*Bioinformatics*, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

*Genome Analysis*

# Comparing Complex Variants In Family Trios

Berke Ç. Toptaş[1,2], Goran Rakocevic[1,2], Péter Kómár[1,2] and Deniz Kural[1,2,*]

[1]Seven Bridges Genomics, 1 Main Street, Cambridge, MA 02142 USA

[2]SBGD Inc, 215 First Street, Cambridge MA 02142 USA

[*]To whom correspondence should be addressed

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** Several tools exists to count Mendelian violations in family trios using the naive variant comparison method where variants at same locations are compared. Naive variant comparison however, fails to assess regions where multiple variants need to be examined together. This reduces the accuracy of existing Mendelian violation checking tools.

**Results:** We introduce VBT, a trio concordance analysis tool that identifies Mendelian violations in family trios using a well known variant comparison algorithm to resolve variant representation differences. We show that VBT outperforms all previous trio comparison tools.

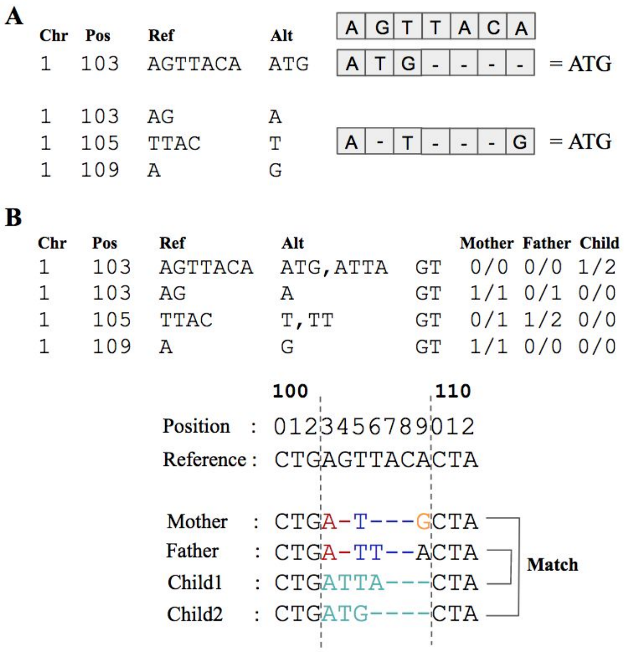
**Availability:** Pending
**Contact:** deniz.kural@sbgdinc.com
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1. Introduction

Recent technological advancements enabled a rapid progress in our understanding and characterization of the human genome, assessing the scale and extent of genomic variation present in the population(The 1000 Genome Project Consortium et al., 2012), as well as a creation of a vast body of knowledge related to the functioning of human body and its diseases(Jamuar et al., 2015).

The prevailing paradigm in the field  has been based around a reference genome, and genetic variants, which are defined as differences at particular sites in DNA sequence. One category of genetic variants of particular interest are the so-called de novo mutations, which are not inherited from parents to offspring, but arise between generations. De novo mutations occur with relatively low frequencies ($1.2x10^{-8}$)(Conrad et al., 2011; Kong et al., 2012; Francioli et al., 2014) compared to average amount of variants a person has. However, since de novo mutations do not undergo the same amount of evolutionary selection as inherited variations, they are often more deleterious. Indeed, de novo mutations have repeatedly been implicated in rare and complex diseases(Hidalgo et al., 2016).

Historically, detection of de novo mutations has been a difficult problem, until the advent of Next Generation Sequencing (NGS) allowed for their detection in either whole genome sequences, across the protein coding regions, or even in specific targeted genes (Ku et al. 2012). Following this approach, typically the DNA samples from both parents and the child are sequenced, and based on this sequencing data we look for variants that do not follow the Mendelian inheritance patterns. A major hurdle in this process arises from the limited accuracy of NGS-based assays. Problems range from amplification bias, adapter contamination, and sequencing errors to read mapping and variant calling issues. Consequently, false variants may be introduced, real variants missed, or the zygosity of a variant changed. Currently, the most accurate methods still produce thousands of errors, even considering only a subset of the human genome which is considered “easy” to process(PrecisionFDA, 2016; Zook et al., 2016). While on the genome scale, these results yield impressive precision and recall numbers (in excess of 99.9%), they indicate a significant amount of noise when one attempts to detect de novo mutations, as errors in any of the three genomes (parents’ and the child's) might cause a putative Mendelian violation, which needs to be further considered. In order to alleviate the problem, specific tools have been proposed that employ specialized statistical models (DeNovoGear (Ramu et al. 2013), PhaseByTransmission (Francioli et al., 2016)).



| Chr | Pos | Ref | Alt |
|---|---|---|---|
| 1 | 103 | AGTTACA | ATG |
| 1 | 103 | AG | A |
| 1 | 105 | TTAC | T |
| 1 | 109 | A | G |

| Chr | Pos | Ref | Alt | | Mother | Father | Child |
|---|---|---|---|---|---|---|---|
| 1 | 103 | AGTTACA | ATG,ATTA | GT | 0/0 | 0/0 | 1/2 |
| 1 | 103 | AG | A | GT | 1/1 | 0/1 | 0/0 |
| 1 | 105 | TTAC | T,TT | GT | 0/1 | 1/2 | 0/0 |
| 1 | 109 | A | G | GT | 1/1 | 0/0 | 0/0 |

**Figure 1: (a)** Representation difference in small indels. **(b)** A toy example of variant representation difference in trios. Naive trio comparison tools marks all 4 records as Mendelian violation. However a consistent combination can be found if they are processed together.

Here we address a related problem connected to the identification of de novo variants in the data from a family trio, one which arises from varying variant representations. Regions with several overlapping variants often have a number of different ways in which they can be represented, all of which conform to the widely accepted VCF standard (Danecek et al., 2011); the same is true for most variants which are complex in nature, and even some simple indels (Figure 1a). The choice of which of the possible representations is produced often depends on the variant context (other nearby variants) and the set of sequencing reads used to identify the variant. If this choice happens to be different between different members of the pedigree, comparing the three sets of calls position by position will result in detection of Mendelian violations, even though the underlying haplotypes are Mendelian compliant (Figure 1b).

Problems related to variant representation has been recognized in the context of benchmarking NGS data processing methods, and numerous approaches have been developed for comparing two sets of results for a single sample(Vcfeval (Cleary et al., 2015), VarMatch(Sun et al., 2017), Hap.py). However, none of these tools are capable of resolving the issue with data from a family trio.

In this paper, we present VBT, a Mendelian violation detection tool that uses advanced variant comparison to deal with ambiguities arising from different variant representations. VBT extends the variant comparison algorithm of vcfeval(Cleary et al., 2015) for trio concordance analysis. We show that VBT outperforms all previous trio comparison methods.

## 2. Methods

### 2.1 Variant Comparison Integration and Ideal Mendelian Equation

For a variant set V, we define the phasing vector, $P_V = \{p_1, p_2, .. p_{|V|}\} \in \{0,1\}^{|V|}$, where the i-th value (0 or 1) indicates whether the first or second allele of the i-th variant is selected for one haplotype. Similarly $P_{V'}$ denotes the opposite phasing vector $\{1\text{-}p_1, 1\text{-}p_2, .. , 1\text{-}p_{|V|}\}$, which indicates the alleles on the other haplotype, not selected by $P_V$. A haplotype function $h(V, P_V)$ is defined(Cleary et al., 2015) to produce the modified reference sequence obtained by applying all variants of V to the reference sequence using the phasing vector $P_V$. vcfeval(Cleary et al., 2015) defines the optimal solution to the variant matching problem as:

$$X^{opt}, P_X^{opt} = \underset{\substack{X \subseteq B,\ Y \subseteq C \\ P_X \in \{0,1\}^{|X|},\ P_Y \in \{0,1\}^{|Y|}}}{\arg\max}\ I[h(X, P_X), h(Y, P_Y)]\ I[h(X, P_{X'}), h(Y, P_{Y'})]\ |X| \quad (1)$$
$$Y^{opt}, P_Y^{opt}$$

where B and C denote baseline and called variant sets, and $X^{opt}$ and $Y^{opt}$ are the sets of variants which maximize the number of matches in baseline and called variant set. I[seq1, seq2] is the indicator function that is 1 if seq1 = seq2, and 0 otherwise.

For VBT, we aimed to extend the definition in eq.(1) for family trios without de-novo mutations to detect Mendelian violations. Let M, F and C represent the sets of variants of mother, father and child, respectively. We define the optimal solution to the trio matching problem as :

$$X^{opt}, Y^{opt}, Z^{opt} = \underset{\substack{X \subseteq M,\ Y \subseteq F,\ Z \subseteq C \\ P_X \in \{0,1\}^{|X|},\ P_Y \in \{0,1\}^{|Y|},\ P_Z \in \{0,1\}^{|Z|}}}{\arg\max}\ I[h(X,P_X), h(Z,P_Z)]\ I[h(Y,P_Y), h(Z,P_{Z'})]\ |Z| \quad (2)$$
$$P_X^{opt}, P_Y^{opt}, P_Z^{opt}$$

where $P_{Z'}$ denotes the opposite of $P_Z$ phasing vector. Eq. (2) maximizes the number of variants in the child that matches with both mother and father. $Z^{opt}$ denotes the set of Mendelian-consistent variants in the child, and the remaining child variants C \ $Z^{opt}$ are marked as Mendelian violations.

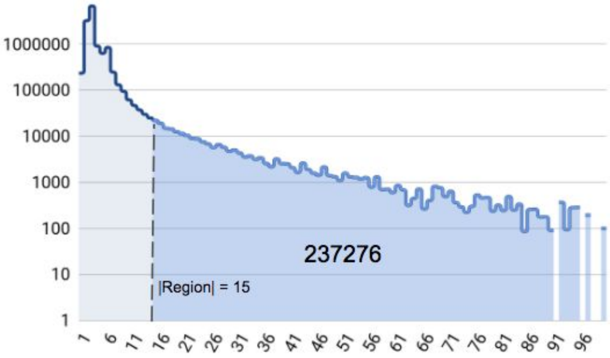


**Figure 2:** Number of total variants vs region size for Central European (CEU) Trio using HaplotypeCaller. Regions are obtained using intermediate results(sync points) of vcfeval in --squash-ploidy mode. Mother-child and father-child vcfs are compared separately and resulting two sync point sets are intersected.

Variants are partitioned into smaller groups during variant comparison with *syncpoints*(Cleary et al., 2015) where each variant subset resides between the two syncpoint and can be processed independently. For each subset, there are $3^{|subset|}$ combinations where for each family member, and each variant r there are three possibilities: excluding r, including r with $p_r = 0$, or including r with $p_r = 1$. Therefore for each region, vcfeval has exponential time and space complexity. Vcfeval uses a cutoff strategy to prevent memory-based runtime errors by skipping regions where the total variant combination exceeds the defined limits. A similar cutoff strategy (ie. |subset| > 15) using eq. (2) would cause to skip more than 200,000 variants as seen in Figure 2.

## 2.2 Search Space Reduction and Same Allele Match Elimination

We aim to decrease the number of possible combinations by separating the two indicator function in eq. (2) for mother-child and father-child variant sets, and optimize the two haplotype sequences separately. After separate processing, if a child variant exists in both haplotype sequences, it can be marked as Mendelian consistent, otherwise it becomes a Mendelian violation. When processing mother-child and father-child variants separately, we need to guarantee that the child's haplotypes use opposite phases $P_Z$ and $P_{Z'}$.

For heterozygous child variants, if one of the two alleles is not present in mother-child and father-child sequences, they should be reported as Mendelian violation. For instance, if genotype of mother is 1/1, father is 0/1 and child is 1/2 for a variant at the same position, then the child variant matches with both parents' variants with allele 1. Allele 2 on the other hand is not present in any of the parents. Although the child variant matches with both sides, it is a Mendelian violation because the same phase is used for both matches. We call this error *same allele matching*.

In a family trio, child variants often match to parent variants with both of their alleles. For these child variants, any of the two allele can be present in the final haplotype sequence, and phase is selected randomly. If, after finding the matches with both parents, the same allele of the child's variant is selected for both haplotype sequences, it would be marked as violation due to the same allele matching error even though, in this case, we could flip the phasing with one parent without breaking the match and resolve the inconsistency. To identify these variants, we apply the duo comparison function Eq. (1) to mother-child and father-child variants:

$$\begin{array}{ll} X^{opt}, P_X{}^{opt} & = \arg\max \; I[h(X, P_X), h(Z, P_Z)]\, I[h(X, P_{X'}), h(Z, P_{Z'})]\, |Z| \quad (3) \\ Z1^{opt}, P_{Z1}{}^{op} & \quad X \subseteq M,\; Z \subseteq C \\ & \quad P_X \in \{0,1\}^{|X|},\; P_Z \in \{0,1\}^{|Z|} \end{array}$$

$$\begin{array}{ll} X^{opt}, P_X{}^{opt} & = \arg\max \; I[h(Y, P_Y), h(Z, P_Z)]\, I[h(Y, P_{Y'}), h(Z, P_{Z'})]\, |Z| \quad (4) \\ Z2^{opt}, P_{Z2}{}^{opt} & \quad Y \subseteq F,\; Z \subseteq C \\ & \quad P_Y \in \{0,1\}^{|Y|},\; P_Z \in \{0,1\}^{|Z|} \end{array}$$

Where Z1 and Z2 are child variants sharing both alleles with mother and father variants respectively. From the remaining variants $M \setminus X^{opt}$ ( =: MM), $F \setminus Y^{opt}$ ( =: FF), $C \setminus Z1^{opt}$ ( =: CC1) and $C \setminus Z2^{opt}$ ( =: CC2); we obtain all child variants sharing a single allele by maximizing a single haplotype sequence, ignoring the alternate phases of variant sets:

$$\begin{array}{ll} XX^{opt}, P_{XX}{}^{opt} = & \arg\max \; I[h(XX, P_{XX}), h(ZZ1, P_{ZZ1})]\, |ZZ1| \quad (5) \\ ZZ1^{opt}, P_{ZZ1}{}^{opt} & \quad XX \subseteq MM,\; ZZ1 \subseteq CC1 \\ & \quad P_{XX}, P_{ZZ1} \end{array}$$

$$\begin{array}{ll} YY^{opt}, P_{YY}{}^{opt} = & \arg\max \; I[h(YY, P_{YY}), h(ZZ2, P_{ZZ2})]\, |ZZ2| \quad (6) \\ ZZ2^{opt}, P_{ZZ2}{}^{opt} & \quad YY \subseteq FF,\; ZZ2 \subseteq CC2 \\ & \quad P_{YY}, P_{ZZ2} \end{array}$$

where, during maximization, $P_{XX}$, $P_{YY}$, $P_{ZZ1}$ and $P_{ZZ2}$ are required to be such that the reference allele ("0") is never used in any comparison. I.e. if a variant has the genotype 2|0, the corresponding phasing is not allowed to take the value 1, because that would correspond to the "0" allele.

The reason for not allowing the reference allele to be used is the ambiguity caused by identical representation of excluded variant and included reference allele. Child variants having reference phasing were always included with the old indicator function regardless of the corresponding parent variant. For example, if genotype of mother is 0/1, father is 2/2 and child is 0/1 for a variant at the same position, mother and child variants would be included because they share both alleles. However in father-child side, father variant would be excluded (ie. the position now becomes reference) and child would be included again with 0 allele. At the end, child variant would be present on both sides and would be marked as Mendelian consistent, while it is a violation in reality. With the above restriction on the phasing vectors, we eliminate this mistake.

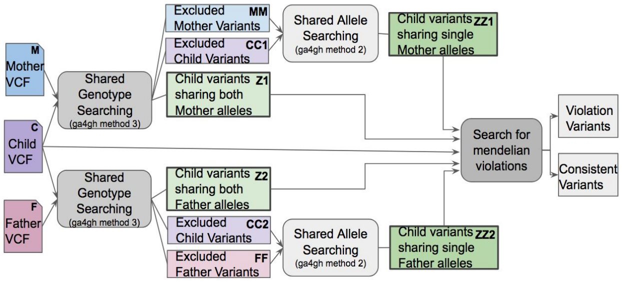


**Figure 3:** VBT pipeline using vcfeval best path algorithm and GA4GH benchmarking standard methods(ga4gh, 2015). Included variants are present in the best common path between parent and child while excluded variants are eliminated from that path.

Once we obtain our four child variant set $Z1^{opt}$, $Z2^{opt}$, $ZZ1^{opt}$ and $ZZ2^{opt}$ with their phasing information $P_{Z1}{}^{opt}$, $P_{Z2}{}^{opt}$, $P_{ZZ1}{}^{opt}$, $P_{ZZ2}{}^{opt}$(Figure 3), we check how many of them exist in both mother-child and father-child side to determine Mendelian violations with the following method:

**Algorithm 1:** Same Allele Match Elimination

```
procedure GETVIOLATIONS
Input: $Z1^{opt}, P_{Z1}{}^{opt}, Z2^{opt}, P_{Z2}{}^{opt}, ZZ1^{opt}, P_{ZZ1}{}^{opt}, ZZ2^{opt}, P_{ZZ2}{}^{opt}$
Output: ConsistentChildList, ViolationChildList
CVars_MC = $Z1^{opt} \cup ZZ1^{opt}$, CVars_FC = $Z2^{opt} \cup ZZ2^{opt}$
CPhases_MC = $P_{Z1}{}^{opt} \cup P_{ZZ1}{}^{opt}$, CPhases_FC = $P_{Z2}{}^{opt} \cup P_{ZZ2}{}^{opt}$
SortByIndex(CVars_MC, CPhases_MC)
SortByIndex(CVars_FC, CPhases_FC)
WHILE varM in CVars_MC AND phaseM in CPhases_MC AND
 varF in CVars_FC AND phaseF in CPhases_FC
    IF varM.Index = varF.Index
        IF IsHomozygous(varM)
            ADD varM to ConsistentChildList
        ELSE IF phaseM $\neq$ phaseF
            ADD varM to ConsistentChildList
        ELSE IF varM $\in$ Z1 OR varF $\in$ Z2
```

```
12              ADD varM to ConsistentChildList
13          ELSE
14              ADD varM to ViolationChildList
15          ENDIF
16          next(varM), next(varF), next(phaseM), next(phaseF)
17      ELSE IF varM.Index < varF.Index
18          ADD varM to ViolationChildList
19          next(varM), next(phaseM)
20      ELSE
21          ADD varF to ViolationChildList
22          next(varF), next(phaseF)
23      ENDIF
24  ENDWHILE
```

For intersection operation, we first merge the shared genotype and shared allele child variant sets keeping the information of belonging sets for each variant at lines (1) and (2) of the pseudocode. Then we sort the merged child variant sets by variant indexes (order in vcf) at lines (3) and (4). At line (7), we check the condition where child variant is homozygous and same allele matching condition is ignored. At line (9) we check whether heterozygous child variants match to parents with different phasings. At line (11) we check if child variant matches to parent with both alleles so that alternative phasing can also be used to avoid same allele match condition. We use *next* command to iterate to the following variant/phase at line (16). At the end, we obtain the list of Mendelian violation and consistent variant list for given set Z1, Z2, ZZ1 and ZZ2.

We process child variants with reference allele as follows:

$$K1 = \{\, r \in Z2^{opt} \cup ZZ2^{opt},\ A_r(0) = a_{REF} \lor A_r(1) = a_{REF} : h(X^{opt} \cup XX^{opt}, P_X^{opt} \cup P_{XX}^{opt})[s_r..e_r] = Ref[s_r..e_r] \,\} \qquad (7)$$

$$K2 = \{\, r \in Z1^{opt} \cup ZZ1^{opt},\ A_r(0) = a_{REF} \lor A_r(1) = a_{REF} : h(Y^{opt} \cup YY^{opt}, P_Y^{opt} \cup P_{YY}^{opt})[s_r..e_r] = Ref[s_r..e_r] \,\} \qquad (8)$$

Where $s_r$ and $e_r$ denote the start and end position of variant r and *Ref* denotes the reference sequence string. $A_r(k)$ is the allele function that represents the allele of variant r with the phase selection of $k \in \{0,1\}$ and $a_{REF}$ is the reference allele of variant. K1 and K2 are the set of consistent child variants that contain a reference allele and they are inserted to the set of consistent variants. Remaining unprocessed child variants (ie. $C \setminus (Z1^{opt} \cup Z2^{opt} \cup ZZ1^{opt} \cup ZZ2^{opt})$) are inserted to the set of violation variants.

Once we obtain decisions of all child variants, we merge mother, father and child vcf as a trio by merging variants at the same position. Then, we apply three post-processing steps on the merged vcf:

(1) Assign Mendelian decision to sites where child has no variant (ie. homozygous ref child variants in merged trio). For each hom-ref child variant, final haplotype sequences of both mother and father are checked. If any of parent haplotype sequences is non-reference at the child variant location, then the variant is marked as violation.

(2) Consolidate the decision for variants affecting the same position in the final haplotype sequence. Consistent variant decisions are changed to violation if there is at least one overlapping violation variant.

(3) Exclude sites where nocall is reported by at least one family member.

## 3. Results

A truth set for trio analysis does not exist for direct result comparison. Instead, we use alternative testing methods to compare VBT and existing tools. In the first scenario, we construct a trio from a single individual by changing its variant representations. Since all three samples belong to the same individual, we expect to see 0 Mendelian violation.

**Table 1.** Violation counts of different tools where the input trio is constructed from a single sample

| Input Sample | VBT | LBL | PBT* |
|---|---|---|---|
| NA12878 | 0 | 62785 | 61977 |
| NA12891 | 0 | 59641 | 60469 |
| NA12892 | 0 | 61134 | 59712 |

Line-By-Line check tools(LBL) includes RTG -mendel, GATK -SelectVariants and Vcftools -mendel. *PhaseByTransmission(PBT), 2x10$^{-2}$ and 1x10$^{-8}$ are used as mutation rate. PBT outputs the number of violation and corrected genotypes separately and we use the sum of these two. Identical results are obtained for both mutation rates.

We use Central European(CEU) individuals as test data by using BWA-MEM(Li, 2013) + FreeBayes(Garrison et al., 2012) to generate vcf files then use Vt norm(Tan et al., 2015) to alter variant representations and generate the “child’s” samples. We run these three test datasets with VBT, naive Mendelian error detection tools and PhaseByTransmission (PBT). For PBT, we used both 2x10$^{-2}$ and 10$^{-8}$ as mutation rates, and obtained the same number of *corrections* plus *mutations*. As seen in Table 1, VBT correctly outputs no violation for all three test data while the other tools output around sixty thousand violations.

In the second experiment, we used CEU trio (NA12878, NA12891 and NA12892) aligned with BWA-MEM to compare trio concordance rate of different variant callers, FreeBayes(fb), GATK(McKenna et al., 2010) UnifiedGenotyper(ug) and HaplotypeCaller(hc) We used vcftools v0.1.14 to merge vcf files of individual samples.

**Table 2.** Violation validation results of different variant callers using CEU trio bam files aligned with BWA.

| | **VBT** | | | **Line-by-Line** | | |
|---|---|---|---|---|---|---|
| | Total | Wrong | Missed | Total | Wrong | Missed |
| **fb** | 282984 | 258 | 2112 | 308150 | 39999 | 16910 |
| **fb+norm** | 280665 | 22 | 1725 | 308610 | 38646 | 12633 |
| **hc** | 221190 | 442 | 247 | 207048 | 118 | 14059 |
| **ug** | 248342 | 241 | 2727 | 244821 | 10170 | 16284 |

Total violation, wrongly assessed violation and missed violation counts for autosomes are provided for VBT and naive line-by-line violation check method. No filtration is applied to the data. To generate this table, it is assumed that no violation exists which both tool failed to detect.

We generated 4 trio vcfs using different variant callers, then we ran VBT and line-by-line checking tool for each trio to compare results of tools.

To test which Mendelian decisions are correct, we implemented a nonlinear validation tool that checks all possible combinations of variant phasings separately in regions between syncpoints. (Supplementary Material, Section 2). Table 2 shows the numbers of total violations, falsely identified violations and missed violations for the two methods, and for the 4 different variant calling pipelines. If we define *recall* of Mendelian violation checking tool as Violation Count / (Violation Count + Missed Violation) and *precision* as (Violation Count - Wrong Violation) / Violation Count, It can be seen from Table 2 that, VBT's recall is higher than the recall of the line-by-line checking method. The precision of line-by-line tools and VBT is closer for HaplotypeCaller because the representations of called variants are more similar  across the samples compared to other variant callers.

## 4. Discussion

In this work, we presented VBT, a mendelian violation detection tool that is capable of comparing complex indels in family trios. We showed with our test scenarios that, VBT has better accuracy than the existing tools.

VBT does not perform any statistical analysis based on genotype likelihood values to identify de novo mutations from set of Mendelian violations. However, due to the comparison algorithm we use, VBT does not alter any variant representation in the given input which makes VBT integratable with other tools doing such analyses.

VBT can also be used for different purposes rather than rare disease studies such as truth-free benchmarking of sequencers/variant callers. (Douglas et al. 2002; Pilipenko et al., 2014; Nutsua et al., 2015). Due to the very small mutation rate in human genome, all Mendelian violations can be considered as sequencing/variant calling errors. Using VBT as trio concordance analysis is useful where there is no truth-set exist and provide more data coverage as opposed to current whole genome truth-sets which are limited to a set of high confidence regions in a few samples, excluding many important regions of the genome (Zook et al., 2014).

VBT resolves variant representation differences in family trios efficiently by maximizing matching child variants with mother and father separately instead of using the ideal trio comparison function(eq 2). This enables covering nearly all regions in datasets and provide VBT a reasonable running time, which varies between 5 and 10 minutes on Amazon c4.4xlarge instance for whole human genome trio concordance analysis. On the other hand, obtained local best paths from mother-child and father-child duos are not always identical to the global optimum. This introduces a small error rate visible on Table 2. Ideal Mendelian function (eq. 2) could be implemented by skipping around 1% of all variants. However, we trade off the accuracy aiming to support whole genome data with the least number of skipped variants.

In order to improve VBT's accuracy, it's wrong decisions can be corrected by post-processing violation regions. In regions where line-by-line check and VBT disagrees, a nonlinear violation check can be performed by generating all possible subsequences for that region, similarly to our violation validation pipeline.

## 5. Acknowledgements

We would like to thank Maxime Huvet, Maria C. Suciu and Amit Jain for their valuable contributions to this work.

## 6. Funding

This work has been supported by the …..

*Conflict of Interest:* none declared.

## 7. References

1000 Genomes Project Consortium et al. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56–65.

Cleary, J.G. et al. (2015) Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv*. doi:10.1101/023754/

Conrad, D.F. et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, 43, 712-714

Danecek, P. et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics*, 27(15), 2156–2158.

Douglas, J.A. et al. (2002). Probability of Detection of Genotyping Errors and Mutations as Inheritance Inconsistencies in Nuclear-Family Data. *Am. J. Hum. Genet*. 70(2) 487-495.

Francioli, L.C. et al. (2017). A framework for the detection of de novo mutations in family-based sequencing data. *Eur. J. Hum. Genet.* 25, 225-233.

Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing. *arXiv*. doi: arXiv:1207.3907

Genome of the Netherlands Consortium. (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet*. 46, 818-825.

Hidalgo, R. A. et al. (2016). New insights into the generation and role of de novo mutation in health and disease. *Genome Biol.* 17, 241.

Jamuar, S.S. and Tan, E.C. (2015) Clinical application of next-generation sequencing for Mendelian diseases. *Hum. Genomics*. doi: 10.1186/s40246-015-0031-5

Kong, A. et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature.* 488, 471-475.

Krøigård, A.B. et al. (2016). Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *Plos One*. doi: 10.1371/journal.pone.0151664

Ku, C. -S. et al. (2012). A new era in the discovery of de novo mutations underlying human genetic disease. *Hum Genomics*. 6, 27.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. doi:  arXiv:1303.3997v2

Liu, X. et al. (2013) Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *Plos One*. doi: 10.1371/journal.pone.0075619

McKenna, A. et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20, 1297-1303.

Pilipenko, V. V. et al. (2014). Using Mendelian inheritance errors as quality control criteria in whole genome sequencing dataset.

Precision FDA Truth Challenge. (2016). https://precision.fda.gov/challenges/truth/results. Accessed September 2017.

Ramu, A. et al. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat Methods*. 10, 985-987.

Shang, J. et al. (2014). Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. *BioMed Res Int.* 2014, 309650.

Sun, C and Medvedev P. (2016). VarMatch: robust matching of small variant datasets using flexible scoring schemes. *Bioinformatics*. 33(9), 1301-1308.

Tan, A. et al. (2015). Unified Representation of Genetic Variants. *Bioinformatics*. 31(13), 2202-2204.

Wei, Q. et al. (2015). A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics*. 31(9), 1375-1381.

Zook, J.M. et al. (2014). Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls. *Nat Biotechnol.* 32, 246-251.

Zook, J.M. et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data, 3.* doi: 10.1038/sdata.2016.25