899
2. Gholami M, Erbe M, Gärke C, Preisinger R, Weigend A, Weigend S, et al. Population genomic analyses based on 1 million SNPs in commercial egg layers. PloS One. 2014; 9 (4): e94509. https://doi.org/10.1371/journal.pone.0094509 PMID: 24739889
3. Liu L, Wang MN, Feng JY, See DR, Chao SM, Chen XM. Combination of all-stage and high-temperature adult-plant resistance QTL confers high-level, durable resistance to stripe rust in winter wheat cultivar Madsen. Theor Appl Genet. 2018; 131 (9): 1835–1849. https://doi.org/10.1007/s00122-018-3116-4 PMID: 29797034
4. Rochus CM, Tortereau F, Plisson-Petit F, Restoux G, Moreno-Romieux C, Tosser-Klopp G, et al. Revealing the selection history of adaptive loci using genome-wide scans for selection: an example from domestic sheep. BMC Genomics, 2018; 19 (1): 71. https://doi.org/10.1186/s12864-018-4447-x PMID: 29357834
5. Zhang MJ, Ren WZ, Sun XJ, Liu Y, Liu KW, Ji ZH, et al. GeneChip analysis of resistant Mycobacterium tuberculosis with previously treated tuberculosis in Changchun. BMC Infect Dis. 2018; 18 (1): 234. https://doi.org/10.1186/s12879-018-3131-8 PMID: 29788948
6. Liu XT, Han SZ, Wang ZH, Gelernter J, Yang B.Z. Variant callers for next-generation sequencing data: a comparison study. PloS One. 2013; 8 (9): e75619. https://doi.org/10.1371/journal.pone.0075619 PMID: 24086590
7. Ni GY, Strom TM, Pausch H, Reimer C, Preisinger R, Simianer H, et al. Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. BMC Genomics. 2015; 16 (1): 824. https://doi.org/10.1186/s12864-015-2059-2 PMID: 26486989
8. Sandmann S, De Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen JH, et al. Evaluating variant calling tools for non-matched next-generation sequencing data. Sci Rep. 2017; 7: 43169. https://doi.org/10.1038/srep43169 PMID: 28233799
9. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, Limborg MT, et al. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Mol Ecol Resour. 2011; 11 Suppl 1: 123–36. https://doi.org/10.1111/j.1755-0998.2010.02943.x PMID: 21429169
10. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annu Rev Med. 2012; 63: 35–61. https://doi.org/10.1146/annurev-med-051010-162644 PMID: 22248320
11. Guo YF, Ding XL, Shen YF, Lyon GJ, Wang K. SeqMule: automated pipeline for analysis of human exome/genome sequencing data. Sci Rep. 2015; 5: 14283. https://doi.org/10.1038/srep14283 PMID: 26381817
12. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015; 5: 17875. https://doi.org/10.1038/srep17875 PMID: 26639839
13. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014; 15 (2): 256–78. https://doi.org/10.1093/bib/bbs086 PMID: 23341494
14. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20 (9): 1297–303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199
15. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017; 33 (13): 2037–9. https://doi.org/10.1093/bioinformatics/btx100 PMID: 28205675
16. Koboldt DC, Zhang QY, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22 (3): 568–76. https://doi.org/10.1101/gr.129684.111 PMID: 22300766
17. Koboldt DC, Larson DE, Wilson RK. Using VarScan 2 for germline variant calling and somatic mutation detection. Curr Protoc Bioinformatics. 2013; 44: 15.4.1–17. https://doi.org/10.1002/0471250953.bi1504s44 PMID: 25553206
18. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907V2. 2012; arxiv.org/abs/1207.3907.
19. Luo RB, Schatz MC, Salzberg SL. 16GT: a fast and sensitive variant caller using a 16-genotype probabilistic model. Gigascience. 2017; 6 (7):1–4. https://doi.org/10.1093/gigascience/gix045 PMID: 28637275

20. Chiara M, Gioiosa S, Chillemi G, D'Antonio M, Flati T, Picardi E, et al. CoVaCS: a consensus variant calling system. BMC Genomics. 2018; 19 (1):120. https://doi.org/10.1186/s12864-018-4508-1 PMID: 29402227

21. Escalona M, Rocha S, Posada D. A comparison of tools for the simulation of genomic next-generation sequencing data. Nat Rev Genet. 2016; 17 (8): 459–69. https://doi.org/10.1038/nrg.2016.57 PMID: 27320129

22. Gézsi A, Bolgár B, Marx P, Sarkozy P, Szalai C, Antal P. VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. BMC Genomics. 2015; 16: 875. https://doi.org/10.1186/s12864-015-2050-y PMID: 26510841

23. Hwang KB, Lee IH, Li H, Won DG, Hernandez-Ferrer C, Negron JA, et al. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. Sci Rep. 2019; 9 (1): 3219. https://doi.org/10.1038/s41598-019-39108-2 PMID: 30824715

24. do Valle ÍF, Giampieri E, Simonetti G, Padella A, Manfrini M, Ferrari A, et al. Optimized pipeline of MuTect and GATK tools to im-prove the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. BMC Bioinformatics. 2016; 17 Suppl 12: 341. https://doi.org/10.1186/s12859-016-1190-7 PMID: 28185561

25. Lawal RA, Al-Atiyat RM, Aljumaah RS, Silva P, Mwacharo JM, Hanotte O. Whole-genome resequencing of red junglefowl and indigenous village chicken reveal new insights on the genome dynamics of the species. Front Genet. 2018; 9: 264. https://doi.org/10.3389/fgene.2018.00264 PMID: 30079080

26. Bassano I, Ong SH, Sanz-Hernandez M, Vinkler M, Kebede A, Hanotte O, et al. Comparative analysis of the chicken IFITM locus by targeted genome sequencing reveals evolution of the locus and positive selection in IFITM1 and IFITM3. BMC Genomics. 2019; 20 (1): 272. https://doi.org/10.1186/s12864-019-5621-5 PMID: 30952207

27. Qanbari S, Rubin CJ, Maqbool K, Weigend S, Weigend A, Geibel J, et al. Genetics of adaptation in modern chicken. PLoS Genet. 2019; 15 (4): e1007989. https://doi.org/10.1371/journal.pgen.1007989 PMID: 31034467

28. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456 (7218): 53–9. https://doi.org/10.1038/nature07517 PMID: 18987734

29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat methods. 2012; 9 (4): 357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

30. Danecek P, Auton A, Abecasis G, Albers CA, Banks Eric, DePristo MA, et al. 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011; 27(15): 2156–8. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

31. Saint-Pé K, Leitwein M, Tissot L, Poulet N, Guinand B, Berrebi P, et al. Development of a large SNPs resource and a low-density SNP array for brown trout (Salmo trutta) population genetics. BMC Genomics. 2019; 20 (1): 582. https://doi.org/10.1186/s12864-019-5958-9 PMID: 31307373

32. Phillips C, Amigo J, Tillmar AO, Peck MA, de la Puente M, Ruiz-Ramírez J, et al. A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel. Forensic Sci Int Genet. 2020; 46: 102232. https://doi.org/10.1016/j.fsigen.2020.102232 PMID: 31986343

33. Cantacessi C, Jex AR, Hall RS, Young ND, Campbell BE, Joachim A, et al. A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. Nucleic Acids Res. 2010; 38 (17): e171. https://doi.org/10.1093/nar/gkq667 PMID: 20682560

34. Fang H, Wu Y, Narzisi G, O'Rawe JA, Barrón LTJ, Rosenbaum J, et al. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 2014; 6 (10): 89. https://doi.org/10.1186/s13073-014-0089-z PMID: 25426171

35. Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 2014; 8 (1): 14. https://doi.org/10.1186/1479-7364-8-14 PMID: 25078893

36. Ghoneim DH, Myers JR, Tuttle E, Paciorkowski AR. Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. BMC Res Notes. 2014; 7: 864. https://doi.org/10.1186/1756-0500-7-864 PMID: 25435282

37. De Summa S, Malerba G, Pinto R, Mori A, Mijatovic V, Tommasi S. GATK hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC Bioinformatics. 2017; 18 Suppl 5:119. https://doi.org/10.1186/s12859-017-1537-8 PMID: 28361668

38. Walker MA, Pedamallu CS, Ojesina AI, Bullman S, Sharpe T, Whelan CW, et al. GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics. 2018; 34 (24): 4287–9. https://doi.org/10.1093/bioinformatics/bty501 PMID: 29982281

39. Brouard JS, Schenkel F, Marete A, Bissonnette N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. J Anim Sci Biotechnol. 2019; 10: 44. https://doi.org/10.1186/s40104-019-0359-0 PMID: 31249686

40. Schnepp PM, Chen MJ, Keller ET, Zhou X. SNV identification from single-cell RNA sequencing data. Hum Mol Genet. 2019; 28 (21): 3569–83. https://doi.org/10.1093/hmg/ddz207 PMID: 31504520

41. Zhao Y, Wang K, Wang WL, Yin TT, Dong WQ, Xu CJ. A high-throughput SNP discovery strategy for RNA-seq data. BMC Genomics. 2019; 20 (1): 160. https://doi.org/10.1186/s12864-019-5533-4 PMID: 30813897

42. Liu Q, Guo Y, Li J, Long JR, Zhang B, Shyr Yu. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics. 2012; 13 Suppl 8: S8. https://doi.org/10.1186/1471-2164-13-S8-S8 PMID: 23281772