



Comparative Analysis for the Performance of Variant Calling Pipelines on Detecting the *de novo* Mutations in Humans

Yu Liang^{1†}, Li He^{2†}, Yiru Zhao³, Yinyi Hao¹, Yifan Zhou¹, Menglong Li¹, Chuan Li³, Xuemei Pu^{1*} and Zhining Wen^{1*}

¹ College of Chemistry, Sichuan University, Chengdu, China, ² Biogas Appliance Quality Supervision and Inspection Center, Biogas Institute of Ministry of Agriculture, Chengdu, China, ³ College of Computer Science, Sichuan University, Chengdu, China

OPEN ACCESS

Edited by:

Zhichao Liu,
The National Center for Toxicological
Research (FDA), United States

Reviewed by:

Dan Li,
University of Arkansas at Little Rock,
United States
Arun Samidurai,
Virginia Commonwealth University,
United States

*Correspondence:

Xuemei Pu
xmpuscu@scu.edu.cn
Zhining Wen
w_zhining@163.com

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Translational Pharmacology,
a section of the journal
Frontiers in Pharmacology

Received: 11 December 2018

Accepted: 21 March 2019

Published: 11 April 2019

Citation:

Liang Y, He L, Zhao Y, Hao Y,
Zhou Y, Li M, Li C, Pu X and Wen Z
(2019) Comparative Analysis
for the Performance of Variant Calling
Pipelines on Detecting the *de novo*
Mutations in Humans.
Front. Pharmacol. 10:358.
doi: 10.3389/fphar.2019.00358

Despite of the low occurrence rate in the entire genomes, *de novo* mutation is proved to be deleterious and will lead to severe genetic diseases via impacting on the gene function. Considering the fact that the traditional family based linkage approaches and the genome-wide association studies are unsuitable for identifying the *de novo* mutations, in recent years, several pipelines have been proposed to detect them based on the whole-genome or whole-exome sequencing data and were used for calling them in the rare diseases. However, how the performance of these variant calling pipelines on detecting the *de novo* mutations is still unexplored. For the purpose of facilitating the appropriate choice of the pipelines and reducing the false positive rate, in this study, we thoroughly evaluated the performance of the commonly used trio calling methods on the detection of the *de novo* single-nucleotide variants (DNSNVs) by conducting a comparative analysis for the calling results. Our results exhibited that different pipelines have a specific tendency to detect the DNSNVs in the genomic regions with different GC contents. Additionally, to refine the calling results for a single pipeline, our proposed filter achieved satisfied results, indicating that the read coverage at the mutation positions can be used as an effective index to identify the high-confidence DNSNVs. Our findings should be good support for the committees to choose an appropriate way to explore the *de novo* mutations for the rare diseases.

Keywords: *de novo* mutation, rare diseases, variant calling pipelines evaluation, gene function, whole-exon sequencing

INTRODUCTION

The genomic structural variations, such as single-nucleotide variants (SNVs), copy-number variants (CNV) and the indels, play important roles in the genetic diseases. The researches in the past decade have discovered the landscape of SNVs in human and the strong causality between the SNVs and the genetic diseases (Ku et al., 2012; Veltman and Brunner, 2012; Boycott et al., 2013). Among the SNVs, the occurrence frequency of *de novo* SNVs (DNSNVs) in germline is as

Abbreviations: AJ, ashkenazi jewish; CNV, copy-number variants; DNSNVs, *de novo* single-nucleotide variants; GATK, genomeanalysis toolkit; GIAB, genome in a bottle; OMIM, online mendelian inheritance in man; PCR, polymerase chain reaction; SNVs, single-nucleotide variants; Ti/Tv, the transition/transversion ratio; WES, whole-exome sequencing; WGS, whole-genomes sequencing.

low as $1.0\sim 3.0 \times 10^{-8}$ SNVs per site per generation (Conrad et al., 2012; Veltman and Brunner, 2012), but this type of mutation is proved to be deleterious and will lead to severe genetic diseases via impacting on the different gene functions. It had been reported in recent studies that the rare sporadic malformation syndromes (Hoischen et al., 2010a,b; Ng et al., 2010) as well as the neurodevelopmental diseases (Hamdan et al., 2014; Turner et al., 2017) were primarily caused by the DNSNVs in single specific genes or a set of genes, elaborating the fundamentality of the *de novo* mutations in the genetic diseases despite of the unclear underlying mechanisms. Therefore, accurately identifying the *de novo* mutations located in the rare-disease-causing genes can be great helpful not only for improving the clinical diagnostics, but also for better understanding the mechanisms in the rare genetic diseases.

Due to the fact that the traditional family based linkage approaches and the genome-wide association studies were unsuited to the detection of *de novo* mutations, the emerging next-generation sequencing technologies such as the whole-genomes sequencing (WGS)/whole-exome sequencing (WES) began to be applied in the researches of genetic diseases (Boycott et al., 2013; Peters et al., 2015; Jin et al., 2017; Turner et al., 2017; Hyrenius-Wittsten et al., 2018). A number of bioinformatics pipelines have subsequently been proposed to call the *de novo* mutations based on the WGS/WES data (McKenna et al., 2010; Li et al., 2012; Koboldt et al., 2013; Kojima et al., 2013; Peng et al., 2013; Ramu et al., 2013; Cleary et al., 2014; Salzberg et al., 2014; Santoni et al., 2014; He et al., 2015; Wei et al., 2015; Francioli et al., 2017; Gomez-Romero et al., 2018; Zhou et al., 2018). The heated discussions have been carried on in recent years about applying these approaches in the diagnostics of rare diseases and the potential clinical implementations (Yang et al., 2013; Lee et al., 2014; Jamuar and Tan, 2015; Bacchelli and Williams, 2016; Krier et al., 2016; Thiffault and Lantos, 2016). However, the occurrence of germline *de novo* mutations is much lower than that of the inherited variations, resulting in the difficulty of discriminating these variants from the noise derived from the procedures of sequencing, reads mapping as well as the variant calling and annotation. So far how the performance of these variant calling pipelines on detecting the *de novo* mutations is still unexplored.

Therefore, in our study, we thoroughly investigated three commonly used trio calling pipelines named GATK (McKenna et al., 2010; Francioli et al., 2017), RTG (Cleary et al., 2014) and VarScan (Koboldt et al., 2013) on the detection of the DNSNVs and found that GATK can detect the DNSNVs in the low GC-content region with a relative low error rate while RTG and VarScan are more suitable for detecting the DNSNVs in the high GC-content region. In refining the calling results of a single pipeline, our proposed filter not only effectively excluded the redundant DNSNVs, but also ensured the transitions/transversions ratio of the results, indicating that the read coverage at the mutation positions of the son's genome and the parents' genomes can be an important index for evaluating the quality of DNSNVs.

MATERIALS AND METHODS

Dataset

The WES data of the Ashkenazi Jewish (AJ) trio set (NA12878) were applied in our study (Zook et al., 2014, 2016). The preprocessed BAM files of the mother (HG004), the father (HG003) and the son (HG002), for which the data preprocessing steps including the reads alignment and duplicates marking had been conducted beforehand, were directly downloaded from the Genome in a Bottle (GIAB) website¹ and used for calling the *de novo* mutations. In order to facilitate the comparison of the similarities and differences of the pipelines and make the point clear, in our study, we only focused on the *de novo* SNVs on the autosomes. It is worth noting that the structure variations, e.g., indels, as well as the mutations on the chromosomes X and Y are also important for the genetic diseases, but it will make the problem more complicated when involving them in the comparisons.

Trio Calling Pipelines

Three pipelines, namely GenomeAnalysisToolKit (version 4.0.5.2) (McKenna et al., 2010; Francioli et al., 2017), RTG (non-commercial version 3.9.1) (Cleary et al., 2014) and VarScan (version 2.3.9) (Koboldt et al., 2013), were applied in this study to call the DNSNVs. In GATK pipeline, the gVCF files for the trio samples were firstly generated by using *HaplotypeCaller* separately and combined into a multi-sample gVCF file through *CombineGVCFs*. Then, the raw SNVs were called by *GenotypeGVCFs* and recalibrated by *VariantRecalibrator* and *ApplyVQSR*. Finally, after deriving the posterior probabilities of the genotypes by *CalculateGenotypePosteriors*, the low quality genotypes were filtered out by *VariantFiltration* and the *de novo* SNVs were annotated by *VariantAnnotator*. In RTG pipeline, the quality calibration files for the BAM files of trio samples were firstly prepared and applied in the subsequent calling procedures. The *de novo* SNVs were called and extracted by RTG with two parameters of *family* and *vcfilter*. As to the Varscan pipeline, the BAM files of trio samples were firstly sorted and combined into a three-sample pileup file. Subsequently, the *de novo* SNVs were called by running the *VarScan trio* command.

Metrics for the Comparison of the Pipelines

We used three metrics namely GC content (Shin et al., 2013), substitution type (Zook et al., 2014) and SNV density (Choi et al., 2018), to evaluate the differences in the calling results generated by three pipelines. A genomic sequence of 100 bases centered on a DNSNV was extracted from the reference genome and the GC content can be calculated via the following equ.

$$\text{GC content (\%)} = \frac{\text{number of bases G and C}}{100} \times 100\% \quad (1)$$

The substitution type includes the point mutations of transition and transversion. The transition refers to the nucleotide changes

¹<ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/>

from a purine to another purine (A↔G) or a pyrimidine to another pyrimidine (C↔T), which occurs more frequently than the transversion in the SNPs. The transversion refers to the changes from purine to pyrimidine, or vice versa (A↔T, A↔C, G↔T, and C↔C). By counting the total number of SNVs in the genomic sequence of 100 bases centered on a DNSNV, the SNV density can be calculated via the equ. (2):

$$\text{SNV density (\%)} = \frac{\text{number of SNVs}}{100} \times 100\% \quad (2)$$

Definition of DNSNV Filter

To distinguish the high-confidence DNSNVs from the noise, the read coverage at the mutation point is an important indicator for this purpose. Based on the concept of signal-to-noise ratio, we not only expect that the number of reads mapped to the reference genome at the mutation point for the parents' datasets is high than that mapped to the mutated sequence as much as possible, but also expect that the number of reads mapped to the mutated sequence for the son's dataset is higher than that mapped to the reference genome. Therefore, to refine the calling results from the pipelines, we proposed a DNSNV filter by considering both read coverage at the mutation point in the son's dataset and the parents' datasets. The filter was defined as the following equ:

$$\text{Score} = \log \left(\frac{|F_{\text{ref}} - F_{\text{alt}}| + |M_{\text{ref}} - M_{\text{alt}}|}{2 \times |S_{\text{alt}} - S_{\text{ref}}|} \right) \quad (3)$$

Where the F_{ref} , M_{ref} , and S_{ref} indicate the number of reads mapped to the reference genome at the i th mutation point for the datasets of father, mother and son, respectively. The F_{alt} , M_{alt} , and S_{alt} indicate the number of reads mapped to the mutated sequence at the i th mutation point for the datasets of father, mother and son, respectively. We assigned the scores for the DNSNVs and rank them by their scores. In this study, we took a non-stringent score (score = 0) as the threshold to investigate the improvement of the calling results.

Genotype Quality

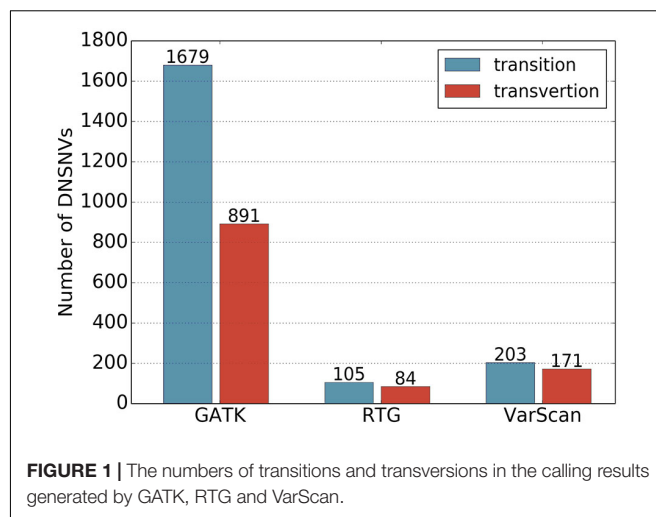
Genotype Quality (GQ) (Zhang et al., 2013) is used to evaluate the filtering results of DNSNVs, which indicates the quality value of the most likely genotype. The quality value refers the possibility of the genotype being present at the site. The larger value means the greater the likelihood of the genotype.

RESULTS

To facilitate the appropriate choice of the trio calling pipelines for detecting the DNSNVs, in our study, we firstly evaluated the results of three commonly used pipelines named GATK, RTG, and VarScan by using the WES data of a father-mother-child trio set from Ashkenazi Jews, and then proposed a filter for removing the redundant DNSNVs from the calling results.

The Landscape of the Calling Results From Three Pipelines

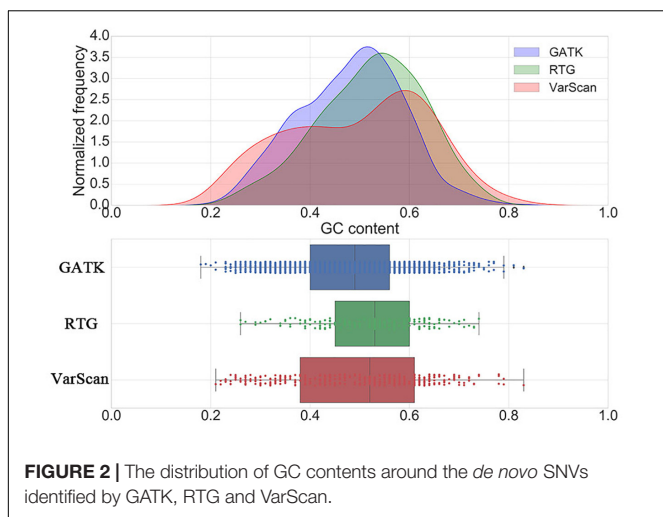
In total, 2570, 189 and 374 DNSNVs in the autosomes were identified by GATK, RTG and VarScan, respectively. In



terms of the quantity of DNSNVs, GATK exhibited higher detecting sensitivity than RTG and VarScan. For the point of mutation types, **Figure 1** showed clear difference between the numbers of transitions and transversions in the calling results of three pipelines. Generally the number of transitions was greater than that of the transversions for all the pipelines. The transition/transversion ratio (Ti/Tv) had been suggested in GIAB study (Zook et al., 2014) to be a metric for evaluating the quality of the novel calling variants based on the assumption that Ti/Tv in novel variants would be similar to that in the common variants. In our results, Ti/Tv for GATK was 1.88 (1679/891), which was higher than those for the other two pipelines [1.25 (105/84) and 1.19 (203/171) for RTG and VarScan, respectively], indicating a lower error rate in the GATK calling results. In addition, we mapped the DNSNVs to the dbSNP database² and separately calculated the overlap rates between the DNSNVs and all the variations, as well as between the DNSNVs and the common variations for the three pipelines (**Supplementary Figure S1**). The overlap rate was calculated by dividing the number of overlapped DNSNVs by the total number of DNSNVs identified by the pipeline. From the figure we can see that GATK achieved the highest overlap rate among three pipelines, indicating the least proportion of DNSNVs involved in the calling results. On the contrary, VarScan achieved the lowest overlap rate, indicating the largest proportion of DNSNVs involved in the calling results. Meanwhile, the Ti/Tv rates of three pipelines indicated the lowest error rate in the calling results of GATK and highest error rate in the results of VarScan. It suggested that VarScan and RTG tend to reveal more DNSNVs in their reported results, but the reliability of them needs to be more carefully validated by further experiments.

Considering the fact that with the increase of GC content, the difficulty of polymerase chain reaction (PCR) amplification in sequencing procedure will increase, which will result in an increase in the error rate of the calling results, we further investigated the distribution of GC contents around the DNSNVs

²<https://www.ncbi.nlm.nih.gov/snp/>



for three pipelines (Figure 2). The majority of the DNSNVs called by GATK located in the regions with GC contents less than 50% while the DNSNVs called by RTG mainly located in the regions with relatively high GC contents (> 50%). This may be the reason that GATK can yield the higher Ti/Tv ratio (1.88) than that (1.25) obtained by RTG. Interestingly, although VarScan yielded a relatively low Ti/Tv ratio (1.19), the distribution of the calling results from VarScan had two peaks that covered the regions with both low (~40%) and high (~60%) GC content. We separately inspected the point mutation types of the DNSNVs called by VarScan in the genome regions with GC content $\geq 50\%$ and GC content $< 50\%$. The Ti/Tv ratios for the high ($\geq 50\%$) and low ($< 50\%$) GC-content regions were 1.32 (116/88) and 1.05 (87/83), respectively. It may indicate that, for the high GC-content region, the VarScan can identify the DNSNVs with even lower error rate than RTG (Ti/Tv = 1.25). As to calling the DNSNVs in the low GC-content region, the performance of VarScan (Ti/Tv = 1.05) is inferior to GATK (Ti/Tv = 1.88).

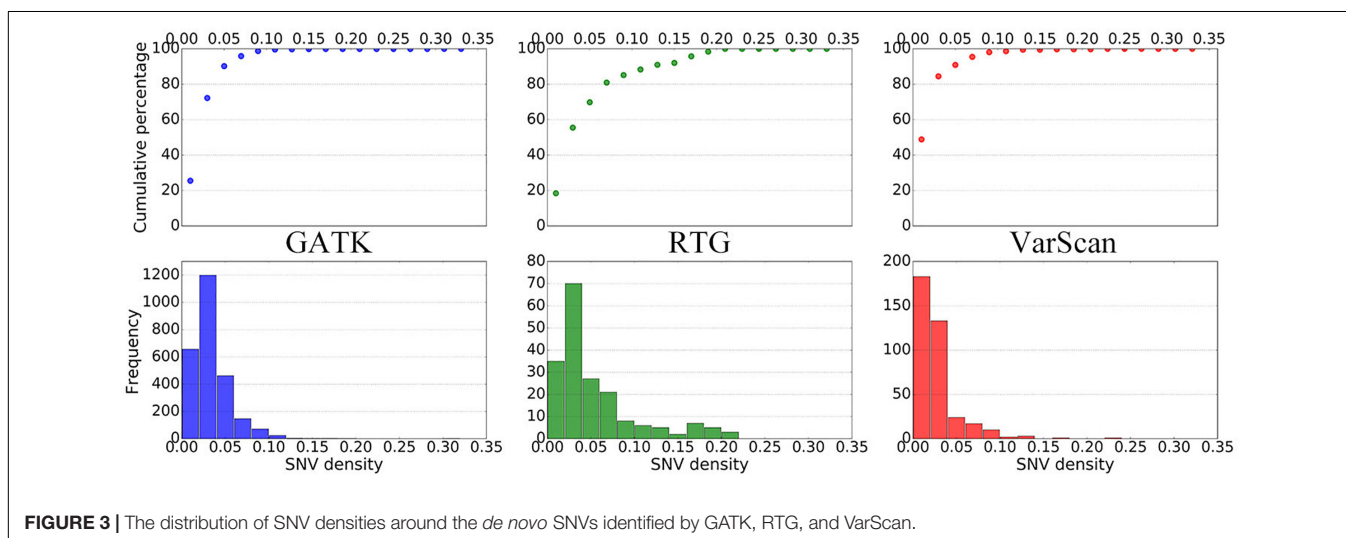
To further elucidate the performance of the three pipelines, we subsequently summarized the distributions of SNV densities

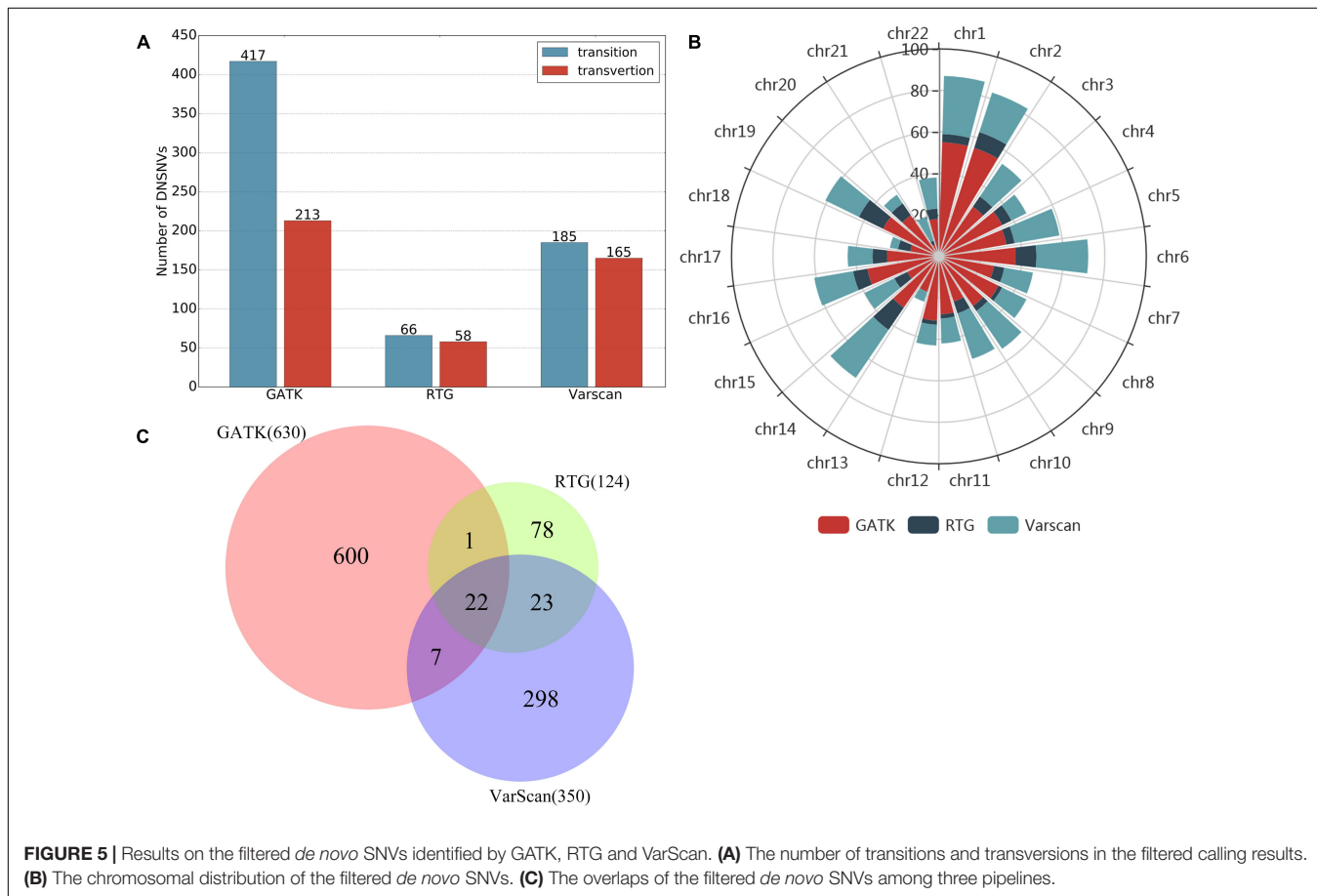
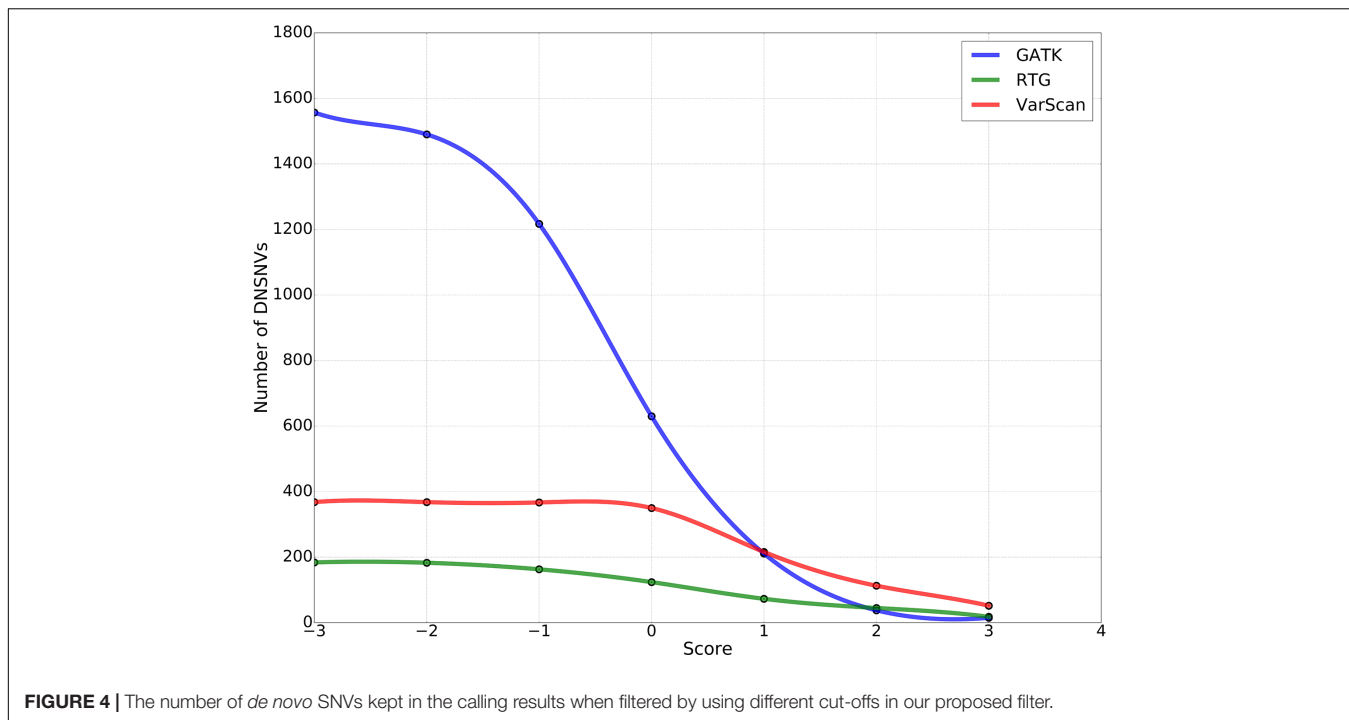
around the DNSNVs (Figure 3). As the SNV density increases, the error rate for calling DNSNVs may increase. It can be seen from the figure that, for the RTG pipeline, only ~70% DNSNVs located in the regions with the SNV density less than 5%, and the accumulated percentage gradually increased to 90% when counting the DNSNVs in the regions with the SNV density less than 15%, indicating a larger error rate may exist in the call results. Compared to RTG, over 90% DNSNVs called by GATK and VarScan located in the regions with the SNV density less than 5%, which may suggest the better quality of the identified DNSNVs.

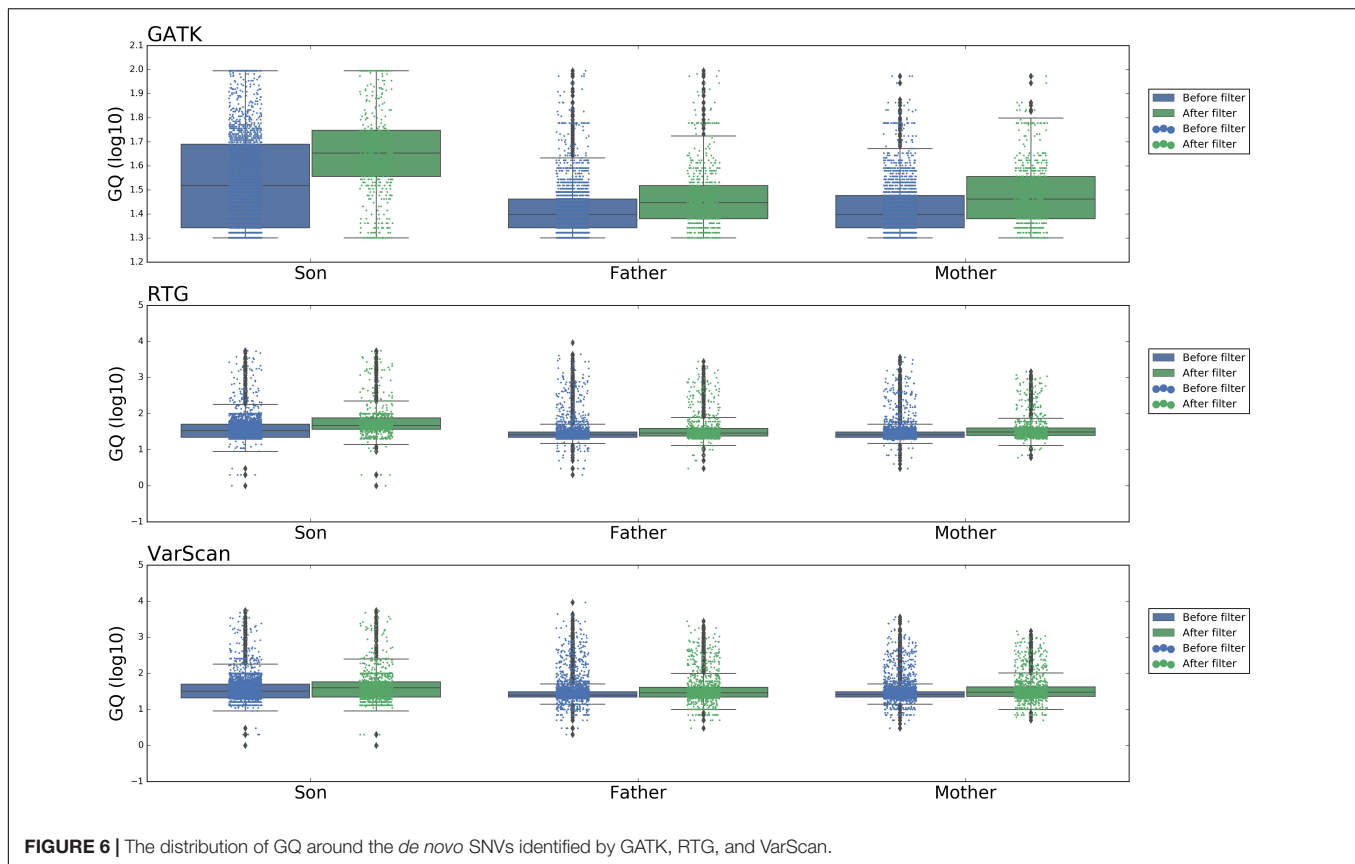
Performance of Our Proposed Filter on Refining the DNSNVs

To obtain the high-confidence DNSNVs, we suggested further refining the calling results of three pipelines by using a proposed filter, which was taken both read coverage at the mutation sites of the son's genome and the parents' genomes into consideration to identify the DNSNVs and can be directly applied to an individual pipeline without reference to the information of other pipelines. Figure 4 showed the numbers of DNSNVs kept by the filter when applying different cut-offs. The number of DNSNVs called by all the pipelines took on a tendency of descension when the cut-off became more stringent, especially for the GATK pipeline. The number of the left DNSNVs called by GATK dramatically decreased from over 1,500 to 15 as the value of the cut-off increased from -3 to 3.

In this study, we just took a non-stringent cut-off (score = 0) as an example for the comparison of the three pipelines. When filtering the DNSNVs with the score > 0 , the numbers of DNSNVs detected by GATK decreased from 2570 to 630. For RTG and VarScan, the numbers of DNSNVs decreased from 189 and 374 to 124 and 350, respectively. Figure 5A showed the number of transitions and transversions in the filtered calling results. The GATK pipeline still yielded the highest ratio (Ti/Tv = 1.96) among three pipelines, which was improved after filtering a number of DNSNVs from the calling results. For the pipelines of RTG and VarScan, since only a small number







of DNSNVs were removed, the quality of final DNSNVs were comparable to that of the original calling results, with a slight decrease in the ratio ($Ti/Tv = 1.14$ and 1.12 for RTG and VarScan, respectively). Our results may suggest that our proposed filter can efficiently reduce the error rate in the DNSNVs from the redundant calling results. In addition, for the calling results with less redundancy, our filter can well maintain the quality of the original results.

The chromosomal distribution of the filtered DNSNVs was shown in **Figure 5B**. The DNSNVs identified by three calling pipelines mainly located on the chromosomes 1, 2, 6, 14, 16, and 19, which indicates that the DNSNVs are more likely to occur on these six chromosomes. Conversely, the DNSNVs rarely occur on chromosomes 13, 18, and 21. **Figure 5C** showed the overlaps of the filtered DNSNVs among three pipelines. Because both RTG and VarScan were adapted for detecting the DNSNVs in the high GC-content region with relative low error rate, close to 36% (45/124) DNSNVs identified by RTG can be found in the results called by VarScan even if 65 out of 189 DNSNVs identified by RTG had been removed out by the filter. When comparing GATK with RTG and VarScan, only about 19% (23/124) and 8% (29/350) DNSNVs called by RTG and VarScan, respectively, were involved in the calling results of GATK. The reason may be that GATK was fit for calling the DNSNVs in the low GC-content region while the ability of RTG and VarScan to identify the DNSNVs in this region is inferior. This should be further validated by using the real disease

samples. Eventually, a total of 22 DNSNVs were detected by all the pipelines.

Figure 6 showed the distribution of DNSNVs GQ of the trio before and after filtering. For GATK, significant difference exists between the calling results before and after filtering. However, for RTG and VarScan, the difference is slight. But for the latter two pipelines, the filtered DNSNVs GQ distribution is still slightly better than that before filtering. In general, the GQ distributions of DNSNVs for the three pipelines after filtering are better than those before filtering.

Biological Relevance of the Overlapped DNSNVs Among Three Pipelines

The 22 overlapped DNSNVs can be detected by three pipelines simultaneously, and should be of relatively high confidence. To investigate the biological relevance of these DNSNVs, we firstly mapped them to the corresponding genes and then explored the associations of the genes with genetic diseases by the Online Mendelian Inheritance in Man (OMIM) database searching and literature survey. The description of 22 DNSNVs as well as the corresponding genes finally identified by all pipelines was listed in **Table 1**. In the results, six out of 22 genes has been proved to be directly associated with the rare genetic diseases. For example, the gene *DTNB* in chromosome 2 is an important protein-coding gene of beta-dystrobrevin. This protein is found to interact directly with dystrophin and the low expression level of it will

TABLE 1 | The 22 overlapped DNSNVs identified by all the trio calling methods and the corresponding genes associated with the diseases.

Chromosome	Position	Substitution	Gene symbol	phenotype
2	25457155	C→T	DTNB*	Muscular dystrophy
2	197791183	C→T	–	–
2	241835203	G→T	–	–
3	4669342	A→T	ITPR1*	Gillespie syndrome Spinocerebellar ataxia 15 Spinocerebellar ataxia 29, congenital non-progressive
3	48603870	G→A	UQCRC1	Predisposition of Alzheimer's disease
3	52547912	C→G	PBRM1*	Clear cell renal cell carcinoma
6	109740508	T→C	FIG4*	Polymicrogyria, bilateral temporooccipital Amyotrophic lateral sclerosis 11 Charcot-Marie-Tooth disease, type 4J Yunis-Varon syndrome
5	140730969	A→G	–	–
8	30703949	C→G	GSR*	Hemolytic anemia due to glutathione reductase deficiency
9	131041047	G→T	LAMC3*	Cortical malformations, occipital
9	131851319	C→G	–	–
11	82698724	T→C	–	–
14	77245340	C→T	TMEM63C	–
14	93170535	G→A	–	–
14	96180196	C→G	–	–
14	106330036	T→A	–	–
14	106993919	A→G	–	–
15	79852462	G→T	MTHFS	Heart defects; lung cancer
17	53638886	G→A	–	–
19	45720935	C→T	FBXO46	–
22	23029544	A→C	–	–
22	23029596	A→T	–	–

*The genes were reported to be directly associated with genetic diseases in the OMIM database.

cause severe Duchenne muscular dystrophy (Blake et al., 1998). The genes *FIG4* and *LAMC3* were reported to be highly correlated with the polymicrogyria and cortical malformations, respectively. Campeau et al. (2013) demonstrated that the inactivation of *FIG4* would result in the central nervous system dysfunction and extensive skeletal anomalies. The research by Barak et al. (2011) exhibited an important role of the gene *LAMC3* in cortical organization. It can be seen that to a certain extent our proposed filter can be helpful for removing the redundant DNSNVs from the calling results.

DISCUSSION

Nowadays, it has been proved that the *de novo* mutations played an important role in human genetic diseases. Based on the next-generation sequencing technology, more and more researches focused on detecting the *de novo* mutations in the rare genetic diseases as well as the potential applications in the clinics for the purpose of improving the clinical diagnosis and better understanding the mechanisms of the genetic diseases. However, it is still a tough work and remains challenging to accurately identify the genomic variants because of the complexity of the sequencing experiments and the variants calling procedures. In

the previous study, Reumers et al. (2012) suggested that an optimized filtering procedure would be helpful for reducing the error rate when detecting the genomic variants with the short-read sequencing data. Considering the fact that the occurrence rate of *de novo* mutations is much lower than those of inherited variations and somatic variants, it is more difficult to distinguish them from the errors in the whole genome. At present, a number of trio calling methods were proposed to detect the *de novo* mutations based on the WGS/WES data, but how the performance of these pipelines on detecting the *de novo* mutations is still unexplored. Therefore, by carefully comparing the results from three commonly used trio calling pipelines, we elucidated that the performance of the three pipelines on calling DNSNVs in high or low GC-content region was different. In addition, based on the read coverage, our proposed filter can be well applied to the single pipeline for refining the calling results.

In this study, we analyzed the calling results from three pipelines named GATK, RTG and VarScan. Generally speaking, GATK can identify the DNSNVs in the low GC-content region with the lowest error rate among the three pipelines while RTG tends to detect the DNSNVs in the high GC-content region. Considering the effect of high GC-content on experimental and computational results of SNV detection, the Ti/Tv ratio achieved by RTG was lower than that achieved by GATK, indicating

a higher error rate in RTG calling results. Therefore, when using single pipeline to identify the DNSNVs, people should pay more attention to the DNSNVs in the GATK calling results that fell into the high GC-content region, or the DNSNVs in the RTG results that fell into the low GC-content region. For VarScan, although the calling results covered a broad region of the GC-content, the Ti/Tv ratio of the DNSNVs in the low GC-content region was only 1.05, which is much lower than that achieved by GATK (Ti/Tv = 1.88). The Ti/Tv ratio of the DNSNVs in the high GC-content region (Ti/Tv = 1.32) was comparable to that achieved by RTG (Ti/Tv = 1.25). So, people still need to carefully validate the DNSNVs detected in the low GC-content region when using VarScan for calling.

For the purpose of removing the redundant DNSNVs, we proposed a filter to refine the calling results for the single pipeline by considering the read coverage at the mutation sites of the son's genome and the parents' genomes. Our results showed that a number of DNSNVs were removed from the GATK calling results when applying a non-stringent cut-off (score = 0) and the Ti/Tv ratio of the left DNSNVs increased, indicating an improvement of the calling results. For the less redundant results, e.g., the DNSNVs detected by VarScan, only a small number of DNSNVs were filtered out and the Ti/Tv ratio did not change significantly. Our findings indicated that the proposed filter might be benefit to the refinement of the DNSNVs identified by current pipelines.

It is worth noting that, considering the fact that the size of the input data is small, the filtering algorithm can be further improved by giving a confidence index score and a statistical test index, (e.g., *p*-value) for the score of a DNSNV, which can be estimated from the mapping probability of the reads and the confidence level of the read coverage. It is helpful for increasing the confidence level of the findings. Moreover, due to the limitation of samples, we only used normal samples for the comparative analysis in this study. The work could be further improved by using genetic disease samples, which would make the evaluation of the variant calling error rate of the specific pipeline more accurate. When constructing the filtering algorithm, we only used the exome sequence as an input. The proposed algorithm can be broadened to the analysis of whole genome sequence by integrating additional steps to trim the whole genome. Additionally, we only discussed the difference of *de novo* SNVs detected by different pipelines. In fact, other types of *de novo* structural variations, such as indels, also play important roles in the biological processes in the genetic diseases. While taking more types of structural variations into account will

make the comparative analysis more complex, it does contribute to a more comprehensive assessment of the performance of existing pipelines. These issues will be further addressed in our future study.

CONCLUSION

In this study, we demonstrated that different pipelines have a specific tendency to detect the DNSNVs in the genomic regions with different GC contents. GATK performed better on detecting the DNSNVs in the low GC-content region while RTG and VarScan are better suited for detecting the DNSNVs in the high GC-content region. To refine the calling results for single pipeline, the read coverage at the mutation positions of the son's genome and the parents' genomes can be considered as an effective index to identify DNSNVs with high confidence. Our findings would be useful for the community to choose the appropriate pipelines and obtain the calling results with high confidence when discovering the *de novo* mutations for the genetic diseases.

AUTHOR CONTRIBUTIONS

ZW designed the experiments. YL, LH, YZ, and YH performed the data analysis. ZW wrote the initial version of manuscript. YZ, YL, and ML prepared all the figures. ZW, CL, and XP discussed the results and revised the manuscript. All authors contributed to discussions regarding the results and the manuscript.

FUNDING

This project was supported by grants from the National Natural Science Foundation of China (Nos. 21575094 and 21573151) and NSAF (No. U1730127).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphar.2019.00358/full#supplementary-material>

FIGURE S1 | The overlap rates between the DNSNVs and the variants in the dbSNP database.

REFERENCES

- Bacchelli, C., and Williams, H. J. (2016). Opportunities and technical challenges in next-generation sequencing for diagnosis of rare pediatric diseases. *Expert Rev. Mol. Diagn.* 16, 1073–1082. doi: 10.1080/14737159.2016.1222906
- Barak, T., Kwan, K. Y., Louvi, A., Demirbilek, V., Saygi, S., Tuysuz, B., et al. (2011). Recessive LAMC3 mutations cause malformations of occipital cortical development. *Nat. Genet.* 43, 590–594. doi: 10.1038/ng.836
- Blake, D. J., Nawrotzki, R., Loh, N. Y., Gorecki, D. C., and Davies, K. E. (1998). beta-dystrobrevin, a member of the dystrophin-related protein family. *Proc. Natl. Acad. Sci. U.S.A.* 95, 241–246. doi: 10.1073/pnas.95.1.241
- Boycott, K. M., Vanstone, M. R., Bulman, D. E., and MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* 14, 681–691. doi: 10.1038/nrg3555
- Campeau, P. M., Lenk, G. M., Lu, J. T., Bae, Y., Burrage, L., Turnpenny, P., et al. (2013). Yunis-Varón syndrome is caused by mutations in FIG4, encoding a phosphoinositide phosphatase. *Am. J. Hum. Genet.* 92, 781–791. doi: 10.1016/j.ajhg.2013.03.020
- Choi, Y., Chan, A. P., Kirkness, E., Telenti, A., and Schork, N. J. (2018). Comparison of phasing strategies for whole human genomes. *PLoS Genet.* 14:e1007308. doi: 10.1371/journal.pgen.1007308
- Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., et al. (2014). Joint variant and *de novo* mutation identification on pedigrees from

- high-throughput sequencing data. *J. Comput. Biol.* 21, 405–419. doi: 10.1089/cmb.2014.0029
- Conrad, D. F., Keebler, J. E., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., et al. (2012). Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* 43, 712–714. doi: 10.1038/ng.862
- Francioli, L. C., Cretu-Stancu, M., Garimella, K. V., Fromer, M., Kloosterman, W. P., Samocha, K. E., et al. (2017). A framework for the detection of *de novo* mutations in family-based sequencing data. *Eur. J. Hum. Genet.* 25, 227–233. doi: 10.1038/ejhg.2016.147
- Gomez-Romero, L., Palacios-Flores, K., Reyes, J., Garcia, D., Boege, M., Davila, G., et al. (2018). Precise detection of *de novo* single nucleotide variants in human genomes. *Proc. Natl. Acad. Sci. U.S.A.* 115, 5516–5521. doi: 10.1073/pnas.1802244115
- Hamdan, F. F., Srour, M., Capo-Chichi, J. M., Daoud, H., Nassif, C., Patry, L., et al. (2014). *De novo* mutations in moderate or severe intellectual disability. *PLoS Genet.* 10:e1004772. doi: 10.1371/journal.pgen.1004772
- He, M., Person, T. N., Hebring, S. J., Heinzen, E., Ye, Z., Schrod, S. J., et al. (2015). SeqHBase: a big data toolset for family based sequencing data analysis. *J. Med. Genet.* 52, 282–288. doi: 10.1136/jmedgenet-2014-102907
- Hoischen, A., van Bon, B. W., Gilissen, C., Arts, P., van Lier, B., Steehouwer, M., et al. (2010a). *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* 42, 483–485. doi: 10.1038/ng.581
- Hoischen, A., van Bon, B. W., Rodriguez-Santiago, B., Gilissen, C., Vissers, L. E., de Vries, P., et al. (2010b). *De novo* nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat. Genet.* 43, 729–731. doi: 10.1038/ng.868
- Hyrenius-Wittsten, A., Pilheden, M., Sturesson, H., Hansson, J., Walsh, M. P., Song, G., et al. (2018). *De novo* activating mutations drive clonal evolution and enhance clonal fitness in KMT2A-rearranged leukemia. *Nat. Commun.* 9:1770. doi: 10.1038/s41467-018-04180-1
- Jamuar, S. S., and Tan, E. C. (2015). Clinical application of next-generation sequencing for Mendelian diseases. *Hum. Genomics* 9:10. doi: 10.1186/s40246-015-0031-5
- Jin, Z. B., Wu, J., Huang, X. F., Feng, C. Y., Cai, X. B., Mao, J. Y., et al. (2017). Trio-based exome sequencing arrests *de novo* mutations in early-onset high myopia. *Proc. Natl. Acad. Sci. U.S.A.* 114, 4219–4224. doi: 10.1073/pnas.1615970114
- Koboldt, D. C., Larson, D. E., and Wilson, R. K. (2013). Using VarScan 2 for germline variant calling and somatic mutation detection. *Curr. Protoc. Bioinformatics* 44, 15.14.11–15.14.17. doi: 10.1002/0471250953.bi1504s44
- Kojima, K., Nariai, N., Mimori, T., Takahashi, M., Yamaguchi-Kabata, Y., Sato, Y., et al. (2013). A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads. *Bioinformatics* 29, 2835–2843. doi: 10.1093/bioinformatics/btt503
- Krier, J. B., Kalia, S. S., and Green, R. C. (2016). Genomic sequencing in clinical practice: applications, challenges, and opportunities. *Dialogues Clin. Neurosci.* 18, 299–312.
- Ku, C. S., Vasilou, V., and Cooper, D. N. (2012). A new era in the discovery of *de novo* mutations underlying human genetic disease. *Hum. Genomics* 6:27.
- Lee, H., Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-Rivera, F., et al. (2014). Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 312, 1880–1887. doi: 10.1001/jama.2014.14604
- Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., et al. (2012). A likelihood-based framework for variant calling and *de novo* mutation detection in families. *PLoS Genet.* 8:e1002944. doi: 10.1371/journal.pgen.1002944
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42, 790–793. doi: 10.1038/ng.646
- Peng, G., Fan, Y., Palculict, T. B., Shen, P., Ruteshouser, E. C., Chi, A.-K., et al. (2013). Rare variant detection using family-based sequencing analysis. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3985–3990. doi: 10.1073/pnas.1222158110
- Peters, B. A., Kermani, B. G., Alferov, O., Agarwal, M. R., McElwain, M. A., Gulbahce, N., et al. (2015). Detection and phasing of single base *de novo* mutations in biopsies from human *in vitro* fertilized embryos by advanced whole-genome sequencing. *Genome Res.* 25, 426–434. doi: 10.1101/gr.181255.114
- Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurler, M. E., Cartwright, R. A., et al. (2013). DeNovoGear: *de novo* indel and point mutation discovery and phasing. *Nat. Methods* 10, 985–987. doi: 10.1038/nmeth.2611
- Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., et al. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat. Biotechnol.* 30, 61–68. doi: 10.1038/nbt.2053
- Salzberg, S. L., Perlea, M., Fahrner, J. A., and Sobreira, N. (2014). DIAMUND: direct comparison of genomes to detect mutations. *Hum. Mutat.* 35, 283–288. doi: 10.1002/humu.22503
- Santoni, F. A., Makrythanasis, P., Nikolaev, S., Guipponi, M., Robyr, D., Bottani, A., et al. (2014). Simultaneous identification and prioritization of variants in familial, *de novo*, and somatic genetic disorders with VariantMaster. *Genome Res.* 24, 349–355. doi: 10.1101/gr.163832.113
- Shin, S. C., Ahn, D. H., Kim, S. J., Lee, H., Oh, T. J., Lee, J. E., et al. (2013). Advantages of single-molecule real-time sequencing in high-GC content genomes. *PLoS One* 8:e68824. doi: 10.1371/journal.pone.0068824
- Thiffault, I., and Lantos, J. (2016). The challenge of analyzing the results of next-generation sequencing in children. *Pediatrics* 137(Suppl. 1), S3–S7. doi: 10.1542/peds.2015-3731C
- Turner, T. N., Coe, B. P., Dickel, D. E., Hoekzema, K., Nelson, B. J., Zody, M. C., et al. (2017). Genomic patterns of *de novo* mutation in simplex autism. *Cell* 171, 710–722.e12. doi: 10.1016/j.cell.2017.08.047
- Veltman, J. A., and Brunner, H. G. (2012). *De novo* mutations in human genetic disease. *Nat. Rev. Genet.* 13, 565–575. doi: 10.1038/nrg3241
- Wei, Q., Zhan, X., Zhong, X., Liu, Y., Han, Y., Chen, W., et al. (2015). A Bayesian framework for *de novo* mutation calling in parents-offspring trios. *Bioinformatics* 31, 1375–1381. doi: 10.1093/bioinformatics/btu839
- Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* 369, 1502–1511. doi: 10.1056/NEJMoa1306555
- Zhang, L., Zhang, J., Yang, J., Ying, D., Lau, Y. L., and Yang, W. (2013). PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. *Bioinformatics* 29, 124–125. doi: 10.1093/bioinformatics/bts627
- Zhou, X., Batzoglou, S., Sidow, A., and Zhang, L. (2018). HAPDeNovo: a haplotype-based approach for filtering and phasing *de novo* mutations in linked read sequencing data. *BMC Genomics* 19:467. doi: 10.1186/s12864-018-4867-7
- 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. *Sci. Data* 3:160025. doi: 10.1038/sdata.2016.25
- Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., et al. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* 32, 246–251. doi: 10.1038/nbt.2835

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Liang, He, Zhao, Hao, Zhou, Li, Li, Pu and Wen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.