RDscan: A New Method for Improving Germline and Somatic Variant Calling Based on Read Depth Distribution

SUNHO LEE,[1,2,†] SEOKCHOL HONG,[1,†] JONATHAN WOO,[1] JAE-HAK LEE,[1] KYUNGHEE KIM,[1] LUCIA KIM,[3] KUNSOO PARK,[2,*] and JONGSUN JUNG[1,*]

[1]Syntekabio Inc, 187 Techno 2-ro, Daejeon 34025, Republic of Korea

[2]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, Republic of Korea

[3]Department of Pathology, Inha University Hospital, Incheon 22332, Republic of Korea

[†]Co-first authors.

*Co-corresponding authors.

ssunho.lee@gmail.com

schong@syntekabio.com

jwhwoo@syntekabio.com

jhlee@syntekabio.com

khkim@syntekabio.com

luciado@inha.ac.kr

kpark@theory.snu.ac.kr

jung@syntekabio.com

ABSTRACT

Several tools have been developed for calling variants from next-generation sequencing data. Although they are generally accurate and reliable, most of them have room for improvement, especially regarding calling variants in datasets with low read depth. In addition, the somatic variants predicted by several somatic variant callers tend to have very low concordance rates. In this study, we developed a new method (RDscan) for improving germline and somatic variant calling in next-generation sequencing data. RDscan removes misaligned reads, repositions reads, and calculates *RDscore* based on the read depth distribution. With *RDscore*, RDscan improves the precision of variant callers by removing false-positive variant calls. When we tested our new tool using the latest variant calling algorithms and data from the 1000 Genomes Project and Illumina's public datasets, accuracy was improved for most of the algorithms. After screening variants with RDscan, calling accuracies increased for germline variants in 11 out of 12 cases and for somatic variants in 21 out of 24 cases. RDscan is simple to use and can effectively remove false-positive variants while maintaining a low computation load. Therefore, RDscan, along with existing variant callers, should contribute to improvements in genome analysis. Source code and binaries, implemented in C++ and supported on Linux, are available for download at https://github.com/satchellhong/RDscan.

Keywords: variant calling, variant filtering, read depth distribution, next-generation sequencing, germline variant, somatic variant

1. INTRODUCTION

Over the past decade, large-scale sequencing of human genomes has been carried out using next-generation sequencing (NGS) technologies (The 1000 Genomes Project Consortium, 2015; Lek et al., 2016; Yi et al., 2010). Variants in human genomes are intimately associated with the genetic causes of many human diseases, as well as the genetic diversity within and among human populations (Lee et al., 2014; Ng et al., 2010). Therefore, identifying variants from NGS data has become a key foundation of human genome analysis.

The accuracy of variant calls for single-nucleotide variants (SNVs) and insertions and deletions (indels) depends on artifacts from sequencing device error, DNA contamination, and read misalignment (Li, 2014; Do and Dobrovic, 2015). These artifacts can lead variant callers to call

artificial unreal variants. To minimize false-positive variant calling frequencies, several algorithms have been developed based on key features such as read depth, base/mapping quality, strand bias, and haplotype (DePristo et al., 2011; Kim et al., 2018; Garrison and Marth, 2012; Rimmer et al., 2014; Lai et al., 2016; Koboldt et al., 2013). Recently, deep learning models such as DeepVariant (Poplin et al., 2018) and NeuSomatic (Sahraeian et al., 2019) have been implemented for variant calling based on these features. However, variant callers for the variants with low read depth and low variant allele frequency (VAF) still have room for improvement in terms of accuracy (Krøigård et al., 2016; Kim et al., 2018; Sahraeian et al., 2019).

To improve variant call accuracy, we propose read depth distribution as a new feature. We expect that the depth distribution of the aligned read set containing a variant should not differ from the depth distribution of the whole read set mapped to the same region, as genomic DNAs are randomly broken into smaller fragments and reads are generated from these fragments by NGS devices regardless of whether a variant is included (King et al., 2006). Hence, this similarity of the read depth distributions can be used to eliminate artificial variants; the variant candidate with read depth distribution similar to the overall read depth distribution should be kept as the true-positive variant, whereas other candidates should be removed as false. We used a similar approach for haplotyping the major histocompatibility complex regions (Ka et al., 2017).

In this study, we developed RDscan, a novel variant filtering method based on read depth distribution that effectively removes artificial variants. This method can improve the accuracy of variant calls relative to any other method. For most variant calling algorithms that we tested, the accuracies of these algorithms were further improved by differentiating real variants from false-positive variants using RDscan.

2. MATERIALS AND METHODS

2.1. WGS data from public genome datasets

To evaluate the accuracy of RDscan for germline variants, we used WGS data for two samples, HG001 (34x) and HG002 (25x), from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015); the detailed dataset information is provided in Supplementary Table S1. The reference standards, the Genome in a bottle (GIAB) truth sets for HG001 (v3.3.2) and HG002 (v4.1), were downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/. The variant calling accuracy was evaluated within the high-confidence region suggested by GIAB.

2.2. In silico mixtures of unrelated germline samples

We used mixed DNA sequencing data from two unrelated individuals (NA12878 and NA12877) generated by Illumina (Kim et al., 2018) to evaluate the accuracy of somatic calls by RDscan. The dataset, consisting of three paired samples with different proportions of normal and tumor DNA purity (100% and 80%, 100% and 20%, and 90% and 80%), was downloaded from https://www.ncbi.nlm.nih.gov/sra/ (see Supplementary Table S1). The truth sets and the confidence region data for verification were downloaded from Illumina.

[Figure 1 about here.]

2.3. Collecting candidate variants

RDscan provides an additional filtering method for eliminating the false-positive variants called by the other variant callers. To collect the germline variants, we used GATK HaplotypeCaller (DePristo et al., 2011), Strelka2 (Kim et al., 2018), and DeepVariant (Poplin et al., 2018). For the somatic variants, we used Strelka2, VarDict (Lai et al., 2016), Mutect2 (Cibulskis et al., 2013), and NeuSomatic (Sahraeian et al., 2019). The collected variants are re-evaluated by the RDscan method, as described below.

2.4. Filtering misaligned reads in repeat region

RDscan prevents erroneous influence on variant calling by preliminarily removing misaligned reads in the repeat regions where one or more sequences appear more than three times repeatedly. Specifically, a read that starts or ends in a repeat region is considered misaligned. As an example, shown in Step 2 of figure 1, a repeat region can consist of nucleotide sequences that repeat TG four times. In this case, we can simultaneously observe two variations at locus *i*, one for having G instead of the reference sequence A, and the other for the AT deletion. However, the top four reads with their ends in the repeat region cannot differentiate the case of variation G from the case of AT deletion because they do not have sequence information to anchor after the repeat region. For reads extending beyond the repeat region, the sequence alignment before and after the repeat region will force the variant to be either G or deletion of AT; otherwise, the misidentification would lead to a positional shift of the sequences after the repeat region. Hence, if the start or end of a read is aligned in a repeat region, we consider the read as having a risk of misalignment, and therefore eliminate it.

2.5. A new variant calling method

In this section, we introduce a new method for determining the reliability of a variant call by comparing the read depth distribution with the variant to the whole read depth distribution aligned over the variant locus. For each candidate variant, RDscan first extracts all reads aligned to the locus of the variant in the corresponding BAM file and then removes misaligned reads in the repeat regions, as described in the previous section. This set of filtered reads is called *ALL*. Among the reads in *ALL*, RDscan extracts *VAR*, a new set of reads containing the variant (Step 3 in Fig. 1). If the variant is an indel, it causes a difference in the aligned area lengths of reads between *ALL* and *VAR*, which adversely affects the calculation of the correlation of the read depth distribution between the two groups. To prevent this, RDscan adjusts alignment of the reads. As shown in Step 4 of figure 1, the reads with the AT deletion have an aligned region longer than other reads by the length of the deleted sequence. In this example, RDscan decreases the aligned area of the reads with the AT deletion by the length of the deletion. Conversely, in the case of an insertion variant, the aligned region is increased by the length of the variant. To compare the read depth distributions between *ALL* and *VAR*, RDscan calculates the read depths for each group. We denote the read depth of the group *X* at locus *i* as D_X^i and the read depth vector of the group *X* as $D_X = [D_X^{sl}, D_X^{sl+1}, \cdots, D_X^{sl+l-1}]$, where *sl* is the start locus of the *Comparison Region* and *l* is the length of the region. The *Comparison Region* is the area in which the reads in *ALL* are aligned. RDscan then uses the following score function to determine the reliability of the variant.

$$RDscore = \begin{cases} 0, & \text{if } corr(D_{ALL}, D_{VAR}) < 0 \\ (corr(D_{ALL}, D_{VAR}))^2, & otherwise \end{cases} \quad (1)$$

The *RDscore* ranges from 0 to 1. The *RDscore* of a variant close to 1 indicates that the variant is reliable. We used ALGLIB library (ALGLIB; https://www.alglib.net/) to calculate the Pearson correlation coefficient (*corr*) between the depths of *ALL* and *VAR*. The exponent value at the equation (1) is to maximize the difference between the *corrs* of true-positive and false-positive variants. Specifically, the reference *corr* for the true-positive variants was set to 0.86 which preserves a sensitivity of 95% and 99.9% for NA12878 sample with coverage of 34x and 300x, respectively (see Fig. 2). The reference *corr* of false-positive variants was set from the average *corr* of the false-positive variants from each algorithm. For the NA12878 sample with coverage of 34x, the average *corrs* of the false-positive variants were 0.231, 0.431, and 0.245 for Strelka2, GATK Haplotype Caller, and DeepVariant respectively. The exponent values of the *RDscore* which maximize the distance between the reference *corrs* of true-positive and false-positive variants are 1.7, 2.5, and 1.8, respectively. As

a result, the average (2.0) of those values is used for the exponent value.

[Figure 2 about here.]

2.6. Variant calling method from paired tumor and normal samples

A variant that is present in a cancerous tissue, but not in matched normal tissue, is called a somatic variant. In practice, the normal sample may contain some tumor cells, or the tumor sample could be contaminated by normal cells; either case could cause real somatic variants to be discarded. To address this problem, RDscan considers the VAF of the tumor and normal samples. First, RDscan calculates the *RDscores* (RD_{tumor} and RD_{normal}) of a variant from the tumor and matched normal samples, respectively, using the equation (1) in the previous section. Then RDscan uses the following score function to determine whether the variant is somatic:

$$RDscore_{somatic} = \max\left(0, RD_{tumor} - \frac{VAF_{normal}}{VAF_{tumor}} \times RD_{normal}\right) \qquad (2)$$

The possible range of the $RDscore_{somatic}$ is 0–1 for a variant; a score closes to 1 means that the variant is somatic variant. For a given variant, VAF_{tumor} and VAF_{normal} represent, respectively, the variant allele frequencies of tumor and normal tissues.

3. RESULTS

3.1. Evaluations of germline variant calling accuracy

We ran RDscan with candidate variant sets from three germline variant callers (Strelka2, GATK HaplotypeCaller, DeepVariant) using two public datasets (HG001 and HG002). Germline variant callers were executed based on the best practices described by the authors (see Supplementary Table S2). Germline variant calling accuracy was evaluated using 'hap.py' (Krusche et al., 2019) relative to the GIAB truth set. In this study, we assumed that germline variants with *RDscore* > 0.5 were true-positive variants.

[Table 1 about here.]

We first compared the original set of variants called by each algorithm with the variants subjected to further filtering with RDscan. The original set of variants consisted of variants that passed all filtering criteria provided by each variant caller. Overall, RDscan improved variant calling accuracy by decreasing the number of false-positive calls (FPs) while minimizing the reduction in true-positive calls (TPs). As shown in Table 1, for SNVs in HG001, the number of FPs was reduced by 69.5% for

Strelka2 (from 438,358 to 133,573), 54.1% for GATK HaplotypeCaller (139,993 to 64,219), and 36.0% for DeepVariant (from 115,512 to 73,911), whereas the numbers of TPs were reduced by 0.65%, 0.74%, and 0.72% respectively. After screening with RDscan, the accuracy (F-score) for SNVs was higher in all six cases, and the accuracy for indels was higher in five out of six cases.

We also show the overall improvement in the performance of these algorithms achieved by RDscan. As shown in figure 3, performance for indels was noticeably improved, whereas for SNVs, RDscan slightly improved the variant calling performance relative to DeepVariant, even though it already had excellent variant calling accuracy. For both datasets, the overall performance of these three algorithms was improved when the variants were screened by RDscan (Supplementary Figure S1).

[Figure 3 about here.]

The additional burden of the quality checks should be relatively moderate. RDscan is simple to use and can effectively remove false-positive variants while still having a low computation load (about 2 hours using Intel Xeon E5620 2.4 GHz, 12 cores, and 64GB memory for one WGS dataset) relative to the runtime of the entire variant call pipeline. The memory usage requirement for RDscan is 5–10 GB.

3.2. Evaluations of somatic variant calling accuracy for paired tumor and normal samples

Next, we evaluated the performance of RDscan for somatic variant calling using in silico mixed datasets and the truth set provided by Illumina (Kim et al., 2018). We used these data to run some of the known somatic variant callers (Strelka2, VarDict, Mutect2, and NeuSomatic), and then ran RDscan to further screen the results obtained with each caller. The somatic variant callers were executed using the best practices provided by the authors (see Supplementary Table S2). We first tried running the latest version (v4.1.8.1) of Mutect2 using the best practice pipeline provided by bcbio-nextgen (Chapman et al., 2021). This pipeline includes the step of removing known germline variants by reference to an external database. Consequently, higher somatic variant calling accuracy is expected for a typical Tumor-Normal dataset. However, for the in silico datasets generated based on the germline variants used for verification in this paper, most of the variants were filtered out by this method, making analysis impossible. Therefore, we used Mutect2 (v4.1.1.0) as an alternative method to proceed only with basic somatic variant calling from the in silico datasets. We used 'som.py' (Krusche et al., 2019) to evaluate somatic variant calling accuracy against the in silico germline mixture truth sets. In this study, variants with $RDscore_{somatic} > 0.3$ were considered true-positive somatic variants.

[Table 2 about here.]

We first compared the original variants of the algorithms with the variants selected by RDscan. The original set of variants consisted of variants that passed all filtering criteria provided by each variant caller. Overall, variant calling accuracies of four algorithms were improved after the variants were screened using RDscan. As shown in Table 2, after screening with RDscan, the call accuracies for SNVs increased in 10 out of 12 cases, and the call accuracies for indels increased in 11 out of 12 cases. Table 2 also shows that most variant callers had difficulty in producing high-accuracy results from samples with low tumor purity. In the case of SNVs, the F-score of the samples with low tumor purity (T:20%, N:100%) decreased by 3.9% for Strelka2, 15.9% for VarDict, 7% for NeuSomatic, and 4.6% for Mutect2 relative to samples with high tumor purity (T:80%, N:100%). Despite the difficulties of variant calling in samples with low tumor purity, RDscan achieved consistent improvement in accuracy over the variant callers we tested. In addition, even with normal samples contaminated by tumor cells (T:80%, N:90%), RDscan improved the accuracy of the variant callers.

[Figure 4 about here.]

Secondly, to show the overall improvement in the performance of these algorithms achieved by RDscan, we compared the changes in precision and recall of the variants according to the score of each algorithm with the changes resulting from application of RDscan to those variants. As shown in figure 4, for both SNVs and indels, RDscan improved the performance of the variant calls from Strelka2. For all three datasets, the overall performance of these four algorithms was improved when the variants were screened with RDscan (see Supplementary Fig. S2).

[Figure 5 about here.]

3.3. Somatic variant calling accuracy of ensemble models

To compare RDscan to the same level methods, the results of using the well-known variant callers as filters were compared with the results of RDscan. We analyzed the accuracies of ensemble models that combined two somatic variant calling algorithms (one as a variant caller and the other as a filter). Six ensemble models were created by combining the four algorithms. For each model with two algorithms, we generated a set of variants that passed the default criteria of both algorithms' filters, and then evaluated them.

We compared 14 results from six ensemble models, four existing algorithms, and four models

analyzed with RDscan. figure 5 shows the top three results for each dataset, in order of accuracy. In the case of SNVs, the combined of Strelka2 and RDscan had the highest accuracy in the two datasets with 80% tumor purity, and the combination of Mutect2 and RDscan had the highest accuracy in the dataset with 20% tumor purity. Indels exhibited similar results to SNVs, but in the case of the dataset with 80% tumor and 100% normal purity, the combination of Mutect2 and Strelka2 had the best accuracy, and the combination of Strelka2 and RDscan had the second highest accuracy. Note that RDscan effectively removes artificial variants, significantly improving variant call accuracy for indels in samples with low tumor purity, where it is difficult to distinguish true variants from artificial variants. The results for all models are in Supplementary Table S3.

3.4. Relationship between read depth coverage, indel length, RDscore, and variant calling accuracy

There are several parameters that influence call accuracy. The major factor would be depth coverage. The depth ranges of the current main technology are 30–70× for WGS and 100–150× for whole-exome sequence (WES). To assess the accuracy of RDscan as a function of depth coverage, we ran RDscan with candidate variant sets from the three germline variant callers using an additional dataset, HG002 (300×), with high depth coverage. Similar to the analysis in the previous section, we compared the original set of variants called and passed by each algorithm with the variants further filtered by RDscan. Supplementary Table S4 shows that additional screening with RDscan is ineffective for datasets with a high depth coverage. We will address this issue in the Discussions section.

RDscan assume that the depth distribution of the aligned read set containing a variant should not differ from the depth distribution of the whole read set mapped to the same region. However, the two distributions could be different for a long indel, because the sequencing quality and mapping quality scores of sequence reads with long indels may be lower than those of sequence reads without variants. To evaluate the RDscore changes according to an indel length, we calculated the RDscores of all heterozygous indels in the GIAB truth sets. Among the 263,034 indels, we analyzed 220,097 indels with at least one sequence read in the HG001 (34×) bam file. Supplementary Fig. S3 shows that the RDscore slightly decreases as the length of indel increases. However, the medians of the RDscores are between 0.8 and 1.0 for all indel lengths, which can be distinguished from artificial variants (In this study, we assumed that germline variants with RDscore < 0.5 were artificial variants).

4. DISCUSSION

Numerous methods have been developed for detecting variants from NGS data. Although the existing variant calling algorithms are generally accurate and reliable, most of them still have room for improvement in terms of accuracy for the variants with low VAF. RDscan removes misaligned reads and repositions reads, and then calculates *RDscore* based on the read depth distribution. By adopting this score, the accuracy of SNV/indel calls can be improved relative to the results obtained using existing variant callers.

Although RDscan can improve variant calling accuracy in most cases, it is important to use it with an understanding of the characteristics of *RDscore*. First, the *RDscores* of the true-positive variants are very densely distributed over a particular value. For example, 98.65% (3,076,963 out of 3,118,927) of the true-positive variants called by GATK HaplotypeCaller for SNVs, have *RDscore* > 0.64. Because variants with *RDscore* > 0.64 already have very high reliability in terms of the correlation (> 0.8) of read depth distribution, variant filtering using *RDscore* > 0.64 can decrease the sensitivity of variant calls. Second, variant filtering with low *RDscore* increases the variant call precision while minimizing the reduction in sensitivity. In particular, in the case of somatic variants in figure 4, precision increases rapidly as *RDscore* increases (blue dots mean *RDscore* = 0, red dots mean *RDscore* = 0.3). Even with a loose *RDscore* criterion, RDscan can find many false calls not identified by existing methods. In this study, we used *RDscore* criteria of 0.5 and 0.3 for germline and somatic variants, respectively. The *RDscore* can range from 0 to 1, but we recommend using an *RDscore* < 0.64. Third, Supplementary Table S4 shows that although RDscan can still increase the precision of variant calls, there is little room for precision improvement, resulting in a slightly decrease in the variant call accuracy for deep sequencing data. For example, the precision of DeepVariant's SNV and indels calls were already very high, at 99.88% and 99.82%, respectively. Therefore, we recommend using RDscan for typical sequencing data with relatively low read depth coverage, rather than for deep sequencing data.

In the last decade, many algorithms have been introduced to achieve high performance using features of NGS data such as read depth, base and mapping quality score, strand bias, and haplotype. In this paper we propose a new feature, read depth distribution. To validate the effectiveness of the read distribution, we first developed our own variant caller (standalone) based on read depth distribution. The accuracy of the variant caller was comparable to, but not better than, the state-of-art variant callers. However, we demonstrated that read depth distribution can increase the accuracy of variat calls when used with other features. Read depth distribution can be used as a key feature of

genomic analysis, along with existing NGS data features.

5. CONCLUSION

RDscan is an SNV/indels filtering tool based on the read depth distribution of an NGS dataset. In this study, we showed that our method could improve the accuracy of germline and somatic variant calls from NGS data. In addition, RDscan is simple to use and can effectively remove false-positive variants while maintaining a low computation load. Therefore, RDscan, along with existing variant callers, will contribute to improvement in genome analysis. Future work should seek to develop new methods based on read depth distribution for sequence analysis other than variant calling and haplotyping.

AVAILABILITY OF DATA AND IMPLEMENTATION

Source code and binaries, implemented in C++ and supported on Linux, are available for download at https://github.com/satchellhong/RDscan.The sequencing data for HG001-HG002 and in silico mixtures used in this study are from the 1000 Genomes Project (http://www.internationalgenome.org/) and the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/), respectively. Detailed descriptions about data can be found in supplementary materials.

ACKNOWLEDGEMENTS

The authors also thank all members of the Precision Medicine Support Center at Inha University Hospital for their generous support in testing the NGS data.

AUTHORS' CONTRIBUTIONS

Sunho Lee: Conceptualization; methodology; software; writing - original draft; writing - review and editing. **Seokchol Hong**: Software; visualization; writing – original draft; writing – review and editing. **Jonathan Woo, Jae-Hak Lee**: Formal analysis; writing – original draft. **Kyunghee Kim**: Data curation; formal analysis. **Lucia Kim**: Conceptualization; resources. **Kunsoo Park**: Conceptualization; methodology. **Jongsun Jung**: Conceptualization; methodology; supervision.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exists.

FUNDING INFORMATION

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2019M3E5D4064683).

SUPPLEMENTARY MATERIALS

Supplementary Figure S1

Supplementary Figure S2

Supplementary Figure S3

Supplementary Table S1

Supplementary Table S2

Supplementary Table S3

Supplementary Table S4

REFERENCES

Chapman, B. et al. 2021. bcbio/bcbio-nextgen. Available at: https://github.com/bcbio/bcbio-nextgen. Accessed October 21, 2020. doi: 10.5281/zenodo.4686097.

Cho, Y., Lee, S., Hong, J.H., et al. 2018. Development of the variant calling algorithm, ADIscan, and its use to estimate discordant sequences between monozygotic twins. *Nucleic Acids Res*. 46(15), 92–92.

Cibulskis, K., Lawrence, M.S., Carter, S.L., et al. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*. 31(3), 213–219.

DePristo, M.A., Banks, E., Poplin, R., et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 43(5), 491–498.

Do, H. and Dobrovic, A. 2015. Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization. *Clinical Chemistry*. 61(1), 64–71.

Garrison, E. and Marth, G. 2012. Haplotype-based variant detection from short-read sequencing. Available at: https://arxiv.org/abs/1207.3907v2. Accessed October 20, 2020.

Ka, S., Lee, S., Hong, J., et al. 2017. HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC bioinformatics*. 18(1), 258–258.

Kim, S., Scheffler, K., Halpern, A.L., et al. 2018. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*. 15(8), 591–594.

King, R.C., Stansfield, W.D., and Mulligan, P.K., eds. 2006. *A dictionary of genetics*, 7th ed. Oxford University Press.

Koboldt, D.C., Larson, D.E., and Wilson, R.K. 2013. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Current protocols in bioinformatics*. 44(1), 15-4

Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., Francisco, M., Moore, B.L., et al. 2019. Best practices for benchmarking germline small-variant calls in human genomes. *Nature biotechnology*. 37(5), 555-560.

Krøigård, A.B., Thomassen, M., Lænkholm, A.V., et al. 2016. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLOS ONE*. 11(3), e0151664–e0151664.

Lai, Z., Markovets, A., Ahdesmaki, M., et al. 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*. 44(11), e108–e108.

Lee, H., Deignan, J.L., Dorrani, N., et al. 2014. Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA*. 312(18), 1880–1880.

Lek, M., Kaczewski, K.J., Minikel, E.V., et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 536(7616), 285–291.

Li, H. 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*. 28(14), 1838–1844.

Li, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*. 30(20), 2843–2851.

Ng, S.B., Buckingham, K.J., Lee, C., et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*. 42(1), 30–35.

Poplin, R., Chang, P.C., Alexander, D., et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*. 36(10), 983–987.

Rimmer, A., Phan, H., et al. 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*. 46(8), 912–918.

Roth, A., Ding, J., Morin, R., et al. 2012. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*. 28(7), 907–913.

Sahraeian, S.M.E., Liu, R., Lau, B., et al. 2019. Deep convolutional neural networks for accurate somatic mutation detection. *Nature Communications*. 10(1), 1–10.

Shiraishi, Y., Sato, Y., Chiba, K., et al. 2013. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res*. 41(7), 89–89.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*. 526(7571), 68–74.

Yi, X., Liang, Y., Huerta-Sanchez, E., et al. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*. 329(5987), 75–78.

Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*. 3(1), 1-26.

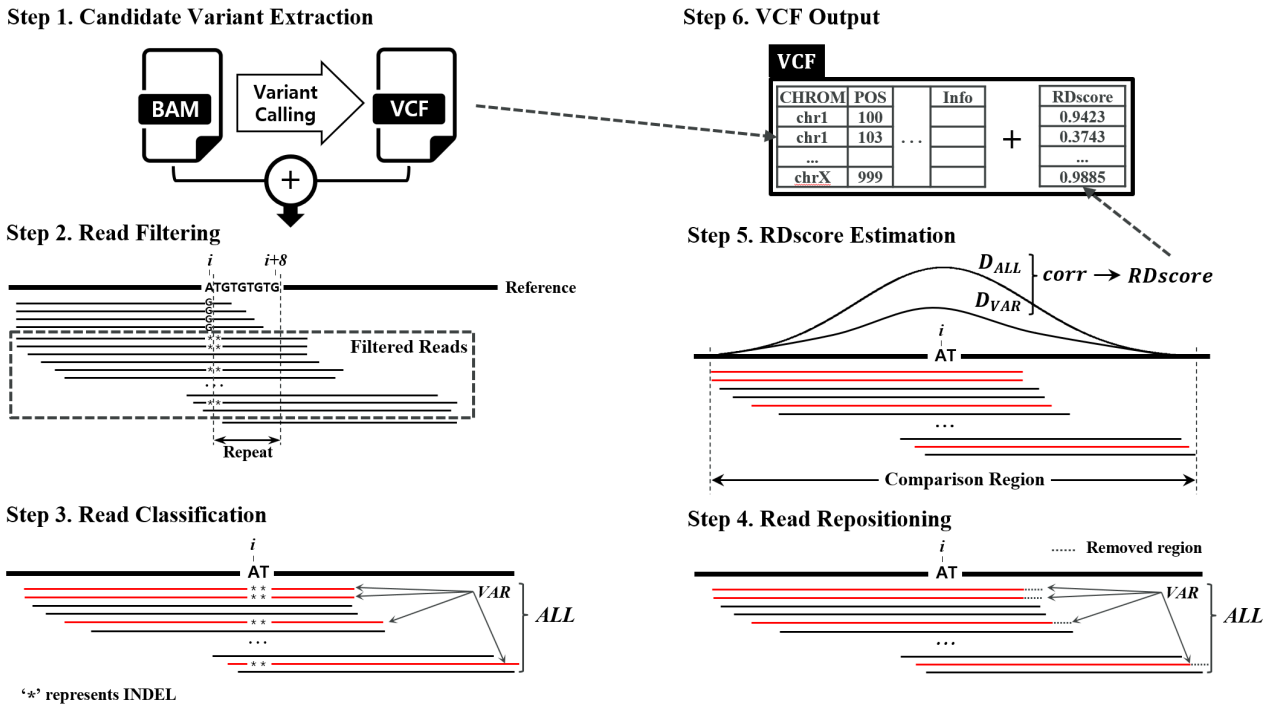

FIG. 1. Overview of RDscan based on read depth distribution. The workflow of RDscan is summarized in six steps. In Step 1, a variant caller produces a VCF file of candidate variants. Among the candidate variants, this figure deals with AT deletion variants occurring at locus *i*. For the reads mapped to locus *i*, in Step 2, RDscan filters the misaligned reads to ensure call accuracy. Reads that start or end in a repeat region are considered misaligned. In this example, the repeat region consists of a nucleotide sequence that repeats TG four times, and the top four reads and the bottom read are removed. In Step 3, all remaining reads are tagged as *ALL*, and all the reads containing the variant are tagged as *VAR*. If the variant is an indel, as shown in Step 4, RDscan adjusts the read alignment to eliminate alignment region length differences between *ALL* and *VAR*. Finally, in Step 5, RDscan estimates the *RDscore* for the AT deletion variant, which represents the correlation depths of *ALL* and *VAR* in the Comparison Region. The Comparison Region is the area in which the reads in *ALL* are aligned. For each candidate variant, the estimated *RDscore* is recorded in the original VCF file through iterations of Steps 2 through 5.

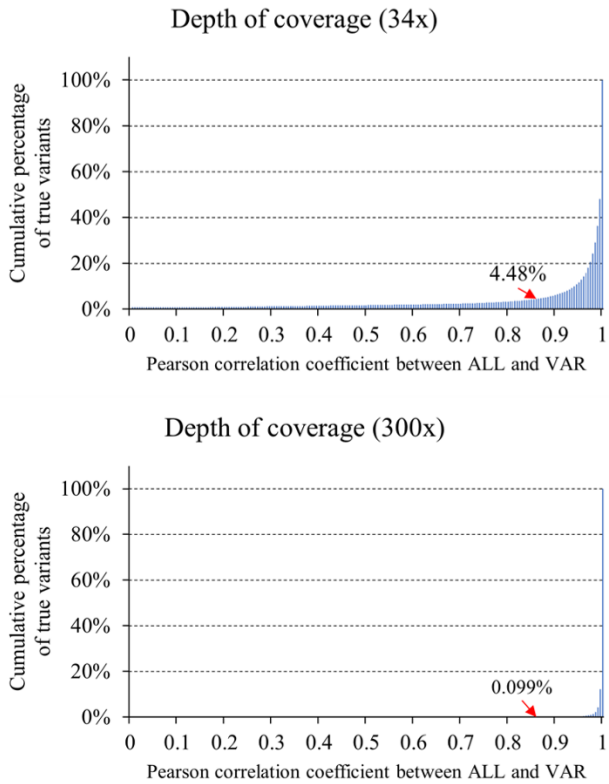

FIG. 2. The cumulative distributions of the Pearson correlation coefficients between *ALL* and *VAR* of the true-positive variants (3,544,295) within the two NGS data with coverage of 34x and 300x for the NA12878 sample.

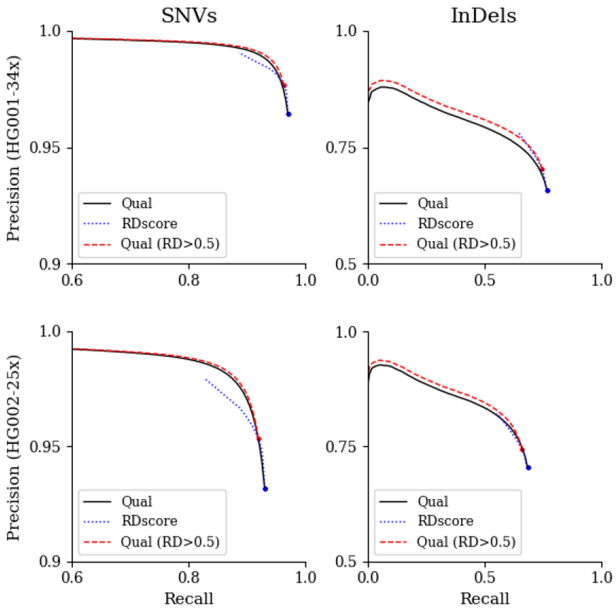

FIG. 3. Variant calling accuracy changes according to DeepVariant and RDscan scores. For two datasets (HG001 and HG002), ROC curve shows the variation in recall and precision of the variant call with a scoring parameter. The scoring parameters used to generate the solid and dotted curves were Qual (DeepVariant) and *RDscore* (RDscan), respectively. The dashed line, Qual (RD>0.5), show the results of applying *RDscore* > 0.5 to the sets of variants according to the Qual parameter of DeepVariant (solid line).

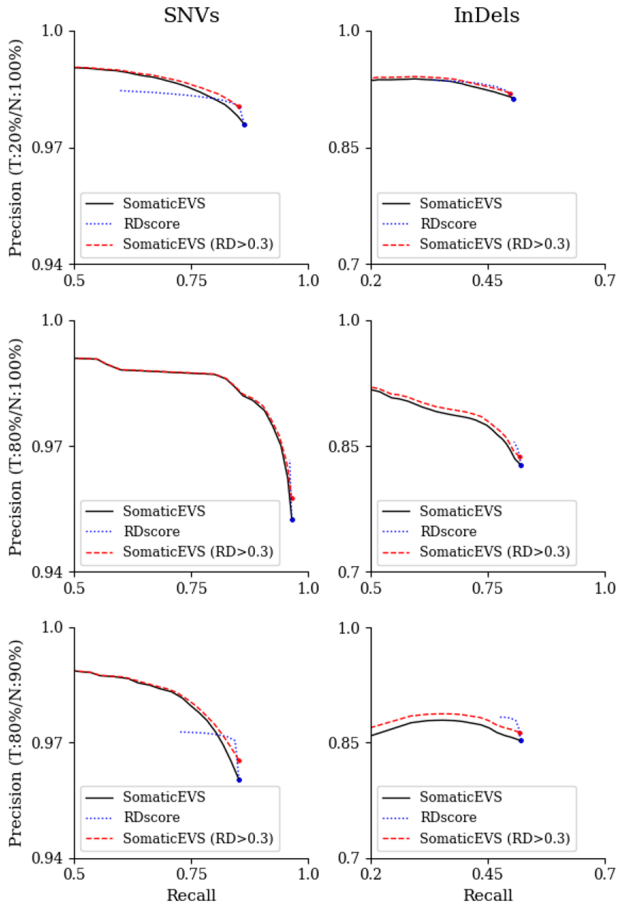

FIG. 4. Changes in somatic variant calling accuracy according to Strelka2 and RDscan scores. For three datasets, ROC curve shows the change of recall and precision for the variant calls according to a scoring parameter. The scoring parameters used to generate the solid and dotted curves were SomaticEVS (Strelka2) and *RDscore* (RDscan), respectively. The dashed line, SomaticEVS (RD>0.3), shows the results of applying $RDscore_{somatic} > 0.3$ for the sets of variants according to SomaticEVS of Strelka2 (solid line).

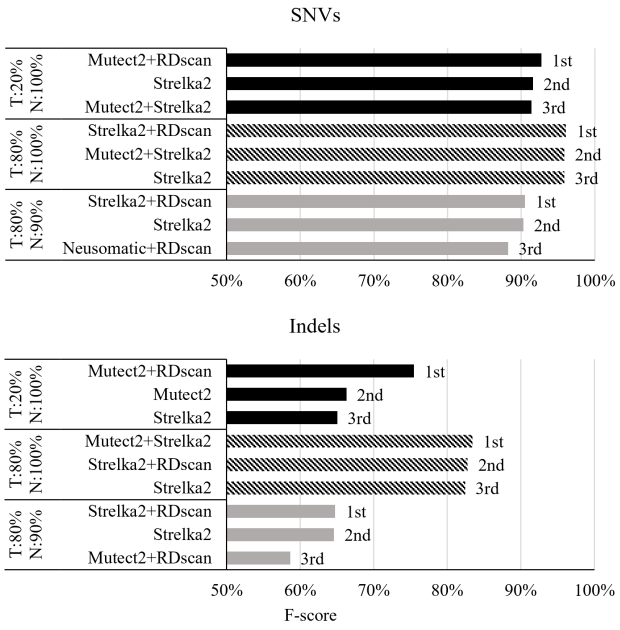

FIG. 5. Somatic variant calling accuracies of ensemble models. To evaluate the performance of ensemble models, 14 models (six ensemble models, four models analyzed with RDscan, and four existing algorithms) were used along with three datasets with proportions of tumor and normal DNA purity of 20% and 100%, 80% and 100%, and 80% and 90%. This figure shows the top three models in order of SNV/indels call accuracy for each dataset.

TABLE 1. Germline variant calling accuracy

Dataset	Variant Caller	Filter	SNVs						Indels					
			True Positive	False Positive	False Negative	Rec. (%)	Pre. (%)	F-scr. (%)	True Positive	False Positive	False Negative	Rec. (%)	Pre. (%)	F-scr. (%)
HG001 (34x)	Strelka2 v2.9.10	PASS	2900173	438358	309142	90.4	86.9	**88.6**	359961	145261	121879	74.7	71.2	**72.9**
		PASS+RD	2881124	133573	328191	89.8	95.6	**92.6**	349426	122261	132414	72.5	74.1	**73.3**
	GATK Haplotype Caller v4.1.8.1	PASS	3118927	139993	90388	97.2	95.7	**96.4**	342216	219032	139624	71.0	61.0	**65.6**
		PASS+RD	3095699	64219	113616	96.5	98.0	**97.2**	331283	155762	150557	68.8	68.0	**68.4**
	Deep Variant v1.0.0	PASS	3112316	115512	96999	97.0	96.4	**96.7**	369765	196180	112075	76.7	65.3	**70.6**
		PASS+RD	3089778	73911	119537	96.3	97.7	**97.0**	360009	155192	121831	74.7	69.9	**72.2**
HG002 (25x)	Strelka2 v2.9.10	PASS	2788428	727723	564390	83.2	79.3	**81.2**	319335	52059	203699	61.1	86.0	71.4
		PASS+RD	2759569	252484	593249	82.3	91.6	**86.7**	310899	45231	212135	59.4	87.3	70.7
	GATK Haplotype Caller v4.1.8.1	PASS	3094172	256578	258646	92.3	92.3	**92.3**	324496	180164	198538	62.0	64.3	**63.2**
		PASS+RD	3057662	123573	295156	91.2	96.1	**93.6**	310722	117435	212312	59.4	72.6	**65.3**
	Deep Variant v1.0.0	PASS	3117943	227805	234875	93.0	93.2	**93.1**	358205	152505	164829	68.5	70.1	**69.3**
		PASS+RD	3082064	150850	270754	91.9	95.3	**93.6**	345302	120230	177732	66.0	74.2	**69.9**

Recall, precision, and F-score are expressed as Rec., Pre., and F-scr., respectively. PASS is the set of variants that passed all filtering criteria provided by each variant caller, and PASS+RD is the set of variants in PASS that passed additional filtering with RDscan. F-scores are calculated by the following equation: $2 \times \frac{(precision \times recall)}{(precision + recall)}$. Bold font indicates that the accuracy of the variants selected by RDscan is higher than the original accuracy of each algorithm.

TABLE 2. Somatic variant calling accuracy

Dataset	Variant Caller	Filter	SNVs						Indels					
			True Positive	False Positive	False Negative	Rec. (%)	Pre. (%)	F-scr. (%)	True Positive	False Positive	False Negative	Rec. (%)	Pre. (%)	F-scr. (%)
Tumor 20% (~110x) Normal 100% (~37x)	Strelka2 v2.9.10	PASS	1092894	26851	173792	86.3	97.6	91.6	102895	9830	100818	50.5	91.3	65.0
		PASS+RD	1078316	21521	188370	85.1	98.0	91.1	101324	8869	102389	49.7	92.0	64.6
	VarDict v1.8.2	PASS	688675	16065	578011	54.4	97.7	**69.9**	46866	8372	156847	23.0	84.8	**36.2**
		PASS+RD	687062	11124	579624	54.2	98.4	**69.9**	46447	2872	157266	22.8	94.2	**36.7**
	NeuSomatic v0.2.1	PASS	976872	12082	289814	77.1	98.8	86.6	52349	3079	151364	25.7	94.4	**40.4**
		PASS+RD	971781	11180	294905	76.7	98.9	86.4	52020	1132	151693	25.5	97.9	**40.5**
	Mutect2 v4.1.1.0	PASS	1167087	147578	99599	92.1	88.8	**90.4**	166506	132258	37207	81.7	55.7	**66.3**
		PASS+RD	1150492	64552	116194	90.8	94.7	**92.7**	154366	51034	49347	75.8	75.2	**75.5**
Tumor 80% (~110x) Normal 100% (~37x)	Strelka2 v2.9.10	PASS	1221652	60742	45034	96.4	95.3	**95.9**	167249	34894	36464	82.1	82.7	**82.4**
		PASS+RD	1221070	54032	45616	96.4	95.8	**96.1**	166715	32429	36998	81.8	83.7	**82.8**
	VarDict v1.8.2	PASS	1099023	33498	167663	86.8	97.0	**91.6**	97715	29101	105998	48.0	77.1	**59.1**
		PASS+RD	1098606	26965	168080	86.7	97.6	**91.8**	97419	22277	106294	47.8	81.4	**60.2**
	NeuSomatic v0.2.1	PASS	1183322	42631	83364	93.4	96.5	**94.9**	97764	14630	105949	48.0	87.0	**61.9**
		PASS+RD	1182991	32205	83695	93.4	97.3	**95.3**	97740	7210	105973	48.0	93.1	**63.3**
	Mutect2 v4.1.1.0	PASS	1194772	59587	71914	94.3	95.2	**94.8**	171710	83012	32003	84.3	67.4	**74.9**
		PASS+RD	1194272	46224	72414	94.3	96.3	**95.3**	168325	55815	35388	82.6	75.1	**78.7**
Tumor 80% (~110x) Normal 90% (~37x)	Strelka2 v2.9.10	PASS	1079088	44545	187598	85.2	96.0	**90.3**	105880	18254	97833	52.0	85.3	**64.6**
		PASS+RD	1078587	38544	188099	85.2	96.5	**90.5**	105528	16695	98185	51.8	86.3	**64.8**
	VarDict v1.8.2	PASS	485305	22976	781381	38.3	95.5	**54.7**	47676	18266	156037	23.4	72.3	**35.4**
		PASS+RD	484945	16619	781741	38.3	96.7	**54.9**	47401	11399	156312	23.3	80.6	**36.1**
	NeuSomatic v0.2.1	PASS	1016718	59879	249968	80.3	94.4	**86.8**	61454	6590	142259	30.2	90.3	**45.2**
		PASS+RD	1016460	21555	250226	80.2	97.9	**88.2**	61439	3102	142274	30.2	95.2	**45.8**
	Mutect2 v4.1.1.0	PAA	584814	30393	681872	46.2	95.1	**62.2**	98159	43597	105554	48.2	69.2	**56.8**
		PASS+RD	584480	22219	682206	46.1	96.3	**62.4**	96282	28482	107431	47.3	77.2	**58.6**

Recall, precision, and F-score are expressed as Rec., Pre., and F-scr., respectively. PASS is the set of variants that passed all filtering criteria provided by each variant caller, and PASS+RD is the set of variants in PASS that passed additional filtering with RDscan. F-scores are calculated by the following equation: $2 \times \frac{(precision \times recall)}{(precision + recall)}$. Bold font indicates that the accuracy of the variants selected by RDscan is higher than the original accuracy of each algorithm.