Effect of Database Size in the Genetic Variants Calling

Sunhee Kim[1], Young-suk Lee[2] and Chang-Yong Lee[1]

[1]*The Department of Industrial and Systems Engineering, Kongju National University, Cheonan 330-717, South Korea*

[2]*Center for RNA Research, Institute for Basic Science, Seoul 151-742, South Korea*

Keywords: Base Quality Score Recalibration, Single Nucleotide Polymorphism, Variant Calling, Database.

Abstract: The base quality score recalibration (BQSR) is an important step in the variant calling from high-throughput sequence data. Motivated by the fact that BQSR necessarily requires a database of known variants such as the dbSNP, we present an extensive analysis on BQSR results for human and rice genome. We showed that the recalibration results depended on the size of the database: the more variants are there in the database, the larger averaged value of the recalibrated base quality scores is obtained. This implies that the recalibrated quality score is lower than it should be when the number of variants in the database is not large enough. Based on the finding that the size of the database should play a crucial role in BQSR, we proposed a method to create a database when the size of a database is not large enough for BQSR results to be reliable. We demonstrated that, in the case of human, the database constructed by the proposed method generated almost the same results as the human dbSNP. In the case of rice, however, we showed that the proposed database is more reasonable than the rice dbSNP.

1 INTRODUCTION

The high-throughput sequencing, such as the next generation sequencer (NGS), generates unprecedentedly large amount of genetic data at a low cost (Metzker, 2010). Massive sequence data are mainly used to identify various genetic variations including the single nucleotide polymorphism (SNP) with the help of relevant bioinformatic tools. The most well known software for the SNP calling is the Genome Analysis Toolkit (GATK), an open pipeline provided by the Broad Institute (DePristo and et al., 2011; der Auwera and et al., 2013).

While a NGS platform produces much more sequence data than the traditional Sanger sequencing does, generated reads are shorter than the Sanger's in length and of inferior quality containing more sequencing errors. When a NGS platform calls a base, they also provide an estimated quality of the base in terms of Phred score (Ewing and et al., 1998; Ewing and Green, 1998),

$$Q_{phred} = -10\log_{10} p(\varepsilon) \ , \qquad (1)$$

where $p(\varepsilon)$ is the error rate, the probability of observing an incorrectly called base. Q_{phred} is an integer and known as the base quality score.

Studies have demonstrated that the Phred-scaled base quality scores issued by a NGS platform are often inaccurate and deviate from true error rates (Li and et al., 2009b; Brockman and et al., 2008). Base quality scores are prone to various sources of systematic (i.e., non-random) and technical errors. Examples include the general trend of increased sequencing errors in later sequence cycles, dinucleotide contents error, and errors due to manufacturing flaws in the equipment. In addition, each platform has a specialized base calling scheme in its sequencing workflow that leads to sequencing errors. The accuracy of a base quality score is important because the downstream analyses of the variant calling rely heavily on the per-base quality score (DePristo and et al., 2011). Thus, under- or over-estimated base quality scores may result in inaccurate (i.e., false positive or false negative) variant calls.

Because the correct estimate of the quality score is essential in the variant calling, GATK provides a recalibration tool for the base quality score, named the base quality score recalibration (BQSR) (GATK, 2018). As a data pre-processing step of sequencing-by-synthesis reads in an aligned sequence of SAM/BAM file (Li and et al., 2009a), BQSR detects systematic errors in the estimated quality score of each base and recalibrates the base quality score, not the base call itself. BQSR is versatile in that it can be applied to the sequencing data

Kim, S., Lee, Y. and Lee, C.
Effect of Database Size in the Genetic Variants Calling.
DOI: 10.5220/0007413402090215
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 209-215
ISBN: 978-989-758-353-7

of various platforms. Besides BQSR, the recalibrating quality of nucleotides (ReQON) (Cabanski and et al., 2012) providesd by a package written in R (R-project, 2018) claims that it also performs the recalibration, which also utilizes known SNP database. Another study related to the base quality recalibration is RIG (Recalibration and Interrelation of Genomic Sequence Data) that is a workflow to generate collection of variants from various available genomic resources (R. McCormick and Mullet, 2015). In addition, it was claimed that the recalibration was solved without an external SNP database (Chung and Chen, 2017).

BQSR is a method of adjusting platform-provided base quality scores to be more accurate by using an external resource (or database) of known variants, such as the dbSNP (Sherry and et al., 2001). It is a method of adjusting Phred quality score to be more accurate by examining every base in the BAM/SAM file. The underlying assumption of BQSR is that any mismatched base with the reference genome at a position not listed in the database is an error. That is, BQSR assumes that a mismatched base listed in the database is correctly sequenced real variants, whereas a mismatch not listed in the database is regarded as a sequencing error. In addition to the database, BQSR groups bases into different categories with respect to various covariates, such as the machine cycle, the base position in a read, and the dinucleotide context, then takes the mismatch rate for each category into account in the recalibration.

The basic assumption of BQSR, any potential variant in a read that is not listed in the database is an error, is reasonable and valid in the statistical sense only if we have enough and accurate information about known variants in the database. If the size of the database is smaller than it should be, the number of false negative (i.e., variants that are incorrectly identified as errors) would increase. Thus, the usefulness of BQSR heavily relies on the number and quality of reported variants in the database. This means that when the database is incomplete, mismatched bases are less likely to be identified by the database; as a result, the quality of the bases will be inferred to be lower than it actually is (Wang and et al., 2015). Thus, it may be ineffective to run BQSR when the database of known variants is not comprehensive. In this respect, it is imperative to find a way of recalibrating the base quality score for species of having not enough information about known variants.

This naturally leads to the following questions. Is the number of known variants in the database of an organism enough to trust BQSR results? What can we do when we do not have enough known variants in the database? In this study, we try to answer these questions. To this end, we used NGS data and the dbSNPs of human and rice, and performed various empirical studies for BQSR to investigate the above questions. Although BQSR step in GATK provides a room to add new covariates to the recalibration, for simplicity, we do not add any new covariate besides default ones provided by GATK.

2 MATERIALS AND METHODS

2.1 Data Acquisition

We used the genome-wide NGS data of FASTQ (Cock and et al., 2010) files for human and rice, each of which is obtained from the 1000 Genomes Project (Human-Genomes, 2015) and the 3000 rice genomes project (Rice-Genomes, 2014), respectively. The 1000 Genomes Project is an international research effort to establish a detailed catalogue of human genetic variation by resequencing about one thousand participants from a number of different ethnic groups. The 3000 rice genomes project is also an international effort of resequencing a core collection of 3,000 rice accessions from 89 countries. The data can be freely downloaded at http://www.internationalgenome.org/ for the 1000 Genomes Project, and http://dx.doi.org/10.5524/200001 for the 3000 Rice Genomes Project. For the empirical studies, we randomly selected 10 samples from each project. The estimated average coverage over 10 resequenced samples is about 5x and 9x for human and rice, respectively.

Variants are called against the Nipponbare IRGSP-1.0 reference genome for rice and the GRCh38 for human. The GRCh38, the Genome Reference Consortium Human genome build 38, is built from reference sequences of different individuals, not just from one individual's genome sequence. Whereas, the IRGSP-1.0, built from one accession's genome, is assembled by the Oryza sativa Japonica Group (Japanese rice) genome from National Institute of Agrobiological Sciences. The reference sequences for both can be obtained from Refs. (Human-Reference, 2018) and (Rice-Reference, 2018), respectively.

2.2 dbSNP

As for the database of known variants, we use the dbSNP (Database, 2018) as GATK does. The dbSNP is a repository of all known variants, such as SNPs and

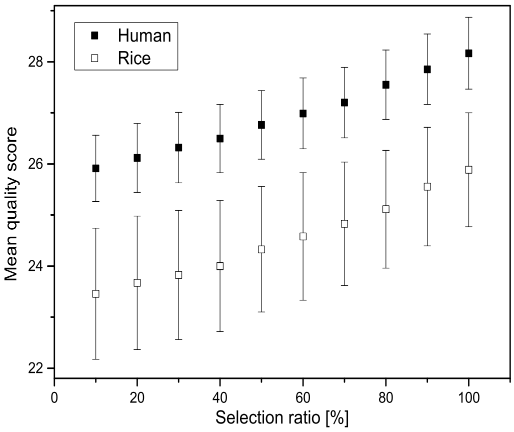

Figure 1: Plots of the average $\bar{\bar{Q}}$ over 10 individuals (or accessions) versus the percentage of selected variants from the dbSNP: the human (□) and the rice (◯). The error bars are the corresponding standard deviations s_Q.

indels, open to the public. The dbSNP is a collection of VCF files generated from various resources. Initially constructed for human, the dbSNP has extended to other organisms. We used the dbSNP build 151 for human (Homo sapiens) that consists of about 318,739,000 variants and the dbSNP build 151 for rice (Oryza sativa) that consists of about 12,185,000 variants. These were the latest version when most of data analyes were carried out.

A base in a BAM/SAM file whose poistion is listed as a variant in the database is not considered as a variant when the base matches with the reference sequence. In this case, the variant in the database is void in the sense that it does not play any role as far as BQSR is concerned. On the other hand, the position of a mismatched base can be in the list in the database as a variant. In this case, we define the variant as an "effective" variant. Thus, variants in the database can be either effective or non-effective, and it is effective variants that affect BQSR. Thus, from the perspective of BQSR, variants in a database (e.g. dbSNP) can be classified into two categories: either effective or not.

2.3 Characteristics of Mean Recalibared Base Quality Score

To answer a question about the sufficiency of variants in the database, we investigated how the size of a database affected BQSR results. This can be accomplished by examining the dependence of BQSR results on the number of variants used as the database. Specifically, we constructed 10 test databases of different sizes, each of which is composed of variants randomly selected out of the dbSNP from 10% to 100% at a 10% interval. We then peformed BSQR by using test databases of different number of variants. In this way, each database and BQSR result can be identified by its selection ratio.

For a given selection ratio, let Q_{ij} be the recalibrated quality score of base j of sample (individual or accession) i. Then, the mean recalibrated quality score $\bar{Q}_i$ of sample i over all bases is given as

$$\bar{Q}_i \equiv \frac{1}{m_i} \sum_{j=1}^{m_i} Q_{ij} \ , \qquad (2)$$

where m_i is the number of bases in sample i. With $\bar{Q}_i$'s, we can define the sample mean $\bar{\bar{Q}}$ and the sample variance s_Q^2 of $\bar{Q}_i$'s as

$$\bar{\bar{Q}} \equiv \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_i \ \text{ and } \ s_Q^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} \left\{ \bar{Q}_i - \bar{\bar{Q}} \right\}^2 , \qquad (3)$$

where n is the number of samples, and in our case, $n = 10$.

The results of BQSR are shown in Fig. 1, in which we plot $\bar{\bar{Q}}$ and s_Q (errorbars) in Eq. (3) versus the selection ratio. For both human and rice, the mean recalibrated quality score increases as the ratio increases. This is expected because, according to the underlying assumption of BQSR as stated in Section 1, the more variants we have in the database, the less number of bases in the BAM/SAM file are regarded as errors which results in higher recalibrated quality scores.

In addition, we find that s_Q, the sample standard deviation of $\bar{Q}_i$'s, of rice is about two times larger than that of human. This is mainly due to the characteristic of the reference sequences and the selected samples. As stated in Section 2.1, because the human reference sequence built from sequences of different individuals, the reference sequences contains sequences from different origins of country. As a result, the degree of mismatches with the reference sequence spread more or less uniformly across different origin of countries. In contrast, the rice reference sequence is built from a single accession of Nippobare Japonica. Thus, rice samples of different variety groups, such as Indica anothe major rice cultivar, from Japonica may include much more mismatches (and thus errors) than Japonica accessions. This in turn leads to a large deviation of the average quality scores over different accessions.

More importantly, from Fig. 1 we find that the mean recalibrated quality score of human gemome is larger than that of rice genome for all selection ratios. This result is in contradiction to the fact that, before the recalibration, the average quality scores for both human and rice are about the same: 33.02 ± 3.74 for human and 36.07 ± 0.63 for rice. Because there is no reason to assume that the base quality scores of human genome are higher than those of rice genome, it is not reasonable to have a noticable gap between the mean recalibrated base quality scores of human and rice.

One plausible explanation for the gap is that the number of (effective) variants in the rice dbSNP is either not large enough for BQSR to be valid, or far smaller than that in the human dbSNP. Because the larger is the size of the database, the higher the recalibrated base quality scores are, BQSR scheme with a database having not enough number of variants may under-estimate the quality score in the recalibration step. While the genome size of rice is about one eighth of that of human, the number of variants in the rice dbSNP is about one twenty fifth of those in the human dbSNP. This indirectly indicates that the size of the rice dbSNP is significantly small relatively to that of the human dbSNP.

The above finding raises an important question: what do we do if we do not have enough number of variants as the database for the recalibration? In what follows, we suggest a recalibration method when the number of variants in a database is not enough.

2.4 Proposed Method of Creating Database

We propose a method of constructing a database as an alternative to an existing database for BQSR when there is no database or the existing database does not contain enough number of variants. The basic idea of the proposed method is to create a new database, similar to the suggestion by GATK forum (Bootstrap, 2018), as follows. We first perform the variant calling by using a variant calling pipeline, such as GATK, without BQSR step to obtain a variant call format (VCF) file that contains variants called by the pipeline without BQSR step. We then perform again the variant calling including BQSR step by using the VCF file as the database. In short, the VCF file obtained without BQSR step plays a role of an alternative database to the existing database, such as the dbSNP. As the variant calling pipeline creates the database for itself, we will call the VCF file as the dbSELF.

The above method can be generalized by an iterative aggregation, in which the dbSELF is updated by performing the variant calling pipeline repeatly. There are two ways to update the dbSELF. The first method is to accumulate the variants called by the pipeline to the current dbSELF, and the second is to replace the current dbSELF with the most recent VCF file. Thus, the dbSELF generated by the accumulation method increases in size, whereas, in the replacement method, the VCF file obtained from the *n*th variant calling is the dbSELF for the $(n+1)$th variant calling. In general, the number of variants in the dbSELF of the accumulation method is larger than that of the replacement method. Note that the number of variants in the dbSELF depends on sample (individual or accession). The procedure of the proposed method is schematically depicted in Fig. 2.

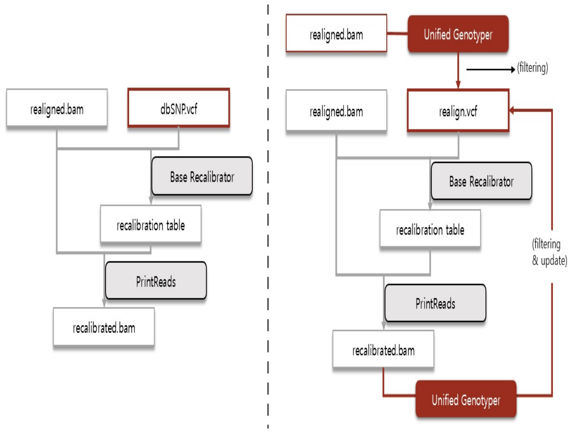

Figure 2: The schematic flow chart of current BQSR setp (left) and the proposed step (right).

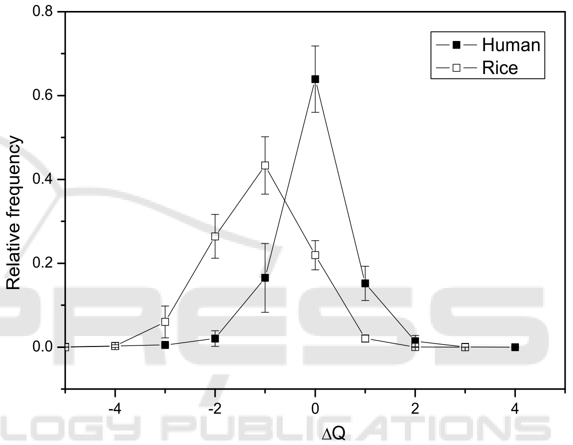

Figure 3: The relative frequency of Δ_{ij}'s for human and rice. The plots are averaged distribution over 10 samples with error bars representing the standard deviation.

3 EXPERIMENTAL RESULTS AND DISCUSSION

In order to assess the proposed method, we adopted GATK and performed the variant calling by using two different databases: the dbSNP and the dbSELF. We analyzed basewise difference in the recalibrated base quality scores obtained by the two databases. The dif-

Table 1: Statistics of dbSELF and dbSNP. N_{total} and N_{eff} are the number of variants and the effective variants in the dbSNP, respectively. N_{ebase} and N_{error} are the number of effective bases and errors, respectively; the error rate is defined as the ratio of the number of errors to the total number of bases in raw data. All quantities, except the number of variants in the dbSNP, are averaged over 10 samples, and the corresponding standard devations are in the parentheses.

	Database		Genotype		
Human	N_{total}	N_{eff}	N_{ebase}	N_{error}	Error rate (%)
dbSNP	318,739,162	5,965,063 (857,126)	17,919,487 (4,950,378)	14,570,351 (3,424,599)	0.11 (0.04)
dbSELF	2,824,914 (571,181)	2,822,702 (570,685)	14,316,273 (4,649,247)	18,173,565 (4,020,284)	0.14 (0.05)
Rice	N_{total}	N_{eff}	N_{ebase}	N_{error}	Error rate
dbSNP	12,185,568	1,411,561 (955,475)	8,936,172 (6,314,281)	7,410,641 (3,009,390)	0.24 (0.12)
dbSELF	1,505,469 (1,001,366)	1,464,242 (981,338)	10,815,057 (6,969,199)	5,531,758 (2,333,059)	0.18 (0.09)

ference is expressed as $\Delta Q_{ij} \equiv Q_{ij}^{SNP} - Q_{ij}^{SELF}$, where Q_{ij}^{SNP} and Q_{ij}^{SELF} are the recalibrated quality scores of base j in sample i obtained by using the dbSNP and the dbSELF, respectively. Note that ΔQ_{ij} is an integral value as the base quality score is an integer.

Figure 3 shows relative frequency distributions of ΔQ_{ij}'s for both human and rice averaged over 10 samples together with the standard deviations represented by the error bars. From Fig. 3, we see that the disribution of ΔQ_{ij}'s for human is symmetric about $\Delta Q_{ij} = 0$, and the majority of bases (about 64%) have $\Delta Q_{ij} = 0$ and more than 95% of bases have $|\Delta Q_{ij}| \leq 1$. This means that more than 95% of the recalibrated base quality scores obtained by using two different databases are the same or differ by one Phred score. This result suggests that the dbSELF can serve a reasonably good alternative to the dbSNP.

In the case of rice, however, whereas about 22% of bases have $\Delta Q_{ij} = 0$, more than 70% of bases have their recalibrated quality scores obtained by the dbSELF higher than those obtained by the dbSNP. Considering that BQSR with the rice dbSNP underestimates the recalibrated quality score compared to the human dbSNP as discussed in Section 2.3, the rice dbSELF can alleviate, at least in part if not entirely, the under-estimate of the recalibrated scores. In this sense, the rice dbSELF may substitute for the rice dbSNP for a better BQSR result.

As stated in Section 1, the genotype of a base is regarded as an error when the base in a BAM/SAM does not match with the reference at a position not listed in the database. As a complementary to the error, we define an effective base as a mismatched base that is identified by an effective variant listed in the database. Thus, a mismatched base is either an error or an effective base. In Table 1, we list the number of variants and effective variants in the two databases, together with the statistics of effective bases and errors. In addition, we estimate the error rate, which is the ratio of the number of errors to the total number of genotyped bases in the raw data (i.e., FASTQ file). Because all quantities, except the number of variants in the dbSNP, depend on samples, we report in Table 1 the mean and the standard deviation of the quantities over 10 samples. Note that the standard deviations of the numbers of variants, effective variants, and effective bases for rice are larger than those for human regardless of the database. This is due to the characteristics of the reference sequence discussed in 2.3.

We see from Table 1 that, for both human and rice, while the dbSELF contains less number of variants than the dbSNP, almost all variants in the dbSELF are effective variants. This is expected because the dbSELF is nothing but a set of variants called from samples without BQSR step. In the case of human, we find that the dbSELF contains far less number of variants (about 0.8%) than the dbSNP does. However, more than 99% of variants in the dbSELF are effective variants, whereas only about 2% in the dbSNP are effective. More importantly, although the dbSELF has less than a half as many effective variants as the dbSNP has, the error rates obtained by using the two databases differ by only 0.03%. This difference is not a significant compared to the difference in the number of effective variants.

In the case of rice, we can see from Table 1 that the dbSELF contains more effective variants (about 4%) than the dbSNP, although the dbSELF contains less number of variants (about 12%) than the dbSNP. While about 12% of variants in the dbSNP are effective, more than 97% of variants in the dbSELF are effective. The fact that the rice dbSELF identifies more effective variants is a primary reason that BSQR using the dbSELF gives higer recalibrated scores on average than the dbSNP does. In addition, the dbSELF generates more effective bases than the dbSNP does; as a result, the error rate using the dbSELF is smaller than that using the dbSNP. This basically yields higher Q_{ij}^{SELF} on average than Q_{ij}^{SNP}.

Note that there is no reason in prior that the error rate of rice should be greater than that of human. Rather, we should expect about the same error rate for both human and rice. In this sense, the fact that the error rate using the dbSELF is almost comparable

Table 2: The results of the number of effective variants and the recalibrated base quality score from the iteration method. N_{acc} and N_{rep} stand for the number of effective variants from the accumulation and the replacement methods, respectively. V_{acc} and V_{rep} represent the number of called variants in the VCF file from the accumulation and the replacement methods, respectively. Q_{acc} and Q_{rep} are the recalibrated quality scores for the accumulated and replaced database, respectively. All quantities are the averaged quantities over 10 samples and the standard deviations are in parantheses.

Human					
Iteration	1	2	3	4	5
N_{acc}	2,822,702 (570,685)	2,846,302 (548,120)	2,849,574 (546,712)	2,850,760 (546,425)	2,851,651 (545,974)
N_{rep}	2,822,702 (570,685)	2,572,609 (496,302)	2,588,531 (499,300)	2,591,406 (498,604)	2,591,033 (499,800)
V_{acc}	2,574,517 (496,981)	2,609,946 (500,232)	2,611,333 (502,151)	2,611,592 (499,787)	2,611,857 (500,988)
V_{rep}	2,574,517 (496,981)	2,590,469 (499,992)	2,593,340 (499,282)	2,592,973 (500,495)	2,595,153 (499,734)
Q_{acc}	28.22 (1.38)	28.85 (1.71)	28.84 (1.72)	28.86 (1.69)	28.86 (1.72)
Q_{rep}	28.22 (1.38)	28.61 (1.62)	28.62 (1.68)	28.64 (1.67)	28.66 (1.66)

Rice					
Iteration	1	2	3	4	5
N_{acc}	1,464,242 (981,338)	1,465,483 (981,566)	1,465,981 (981,858)	1,466,151 (981,975)	1,466,228 (982,017)
N_{rep}	1,464,242 (981,338)	1,404,392 (913,189)	1,430,244 (946,967)	1,431,509 (948,436)	1,431,252 (948,166)
V_{acc}	1,441,154 (928,358)	1,471,463 (967,600)	1,471,935 (968,295)	1,472,026 (968,474)	1,471,632 (967,802)
V_{rep}	1,441,154 (928,358)	1,468,866 (964,458)	1,470,220 (966,029)	1,469,945 (965,743)	1,470,180 (965,956)
Q_{acc}	27.02 (2.30)	28.68 (2.10)	28.63 (2.10)	28.65 (2.08)	28.63 (2.11)
Q_{rep}	27.02 (2.30)	28.47 (2.18)	28.54 (2.15)	28.51 (2.15)	28.53 (2.16)

with that for human supports that the rice dbSELF is a reasonably good alternative to the rice dbSNP.

We assessed two different iteration methods, accumulation and replacement, by updating the database.The results are shown in Table 2. From Table 2 we found the following properties that both human and rice have in common. First, the number of effective variants obtained from the accumulation method increases steadily, without any sudden change, which is expected. With the replacement method, in contrast, the number of effective variants decreases at the second iteration and increases back at the third iteration. In particular, the first iteration produces the largest number of effective variants. Of course, in all iterations, the accumulation method yields more variants than the replacement method. Second, the average scores obtained by the accumulation method are slightly higher than those obtained by the replacement method for all iterations. This is expected because, by defintion, the accumulation method yields more effective variants in the database, although not much, than the replacement method does.

4 CONCLUSIONS

In this study, we investigated the validity of the dbSNP for BQSR. We found that the recalibration results were closely related to the size of dbSNP. This implied that BQSR results might not be reliable when the size of the database is not large enough. Based on the finding that the size of the database should play a crucial role in BQSR, we proposed a method to create a database when the size of a database is not large enough. We demonstrated that, in the case of rice, the proposed method of construction a database is more reasonable than the rice dbSNP. This suggests that the propsed method can be applied to the variant callings of other species for which the size of the database is not large enough.

ACKNOWLEDGEMENTS

This work was supported by a Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (NRF-2018R1D1A3B07042338).

REFERENCES

Bootstrap (2018). https://www.broadinstitute.org/gatk/guide/article?id=44. (Accessed: 2010-09-30).

Brockman, W. and et al. (2008). Quality scores and snp detection in sequencing-by-synthesis systems. *Genome Res.*, 18(5):763–770.

Cabanski, C. and et al. (2012). Reqon: a bioconductor package for recalibrating quality scores from next-generation sequencing data. *BMC Bioinformatics*, 13:221.

Chung, J. and Chen, S. (2017). Lacer: accurate: base quality score recalibration for improving variant calling from next-generation sequencing data in any organism. *https://www.biorxiv.org/content/early/2017/04/27/130732*.

Cock, P. and et al. (2010). The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res.*, 38(6):1767–1771.

Database (2018). http://www.ncbi.nlm.nih.gov/dbvar/content/org_summary/. (Accessed: 2010-03-30).

DePristo, M. and et al. (2011). A framework for variation discovery and genotyping using next generation dna sequencing data. *Nat. Genet.*, 43(5):491–498.

der Auwera, G. V. and et al. (2013). From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, 11:11.10.1.–11.10.33.

Ewing, B. and et al. (1998). Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Res.*, 8(3):175–185.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res.*, 8(3):186–194.

GATK (2018). https://software.broadinstitute.org/gatk/documentation/tooldocs/3.8-0/org_broadinstitute_gatk_tools_walkers_bqsr_BaseRecalibrator.php. (Accessed: 2010-09-30).

Human-Genomes (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Human-Reference (2018). ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/. (Accessed: 2010-01-30).

Li, H. and et al. (2009a). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Li, R. and et al. (2009b). Snp detection for massively parallel whole-genome resequencing. *Genome Res.*, 19(6):1124–1132.

Metzker, M. (2010). Sequencing technologies-the next generation. *Nat. Rev. Genet.*, 11(1):31–46.

R. McCormick, S. T. and Mullet, J. (2015). Rig: Recalibration and interrelation of genomic sequence data with the gatk. *G3*, 5(4):655–665.

R-project (2018). https://www.r-project.org/. (Accessed: 2010-10-05).

Rice-Genomes (2014). The 3,000 rice genomes project. *GigaScience*, 3:7–12.

Rice-Reference (2018). http://rapdb.dna.affrc.go.jp/download/archive/irgsp1/IRGSP-1.0_genome.fasta.gz. (Accessed: 2010-02-30).

Sherry, S. and et al. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res.*, 29(1):308–311.

Wang, J. and et al. (2015). Variant calling using ngs data in european aspen (populus tremula). In Sablok, G. and et al., editors, *Advances in the understanding of biological sciences using next generation sequencing (NGS) approaches*, pages 43–61. Springer, New York.