Bioinformatics, 36(11), 2020, 3549–3551
doi: 10.1093/bioinformatics/btaa116
Advance Access Publication Date: 21 April 2020
Applications Note

Genome analysis

dv-trio: a family-based variant calling pipeline using DeepVariant

Eddie K. K. Ip[1,2], Clinton Hadinata[1], Joshua W. K. Ho[1,2,3] and Eleni Giannoulatou [1,2,*]

[1]Victor Chang Cardiac Research Institute, Sydney, Australia, [2]St. Vincent's Clinical School, UNSW Sydney, Sydney, Australia and [3]School of Biomedical Sciences, The University of Hong Kong, Hong Kong, China

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 18, 2019; revised on January 31, 2020; editorial decision on February 15, 2020; accepted on April 17, 2020

Abstract

Motivation: In 2018, Google published an innovative variant caller, DeepVariant, which converts pileups of sequence reads into images and uses a deep neural network to identify single-nucleotide variants and small insertion/deletions from next-generation sequencing data. This approach outperforms existing state-of-the-art tools. However, DeepVariant was designed to call variants within a single sample. In disease sequencing studies, the ability to examine a family trio (father-mother-affected child) provides greater power for disease mutation discovery.

Results: To further improve DeepVariant's variant calling accuracy in family-based sequencing studies, we have developed a family-based variant calling pipeline, dv-trio, which incorporates the trio information from the Mendelian genetic model into variant calling based on DeepVariant.

Availability and implementation: dv-trio is available via an open source BSD3 license at GitHub (https://github.com/VCCRI/dv-trio/).

Contact: e.giannoulatou@victorchang.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DeepVariant is a novel variant caller which converts pileups of sequence reads into images and uses a deep neural network to identify single-nucleotide variants (SNVs) and small insertions/deletions (INDELs) from next-generation sequencing data (Poplin *et al.*, 2018). This approach outperforms existing state-of-the-art tools. However, DeepVariant was designed to call variants within a single sample. In disease sequencing studies, the ability to examine a family trio, (father-mother-affected child), provides greater power for disease mutation discovery.

By simply merging the variant calls of individual samples produced by a variant caller can lead to high Mendelian errors—a situation where the genotype (GT) in the child cannot be explained by the parent's GTs through Mendelian inheritance. This is because simple merging assumes that variants that are not called within one sample are homozygous for the reference allele leading to erroneous calls. Joint calling of samples alleviates this problem but this is not currently fully supported by DeepVariant.

However, one major advantage of DeepVariant is that it produces well-calibrated GT likelihoods. Therefore, we reasoned that joint calling can be improved by incorporating the GT likelihood information within a family. In this paper, we present an open source pipeline, dv-trio, that enables family-based variant calling using DeepVariant with improved accuracy.

2 Materials and methods

The workflow of dv-trio consists of three main steps: individual variant calling using DeepVariant, family trio joint calling using Genome Analysis Toolkit (DePristo *et al.*, 2011) and Mendelian error correction with FamSeq (Peng *et al.*, 2014) (Supplementary Fig. S1A). It has been developed to work on a local high-performance computing environment or from a cloud-based virtual machine. However, due to the computational load required by DeepVariant, a cloud-based approach is recommended.

2.1 Individual variant calling using DeepVariant

The input for dv-trio is passed as a text file via the '-i' option. It contains the location of the three BAM files for each sample of the family. A reference is also provided via the '-r' option, along with a dbSNP VCF via the '-d' option. Each sample BAM file is then processed by DeepVariant to create a genomic Variant Call Format file (gVCF), as well as a VCF file.

2.2 Family trio co-calling using Genome Analysis Toolkit

Following the creation of gVCFs from DeepVariant, dv-trio utilizes GATK's GenotypeGVCFs functionality to joint call a family trio using the gVCFs of the three family samples.

2.3 GT update based on Mendelian inheritance

Taking the GATK joint-called trio VCF as input, dv-trio applies FamSeq to update the GTs taking into account Mendelian inheritance. This is performed using a Bayesian network implemented in FamSeq, which utilize DeepVariant's well-calibrated GT likelihood. FamSeq allows the user to determine the amount of contribution to the GT update that is based on the family pedigree, using a Likelihood Ratio Cutoff (LRC) value, where 1 (dv-trio default) denotes using family pedigree for all variants, to 0, where an individual-based method is used. dv-trio allows the user to alter the LRC value via the '-t' option.

The final output from dv-trio is a family VCF with the samples' GT fields representing the FamSeq updated GTs. The intermediate outputs from each step, are retained for the user. All output files can be saved to a user-specified cloud-storage location using the '-b' argument of dv-trio.

3 Results

To compare the performance of dv-trio with other trio calling approaches, we used Genome in a Bottle Consortium's (GIAB) Ashkenazim Trio genome dataset (HG002—proband, HG003—father, HG004—mother) (Zook *et al.*, 2016). Three approaches were compared against dv-trio: (i) we used DeepVariant to call and generate a VCF for each sample in the Ashkenazim Trio; then the three samples VCFs were merged to create a family trio VCF using bcftools (Supplementary Fig. S1B); (ii) we used DeepVariant to call and generate a gVCF for each sample in the Ashkenazim Trio; then the three samples gVCFs were merged to create a family trio VCF using GATK4 (Supplementary Fig. S1C); and (iii) we called the trio using GATK4 best practices (Supplementary Fig. S1D). The computational performance of all these approaches is available in Supplementary Table S1.

To benchmark these approaches, a Mendelian error rate was calculated using the ancestry and kinship toolkit Mendel tool (Arthur *et al.*, 2017) for each trio VCF. The performance of the variant calling for each sample was also assessed using the Illumina hap.py tool (Krusche *et al.*, 2019), against the GIAB gold standard truth dataset within GIAB high-confidence regions.

We demonstrated that the Mendelian error rate was reduced by 60% for a dv-trio called trio VCF over a trio VCF created by merging individual DeepVariant VCFs, 19% over a GATK co-called trio VCF and 34% against a GATK Best Practices workflow co-called trio VCF (Fig. 1, Supplementary Table S2). The advantage of this assessment is that it can evaluate the performance of each approach across the whole genome rather than restricted within the GIAB high-confidence set. Interestingly the Mendelian error rate within GIAB high-confidence regions only is very small for all approaches with dv-trio outperforming all other methods (Supplementary Fig. S2).

The highest proportion of Mendelian errors occurs when parental GTs are RR-AA (Fig. 1). A proband with either RR or AA GTs would generate a Mendelian error. The total number of Mendelian compatible GTs across the trio is smaller for this category (with the exception on AA–AA), hence the error rate can be higher. In addition, most other categories of parental GTs would result in a Mendelian error if the child is either RR or AA but not both, also contributing to a smaller error rate. Finally, variant callers are likely to default to a RR GT when there is not enough evidence for a non-reference GT call.

We performed additional tests using genome and exome data from the 1000 Genomes CEPH (Centre d'Etude du Polymorphisme Humain—Utah Residents with Northern and Western Ancestry) Trio which also confirmed dv-trio's lowest Mendelian error rate

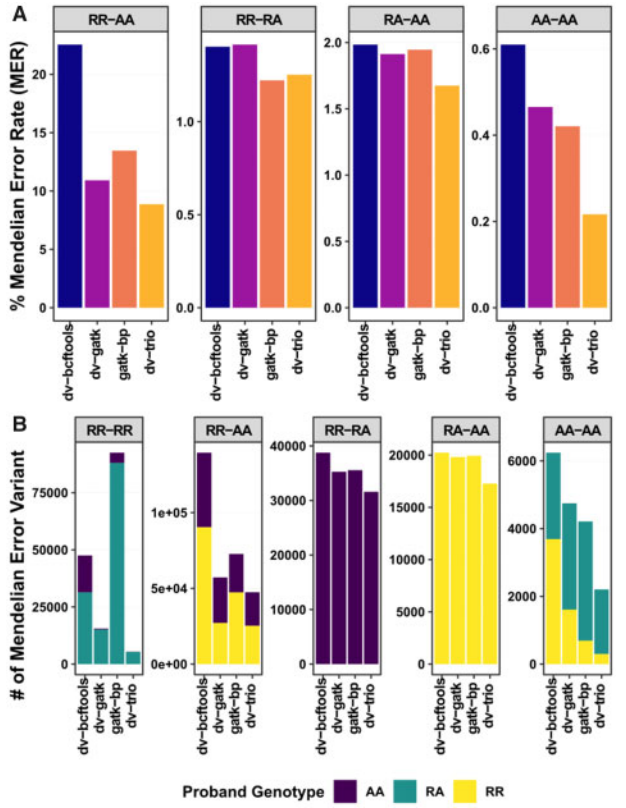

Fig. 1. Benchmark comparisons of alternative trio calling approaches using GIAB Ashkenazim Trio genome data. (**A**) Mendelian error rate. Error rates for GT calls for HG002 proband based on parental GTs (paternal/maternal). (**B**) Mendelian error breakdown by proband GT. Number of variants with Mendelian errors broken down by HG002 proband's GT. R, reference allele; A, alternative allele; dv-bcftools, DeepVariant VCFs + bcftools; dv-gatk, DeepVariant gVCFs + GATK4; gatk-bp, GATK4 best practices

(Supplementary Figs. S3 and S4) (Eberle *et al.*, 2017). We have also evaluated the performance of all approaches for SNVs and INDELs separately (Supplementary Figs. S5–S10). We found that dv-trio performs best for SNVs, whereas for INDELS its performance is equivalent to dv-gatk (creation of DeepVariant gVCFs and merging using GATK4). This is because FamSeq does not currently update the GTs of INDELs. Although no single approach performs consistently better than the others across all categories of parental GTs, both dv-trio and dv-gatk perform best overall. Finally, we investigated the effect of GT quality in Mendelian error rates for all trio calling approaches (Supplementary Fig. S11). By eliminating variants with low GT quality (GQ $<$ 20), the Mendelian error rate is decreased for all approaches with dv-trio exhibiting the lowest error rate.

The precision, recall and F1-measure for proband (Supplementary Fig. S12A), father (Supplementary Fig. S12B) and mother (Supplementary Fig. S12C) showed that when the samples are assessed individually within their GIAB high-confidence regions, DeepVariant still outperforms all existing approaches. However, dv-trio achieves the highest performance of all joint-calling approaches (DeepVariant+GATK4 and GATK4 Best Practices) while achieving marginally similar performance to the single-sample calling of DeepVariant.

4 Conclusions

We have developed a family-based calling pipeline that maintains DeepVariant's high performance and improves joint calling by taking into account the Mendelian genetic model. By interrogating the Mendelian error rates, we show that dv-trio outperforms all other current joint-calling approaches.

Funding

This work was supported by an Australian Postgraduate Award (UNSW) [E.K.K.I.], a National Health and Medical Research Council Career Development Fellowship [1105271 to J.W.K.H.], National Heart Foundation of Australia Future Leader Fellowships [101204 to E.G. and 100848 to J.W.K.H.] and a NSW Health Early-Mid Career Fellowship [E.G.].

Conflict of Interest: none declared.

References

Arthur,R. *et al*. (2017) AKT: ancestry and kinship toolkit. *Bioinformatics*, **33**, 142–144.

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

Eberle,M.A. *et al*. (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res*., **27**, 157–164.

Krusche,P. *et al*. and The Global Alliance for Genomics and Health Benchmarking Team. (2019) Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol*., **37**, 555–560.

Peng,G. *et al*. (2014) FamSeq: a variant calling program for family-based sequencing data using graphics processing units. *PLoS Comput. Biol*., **10**, e1003880.

Poplin,R. *et al*. (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol*., **36**, 983–987.

Zook,J.M. *et al*. (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, **3**, 160025.