Comparing Complex Variants In Family Trios

Authors: Berke Cagkan Toptas, Goran Rakocevic, Péter Kómár, and Deniz Kural

Abstract

Background: Many bioinformatic pipelines have recently been developed to process Next Generation Sequencing (NGS) data. To compare accuracy and performance of different pipelines, truth-sets can be used. However, current truth sets are limited to high confidence regions and most of the pipelines have already greater than 99% accuracy scores. As an alternative to truth-set-based analysis, Mendelian Inheritance rules can be used to assess variant calling accuracy using family trios. Several tools exists to count Mendelian violations in a family trio using a 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 counts Mendelian violations in family trios using a well known variant comparison algorithm to resolve variant representation differences. To test VBT, we construct a family trio using a single individual by altering variant representations. For that generated trio, VBT outputs 0 Mendelian violations where other tools output incorrect ~60,000 Mendelian violations. In addition, we implement a validation pipeline to compare accuracy of VBT and other tools for outputs of different variant callers. VBT correctly identify ~99% of Mendelian violations in whole genome of Central European (CEU) trio while tools using naive variant comparison could identify less than 93%.

Conclusion: The advanced variant comparison method used by VBT can resolve different variant representations and accurately count Mendelian violations in a trio. VBT outperforms all previous Mendelian violation checking tools despite the error rate.

Keywords: Mendelian Inheritance, Violation, Truth-free, Benchmarking, trio, Variant Comparison, Variant Representation

Background

Next Generation Sequencing (NGS) has 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[1], as well as a creation of a vast body of knowledge related to the functioning of human body and its diseases. First steps are also being undertaken towards a direct clinical application of this technology[2][3].

A key step in an NGS experiment is the so-called *secondary analysis*, a data processing step during which data produced by the sequencers are translated into a set of *variant calls*, sites where the sequenced sample differs from the reference genome. For this purpose, a variety of bioinformatics pipelines consisting of aligners[4] and variant callers[5][6] have been developed and results are represented in VCF file format[7]. To assess the accuracy of different variant calling pipelines, several whole genome truth-sets, such as the ones developed by the Genome in a Bottle Consortium (GIAB)[8], can be used.

Analyses based on the GIAB truth-sets are limited to a set of high confidence regions in a few samples, excluding many important regions of the genome[9]. Furthermore, most current variant calling methods produce results with very high degree of overlap with the consensus genotype set[10] and differ at very few loci in the high-confidence regions[11], which renders the comparison between different methods very difficult.

As an alternative to truth-set-based analysis, Mendelian inheritance rules[12] can be used to assess variant calling pipeline accuracy. Due to the very small mutation rate ($\sim 10^{-8}$ per locus)[13][14][15] in human genome, all Mendelian violations can be considered sequencing/variant calling errors.

Several tools exist (RTG mendel, GATK SelectVariants, Vcftools mendel, PhaseByTransmission[18]) that count Mendelian violations using the naive line-by-line variant comparison. In this approach, each record in the merged trio vcf is processed independently, only variants with coinciding reference positions are analyzed together. This method fails to provide an accurate analysis in cases where consecutive records describe overlapping variants. The specification of VCF files allows representing variants in several different ways[16][17]. Equivalent variants could be represented by different vcf records, which can result in wrong Mendelian assessments (Supplementary Material, Section 1).

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[16] for trio concordance analysis. We show that VBT outperforms all previous trio comparison methods.

Methods

Variant Comparison Integration and Ideal Mendelian Function

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[16] to produce the modified reference sequence obtained by applying all variants of V to the reference sequence using the phasing vector P_V. vcfeval[16] defines the optimal solution to the variant matching problem as:

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

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 :

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

where $P_{Z'}$ denotes the opposite of phasing vector P_Z. 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 $C \setminus Z^{opt}$ are marked as Mendelian violations.

Figure 1: *Number of Regions vs number of variants in each region for Central European (CEU) Trio using UnifiedGenotyper is obtained using intermediate results(sync points) of vcfeval in --squash-ploidy mode. Mother-child and father-child vcfs are compared separately and resulting two sets are intersected.*

Variants are partitioned into smaller groups during variant comparison with *syncpoints*[16] 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 from comparison as seen in Figure 1.

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 maximize the two haplotype sequences separately. After separate processing, if a child variant exists in both haplotype sequences, it can be marked 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 different phases P_Z and $P_{Z'}$.

Search Space Reduction and Same Allele Match Elimination

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 at father and mother variants. Although the child variant matches to both side, it is a Mendelian violation. We name this error *same allele matching*.

In a family trio, many child variants match to parent variants with both of their alleles. For those variants, any of the two allele can be present in the final haplotype sequence. If the same allele of child variant is selected for both haplotype sequence, it would be marked as violation due to the same allele matching error although matching allele could be replaced. To identify those child variants that falls in same allele matching error, we apply the original comparison function Eq. (1) to mother-child and father-child variants:

$$X^{opt}, P_X{}^{opt}, Z1^{opt}, P_{Z1}{}^{opt} = \arg\max\ I[h(X, P_X), h(Z, P_Z)]\ \ I[h(X, P_{X'}), h(Z, P_{Z'})]\ |\ Z\ | \qquad (3)$$
$$X \subseteq M,\ Z \subseteq C$$
$$P_X \in \{0,1\}^{|X|},\ P_Z \in \{0,1\}^{|Z|}$$

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

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

$$XX^{opt}, P_{XX}{}^{opt}, ZZ1^{opt}, P_{ZZ1}{}^{opt} = \arg\max\ I[h(XX,P_{XX}),h(ZZ1, P_{ZZ1})]\ |\ ZZ1\ | \qquad (5)$$
$$XX \subseteq MM,\ ZZ1 \subseteq CC1$$
$$P_{XX}, P_{ZZ1}$$

$$XX^{opt}, P_{XX}{}^{opt}, ZZ2^{opt}, P_{ZZ2}{}^{opt} = \arg\max\ I[h(YY,P_{YY}),h(ZZ2 ,P_{ZZ2})]\ |\ ZZ2\ | \qquad (6)$$
$$YY \subseteq FF,\ ZZ2 \subseteq CC2$$
$$P_{YY}, P_{ZZ2}$$

where 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.

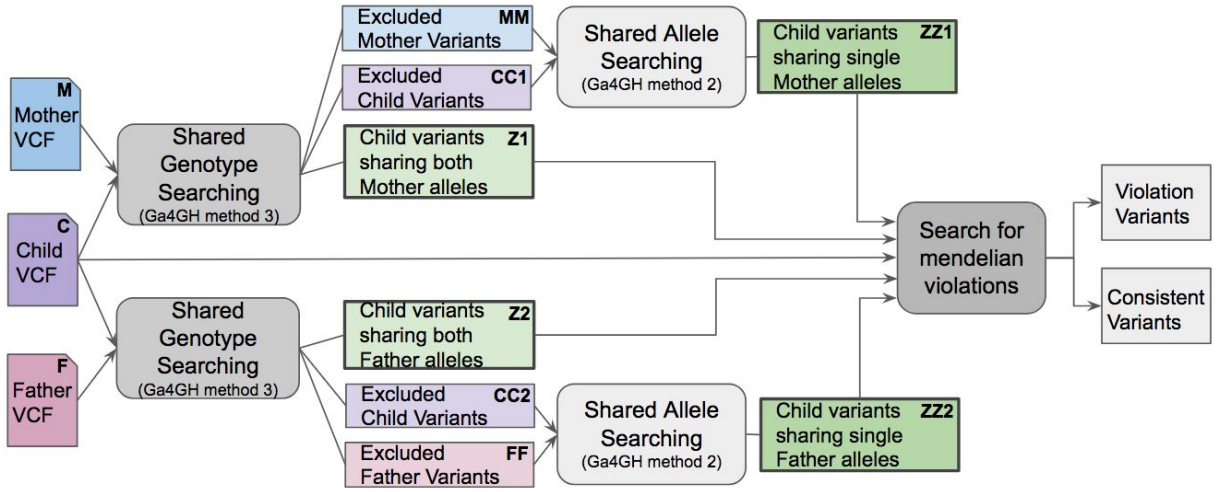

Figure 2: *VBT pipeline using vcfeval best path algorithm and GA4GH benchmarking standard methods[20] . Included variants are present in the best common path between parent and child while excluded variants are eliminated from that path.*

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.

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 2), 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 (Z1^opt, P_Z1^opt, Z2^opt, P_Z2^opt, ZZ1^opt, P_ZZ1^opt, ZZ2^opt, P_ZZ2^opt)
1       OUTPUT: ConsistentChildList, ViolationChildList
2       itrMC, itrFC = 0                                        > Child Variant Iterators
3       CVars_MC = Z1^opt ∪ ZZ1^opt, CVars_FC = Z2^opt ∪ ZZ2^opt
4       CPhases_MC = P_Z1^opt ∪ P_ZZ1^opt, CPhases_FC = P_Z2^opt ∪ P_ZZ2^opt
5       SortByIndex(CVars_MC, CPhases_MC)
6       SortByIndex(CVars_FC, CPhases_FC)
7       WHILE itrMC < SIZE(CVars_MC) AND itrFC < SIZE(CVars_FC)
8           IF CVars_MC[itrMC].Id = CVars_FC[itrFC].Id THEN
9               IF IsHomozygous(CVars_MC[itrMC]) THEN
10                  ADD CVars_MC[itrMC] to ConsistentChildList
11              ELSE IF CPhases_MC[itrMC] ≠ CPhases_FC[itrFC] THEN
12                  ADD CVars_MC[itrMC] to ConsistentChildList
13              ELSE IF CVars_MC[itrMC] ∈ Z1 OR CVars_FC[itrFC] ∈ Z2 THEN
15                  ADD CVars_MC[itrMC] to ConsistentChildList
16              ELSE
17                  ADD CVars_MC[itrMC] to ViolationChildList
18              ENDIF
19              INCREMENT itrMC, itrFC
20          ELSE IF CVars_MC[itrMC].Id < CVars_FC[itrFC].Id THEN
21              ADD CVars_MC[itrMC] to ViolationChildList
22                  INCREMENT itrMC
23          ELSE
24              ADD CVars_FC[itrFC] to ViolationChildList
25                  INCREMENT itrFC
26          ENDIF
27      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 (3) and (4) of the pseudocode. Then we sort the merged child

variant sets by variant indexes (order in vcf) at lines (5) and (6). At line (9), we check the condition where child variant is homozygous and same allele matching condition is ignored. At line (11) we check whether heterozygous child variants match to parents with different phasings. At line (13) 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. 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:

$$K1 = \{ r \in Z2^{opt} \cup ZZ2^{opt} : h(X^{opt} \cup XX^{opt}, P_X{}^{opt} \cup P_{XX}{}^{opt})[s_r..e_r] = Ref[s_r..e_r] \} \quad (7)$$
$$A_r(0) = a_{REF} \vee A_r(1) = a_{REF}$$
$$K2 = \{ r \in Z1^{opt} \cup ZZ1^{opt} : h(Y^{opt} \cup YY^{opt}, P_Y{}^{opt} \cup P_{YY}{}^{opt})[s_r..e_r] = Ref[s_r..e_r] \} \quad (8)$$
$$A_r(0) = a_{REF} \vee A_r(1) = a_{REF}$$

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 the three individual samples as a trio. Then, we apply three post-processing stages 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 sequence is checked for both side if the location of variant is equal to the reference sequence.
(2) Unify the decision of variants affecting the same position at the final haplotype sequence. Consistent variant decisions are changed to violation if there are at least one overlapping violation variant.
(3) Exclude sites where nocall is reported by at least one family member

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.

Sample	VBT	LBL	PBT*
NA12878 as Trio	0	62785	61977
NA12891 as Trio	0	59641	60469
NA12892 as Trio	0	61134	59712

Table 1: Violation counts of different tools where the input trio is constructed from a single sample.
Line-By-Line check tools(LBL) includes RTG -mendel, GATK -SelectVariants and Vcftools -mendel
*PhaseByTransmission(PBT), 2E-2 and 1E-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 mutation rates 2E-2 and 1E-8.

We use Central European(CEU) individuals as test data by using BWA-MEM[21] + FreeBayes[22] to generate vcf files then use Vt norm[23] to alter variant representations and generate the "child's" samples. We run this three test datasets with VBT, naive Mendelian error detection tools and PhaseByTransmission(PBT). For PBT, we use 2E-2 as mutation rate which is a rough estimation of Mendelian violation rate in trios. We also run PBT with default mutation rate (1E-8) for comparison. As seen in Table 1, VBT 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[21] to compare trio concordance rate of different variant callers, FreeBayes(fb)[22], GATK[24] UnifiedGenotyper(ug) and HaplotypeCaller(hc) We used vcftools v0.1.14 to merge vcf files of individual samples.

	VBT			Line-by-Line		
Variant Caller	Violation Count	Wrong Assessed Violation	Missed Violation	Violation Count	Wrong Assessed Violation	Missed Violation
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

Table 2: Validation results of different variant callers using CEU trio bam files aligned with BWA-MEM. Total violation, wrongly assessed violation and missed violation counts for autosomes are shown for VBT and naive line-by-line violation check method. No filtration is applied to the data. We assumed that no violation exists which both tools failed to detect.

We generate 4 trio vcfs using different variant callers, then we run 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 in a region selected so that it is independent of the variants in all other regions. (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)', It can be seen from Table 2 that, VBT's recall is higher than the recall of the line-by-line checking method for all pipelines. The result of line-by-line tools and VBT is closer for HaplotypeCaller because the representations of called variants are more similar across the samples.

Discussion

In this work, we presented VBT, a trio concordance analysis tool for benchmarking variant calling pipelines as an alternative to truth-set-based analysis. We resolve 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). That results in covering nearly all regions in datasets and provide VBT a reasonable running time varies between 10 and 15 minutes on Amazon c4.2xlarge instance[25] for whole 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 losing around 1% of vcf records. However, we trade off the accuracy aiming to support whole genome data without skipping any region as opposed to truth-set-based analysis.

In order to improve VBTs accuracy, some of 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 similar to the violation validation pipeline. This can be a future work for VBT.

Acknowledgements

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

Funding

Pending..

References

[1] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; doi: 10.1038/nature09534

[2] Jamuar SS, Tan EC. Clinical application of next-generation sequencing for Mendelian diseases. Human Genomics. 2015; doi: 10.1186/s40246-015-0031-5

[3] Park JY, Kricka LJ, Fortina P. Next-generation sequencing in the clinic. Nature Biotechnology. 2013; doi: 10.1038/nbt.2743

[4] Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and Comparison of Multiple Aligners for Next-Generation Sequencing Data Analysis. BioMed. 2014; doi: 10.1155/2014/309650

[5] Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. Plos One. 2013; doi: 10.1371/journal.pone.0075619

[6] Krøigård AB, Thomassen M, Lænkholm AV, Kruse TA, Larsen MJ. Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. Plos One. 2016; doi: 10.1371/journal.pone.0151664

[7] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The Variant Call Format and VCFtools. Bioinformatics. 2011; doi: 10.1093/bioinformatics/btr330

[8] Benchmarking Reference Implementation. (2015) https://github.com/ga4gh/benchmarking-tools/tree/master/doc/ref-impl/. Accessed September 2017.

[9] Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls. Nature Biotechnology. 2014; doi: 10.1038/nbt.2835

[10] Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre ABR, Chandramohan D, Chen F, Jaeger E, Moshrefi A, Pham K, Stedman W, Liang T, Saghbini M, Dzakula Z, Hastie A, Cao H, Deikus G, Schadt E, Sebra R, Bashir A, Truty RM, Chang CC, Gulbahce N, Zhao K, Ghosh S, Hyland F, Fu Y, Chaisson M, Xiao C, Trow J, Sherry ST, Zaranek AW, Ball M, Bobe J, Estep P, Church GM, Marks P, Kyriazopoulou-Panagiotopoulou S, Zheng GXY, Schnall-Levin M, Ordonez HS, Mudivarti PA, Giorda K, Sheng Y, Rypdal KB, Salit M. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data. 2016; doi: 10.1038/sdata.2016.25

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

[12] Douglas JA, Skol AD, Boehnke M. Probability of Detection of Genotyping Errors and Mutations as Inheritance Inconsistencies in Nuclear-Family Data. American Journal of Human Genetics. 2002; doi: 10.1086/338919

[13] Conrad DF, Keebler JEM, DePristo MA et al: Variation in genome-wide mutation rates within and between human families. Nature Genetics. 2011; doi: 10.1038/ng.862

[14] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WSW, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; doi: 10.1038/nature11396

[15] Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature Genetics 2014; doi : 10.1038/ng.3021

[16] Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, Zook JM, Trigg L, De La Vega FM. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. bioRxiv. 2015; doi:10.1101/023754

[17] Sun C, Medvedev P. VarMatch: robust matching of small variant datasets using flexible scoring schemes. Bioinformatics. 2016; doi: 10.1093/bioinformatics/btw797

[18] Francioli LC, Cretu-Stancu M, Garimella KV, Fromer M, Kloosterman WP, Genome of the Netherlands consortium, Samocha KE, Neale BM, Daly MJ, Banks E, DePristo MA, de Bakker PI. A framework for the detection of de novo mutations in family-based sequencing data. European Journal of Human Genetics. 2017; doi:10.1038/ejhg.2016.147

[19] Wei Q, Zhan X, Zhong X, Liu Y, Han Y, Chen W, Li B. A Bayesian framework for de novo mutation calling in parents-offspring trios. Bioinformatics. 2015; doi: 10.1093/bioinformatics/btu839

[20] Benchmarking Performance Metrics Definitions for SNVs and Small Indels. (2015). https://github.com/ga4gh/benchmarking-tools/blob/master/doc/standards/GA4GHBenchmarkingPerformanceMetricsDefinitions.md. Accessed September 2017.

[21] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013; doi: arXiv:1303.3997v2

[22] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012; doi: arXiv:1207.3907

[23] Tan A, Abecasis GR, Kang HM. Unified Representation of Genetic Variants. Bioinformatics. 2015; doi: 10.1093/bioinformatics/btv112

[24] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; doi: 10.1101/gr.107524.110

[25] Amazon EC2 Instance Types. https://aws.amazon.com/ec2/instance-types/ Accessed September 2017.