BCFtools/trio-dnm

Petr Danecek

Version: November 25, 2020

This document describes methods implemented in the bcftools/trio-dnm plugin. At the time of writing, the plugin implements two *de novo* mutation (DNM) discovery models: the original "genotype-centric" approach, which is also at the core of the Family-aware Illumina Genotype Likelihood-based method (FIGL) and is used by DeNovoGear (DNG) [Conrad *et al.*, 2011, Ramu *et al.*, 2013], and a new "allele-centric" approach which aims to overcome some shortcomings and bugs in DeNovoGear implementation, and which is in these notes referred to as "trio-dnm model".

1 Genotype-centric approach (DeNovoGear model)

The sole purpose of reimplementing the DeNovoGear model was to build trust in the new method by showing that BCFtools can successfully reproduce results from the original model. This method can be invoked by running the plugin with the `--use DNG` option.

The DNG model relies on genotype likelihoods contained in the FORMAT/PL annotation calculated by bcftools/mpileup (originally samtools/mpileup). It is beyond the scope of this document to explain this step in full detail, therefore we only show the final formula and refer the interested reader to [Li, 2010] and `htslib/errmod.c` code.

For simplicity of notation we will assume bi-allelic sites, with $x, y \in \{0, 1\}$ representing the reference (0) and the alternate (1) allele, and consider only diploid genotypes, i.e. $G \in \{xx, xy, yy\}$. Assume we observed in total n reads covering a genomic position, of which k were alternate reads and $n-k$ reference reads. Each of the bases is assigned the error probability ϵ_i derived from the base quality and mapping quality.

Genotype likelihoods express how consistent is the observed data D with all possible genotypes

$$
\begin{aligned}
P(D|00) &= \prod_{i=0}^{k-1} \beta_{ni}^{f_i}(\epsilon_i), \\
P(D|11) &= \prod_{i=k}^{n-1} \beta_{ni}^{f_i}(\epsilon_i), \\
P(D|01) &= \binom{n}{k} 1/2^n \qquad\qquad \text{(because } \epsilon \ll 1/2\text{)},
\end{aligned}
$$

where $\beta_{ni}^{f_i}$ is the conditional probability of observing more than i errors while accounting for their possible dependency using the empirical parameter $f_i = 0.97\eta^l + 0.03$.

These are the input data for the DNG model, for more details see the section "Revised MAQ model" in [Li, 2010]. In addition DeNovoGear requires FORMAT/DP annotation to filter low coverage sites (DP < 10), but since it is not used in the calculation, it is not required by bcftools/trio-dnm.

The DNG model evaluates joint data likelihoods for all possible combinations of trio genotypes P, M, F of the proband, mother and father as follows

$$
L_{P,M,F} = P(D|P, M, F) \cdot P(P|M, F) \cdot P(M, F). \tag{1}
$$

The first term represents the product of genotype likelihoods of the proband, mother and father; the second term represents the transmission probability (0.25, 0.5, or 1 if compatible with Mendelian inheritance or $\mu = 10^{-8}$ for each mutated allele); and the third term is the prior probability of drawing two genotypes M and F from the population. For details see the supplementary materials of [Conrad *et al.*, 2011] and bcftools/trio-dnm code.

The most likely combination of trio genotypes P, M, F incompatible with Mendelian inheritance is then selected and the posterior probability calculated as

$$P(\text{is DNM}) = \frac{L_{P,M,F}}{\sum_{p,m,f} L_{p,m,f}},$$

where summation in the denominator is over all possible genotype combinations, including combinations compatible with Mendelian inheritance. The program adds the rounded phred-scaled value of $1 - P(\text{is DNM})$ as the FORMAT/DNM annotation, bigger values indicate increased likelihood of the variant being a true *de novo* mutation.

In a test performed on 30,898 candidate *de novo* mutations there was perfect correlation between results from DeNovoGear and bcftools/trio-dnm implementations ($R^2 = 1.0$). Note that the DNG implementation in bcftools/trio-dnm uses the same mutation rates for both SNPs and indels and does not attempt to reproduce the calculation for sex chromosomes for reasons given further.

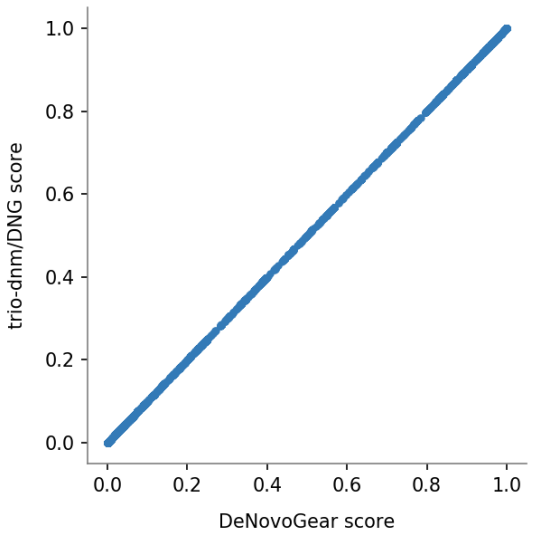

Figure 1. DNM probability as calculated by the original DeNovoGear program and its reimplementation in BCFtools/trio-dnm.

2 Allele-centric approach (trio-dnm model)

The DeNovoGear model suffers from a shortcoming which manifests itself in inflated false positive rate. The underlying reason is that genotype likelihoods were designed for germline variant calling and, to a degree, are robust to mapping and alignment errors. Specifically, if the observed fraction of alternate bases deviates strongly from the expected binomial sampling probability, the likelihood of heterozygous genotype will be smaller than the likelihood of homozygous genotype despite strong presence of the alternate allele. This is a property desired for germline variant calling but is unsuitable for DNM discovery. Indeed, the most common failure mode of the DNG model is the situation where alternate reads are present in both the proband and a parent, but as most likely are deemed the heterozygous genotype in the proband and homozygous genotypes in both parents. Although this behavior may lead to serendipitous discovery of valid biological cases (true mosaic mutations in parents), such discoveries are accidental and indistinguishable from false calls on the basis of call quality.

The revised approach described in this section fixes this problem by sensing the presence of alternate alleles in the parents directly, rather than relying on genotype likelihoods. Genotype likelihoods are used only to determine heterozygous genotype in the proband.

2.1 Allele Quality

Let $\epsilon_{x,i}$ denote the base error as before (i.e. the maximum of base and mapping error), but this time for each base $x \in \{A, C, G, T\}$ separately. For simplicity of notation we limit the description to single nucleotides, but indels are handled the same way. We introduce a new variable a_x which will represent the probability of base x being genuine and not an artifact. Then if the base x is present in reads covering a genomic position k_x times, the probability a_x can be expressed as

$$a_x = \begin{cases} 1 - \prod_{i=0}^{k_x} \epsilon_{x,i} & \text{if } k_x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This quantity is calculated at the bcftools/mpileup step by adding the `-a QS` option and it will be used to calculate a modified trio genotype likelihood by replacing $P(D|P, M, F)$ in equation (1) with a new term $P'(D|P_{xy}, M'_{xy,uv}, F'_{xy,rs})$ as explained next.

Extend the notation of genotype likelihood in the proband by adding subscripts to indicate the genotype in question, and of the parental likelihoods by adding the allelic context of the proband's genotype. For example, $M'_{01,00}$ stands for the likelihood of the homozygous reference genotype in the mother (0/0) conditioned on the heterozygous genotype in the proband (0/1). The modified trio genotype likelihood can be then expressed as

$$P'(D|P_{xy}, M'_{xy,uv}, F'_{xy,rs}) = P_{xy} \cdot M'_{xy,uv} \cdot F'_{xy,rs} \tag{3}$$

with maternal genotype likelihoods calculated as

$$M'_{xy,uv} = \prod_{i=0}^{4} \begin{cases} m_i & \text{if } i \in \{u, v\}, \\ 1 - m_i & \text{if } i \in \{u, v\} \setminus \{x, y\}, \\ 1 - m_i & \text{otherwise, if } u = v, \end{cases} \tag{4}$$

where m_x denotes the allele quality a_x in mother (and similarly for father). The first case checks for a strong presence of the alleles in the parental genotype uv; the second penalizes the presence of the assumed *de novo* allele in the parental genotype; and the third penalizes presence of multiple alleles in homozygous parental genotypes.

2.2 Modifications to DeNovoGear calculations

The math notes in [Conrad *et al.*, 2011] and the DeNovoGear code contain some errors and in some cases lack clarity.

1. The math notes state that conditional on a segregating site with 1 or 2 alternate alleles, genotypes with 1 alternate allele are found with frequency 3/5 and genotypes with 2 alternate alleles are found with frequency 2/5. This is in fact reversed, there are 4 configurations with a single alternate allele and 6 with two alternate alleles, giving frequencies 2/5 and 3/5, respectively:

```
00 01
00 10
01 00
10 00     .. 4 out of 10 = 2/5

00 11
01 01
10 01
01 10
10 10
11 00     .. 6 out of 10 = 3/5
```

2. It is not clear how the term 0.001 (as opposed to 0.002) has been derived in the math notes.
3. The DeNovoGear code deviates from the math notes, setting for example the prior probability of a single alternate allele in parents as

```
g_priors[i] = 0.995 * 0.002 * (3.0 / 5.0) * (4.0 / 5.0) * 0.5;
```

   or triallelic cases as

```
g_priors[i] = 0.002 * 0.002 / 414;
```

   No explanation is provided as for where the terms 0.995, 0.5 or 414 come from.
4. The DNG code for setting chrX priors has multiple bugs. For example, the transmission of chrX from mother to a boy does not depend on the paternal genotype, etc. From this reason bcftools/trio-dnm does not attempt to reimplement this part of the code.

2.3 Revised DeNovoGear priors

Label the reference allele 0 and consider at most three non-reference alleles $x \in \{1,2,3\}$. Set the prior probabilities of sampling reference and alternate genotypes as follows

$P_1 = 0.998$
: the probability of sampling two reference genotypes at sites with one alternate allele common in the population (biallelic sites)

$P_3 = (1-P_1)^2$
: the probability of sampling a non-reference genotype at sites with two different alternate alleles common in the population (triallelic sites)

$P_2 = 1 - P_1 - P_3 \approx 1 - P_1$
: the probability of sampling a non-reference genotype at biallelic sites

$P_4 = 10^{-26}$
: the probability of observing tetra-allelic site

2.3.1 Autosomal chromosomes

The prior probability of sampling four chromosomes from the population for parental genotypes M and F is then as follows

$$P(M,F) = \begin{cases} \text{0 alternate alleles:} & \\ \quad P_1 & M,F = 00,00 \\ \text{1 alternate allele:} & \\ \quad P_2 \cdot (1/15) \cdot (1/3) & M,F \in \{xx,xx\} \text{ where } x \in \{1,2,3\} \\ \quad P_2 \cdot (2/15) \cdot (1/3) & M,F \in \{00,xx\ ;\ xx,00\} \\ \quad P_2 \cdot (4/15) \cdot (1/3) & M,F \in \{0x,xx\ ;\ x0,xx\ ;\ xx,0x\ ;\ xx,x0\} \\ \quad P_2 \cdot (4/15) \cdot (1/3) & M,F \in \{0x,0x\ ;\ 0x,x0\ ;\ x0,0x\ ;\ x0,x0\} \\ \quad P_2 \cdot (4/15) \cdot (1/3) & M,F \in \{00,0x\ ;\ 00,x0\ ;\ 0x,00\ ;\ x0,00\} \\ \text{2 alternate alleles:} & \\ \quad P_3 \cdot (1/19) \cdot (1/3) & M,F \in \{00,xy\ ;\ 0x,xy\ ;\ \dots\} \\ \text{3 or more alternate alleles:} & \\ \quad P_4 & \end{cases}$$

2.3.2 Chromosome X in males

For chromosome X inheritance pattern in males we define analogously

$$\begin{aligned} P_{X,1} &= 0.999 \\ P_{X,3} &= (1 - P_{X,1})^2 \\ P_{X,2} &= 1 - P_{X,1} - P_{X,3} \approx 1 - P_{X,1} \end{aligned}$$

The priors for drawing the maternal genotype then become

$$P_X(M) = \begin{cases} \text{0 alternate alleles:} & \\ \quad P_{X,1} & M = 00 \\ \text{1 alternate allele:} & \\ \quad P_{X,2} \cdot (1/3) \cdot (1/3) & M = xx \text{ where } x \in \{1,2,3\} \\ \quad P_{X,2} \cdot (2/3) \cdot (1/3) & M \in \{0x\ ;\ x0\} \\ \text{2 alternate alleles:} & \\ \quad P_{X,3} \cdot (1/3) & M \in \{xy\ ;\ xz\ ;\ yz\} \end{cases}$$

2.3.3 Chromosome X in females

For chromosome X inheritance in females we approximate as

$$P_{XX}(M,F) = \begin{cases} \text{0 alternate alleles:} & \\ \quad P_1 & M,F = 00,0 \\ \text{1 alternate allele:} & \\ \quad P_2 \cdot (3/7) \cdot (1/3) & M,F \in \{00,x\ ;\ 0x,0\ ;\ x0,0\} \text{ where } x \in \{1,2,3\} \\ \quad P_2 \cdot (3/7) \cdot (1/3) & M,F \in \{0x,x\ ;\ 0x,x\ ;\ xx,0\} \\ \quad P_2 \cdot (1/7) \cdot (1/3) & M,F = xx,x \\ \text{2 alternate alleles:} & \\ \quad P_3 \cdot (1/9) \cdot (1/3) & M,F \in \{0x,y\ ;\ \ldots\ ;\ xx,y\ ;\ \ldots\ ;\ xy,y\ ;\ \ldots\} \end{cases}$$

3 Results

Overall performance

The test data consisted of 50,764 candidate calls from `de_novos.ddd_10k.29-01-2018.tab.gz`. For these plots calls on chrX were included, we note that the graphs look the same with chrX excluded.

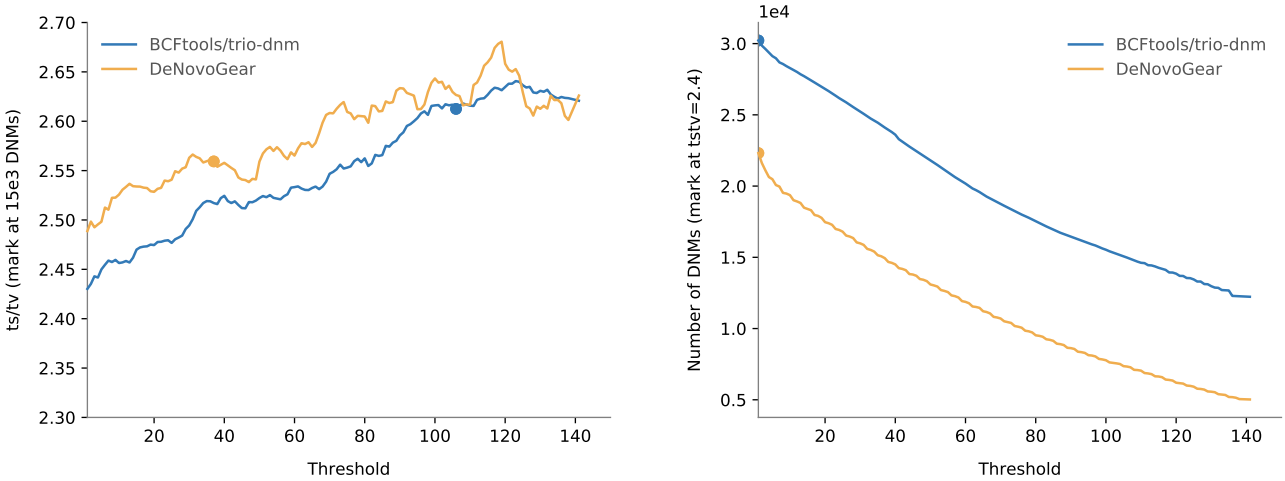

Figure 2. The ts/tv ratio and the number of sites at increasing DNM quality score cutoff. Circles mark points in the graph with the same number of sites and ts/tv, respectively. The qualities calculated by the new model (BCFtools/trio-dnm) are placed higher than those calculated by DeNovoGear and thus contain less noise in the same-sized call set.

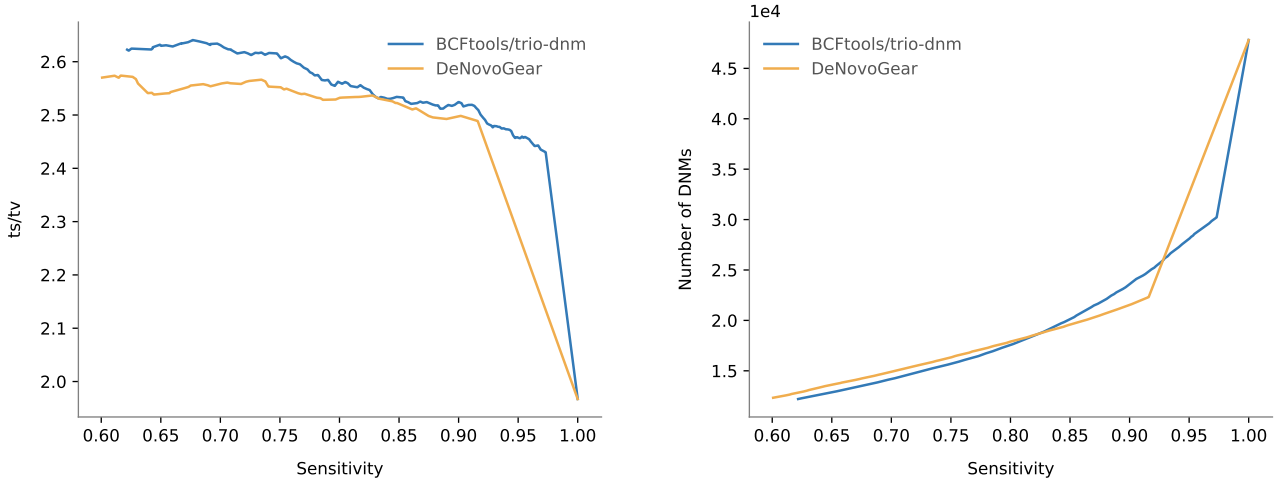

Figure 3. The ts/tv ratio and the number of sites at increasing DNM quality score cutoff plotted as a function of sensitivity with respect to a clean set of 15,213 high-quality *de novo*s from coding/splicing regions. Note that the truth set was generated by filtering the DeNovoGear calls and therefore it should have higher sensitivity.

Cases missed by DeNovoGear

In all IGV screenshots that follow, displayed are from top to bottom father, mother, and the proband. The blue and pink color shows the forward and reverse strand orientation of the reads.

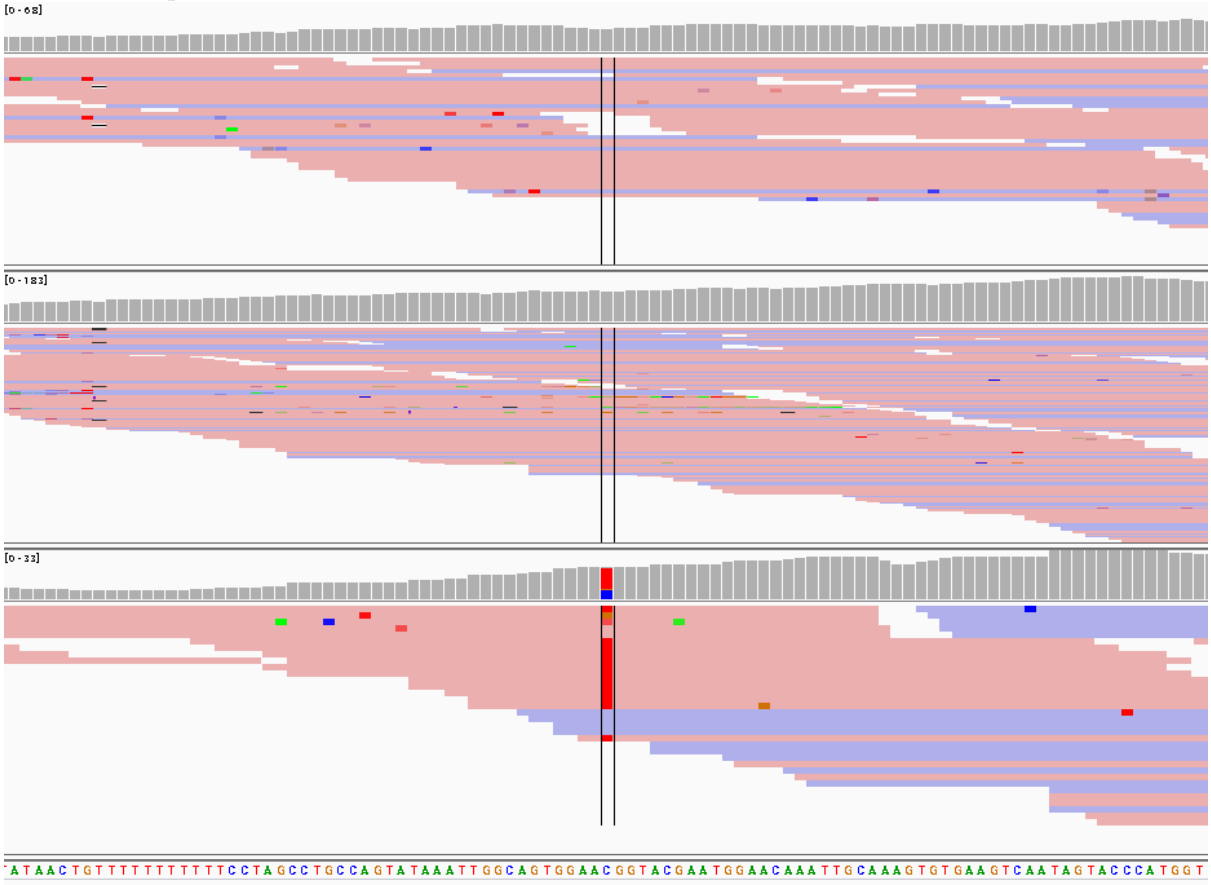

Figure 4. Call missed by DeNovoGear, chromosome X, the proband is male.

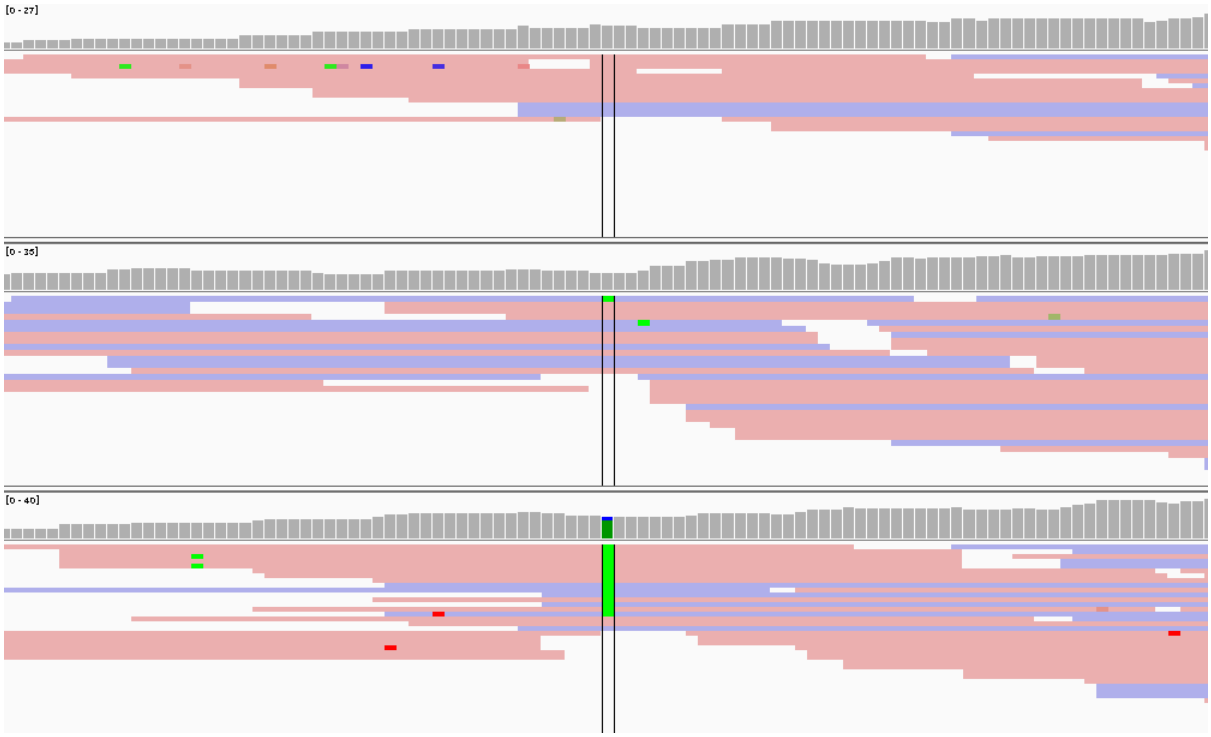

Figure 5. Call missed by DeNovoGear, chromosome X, the proband is female.

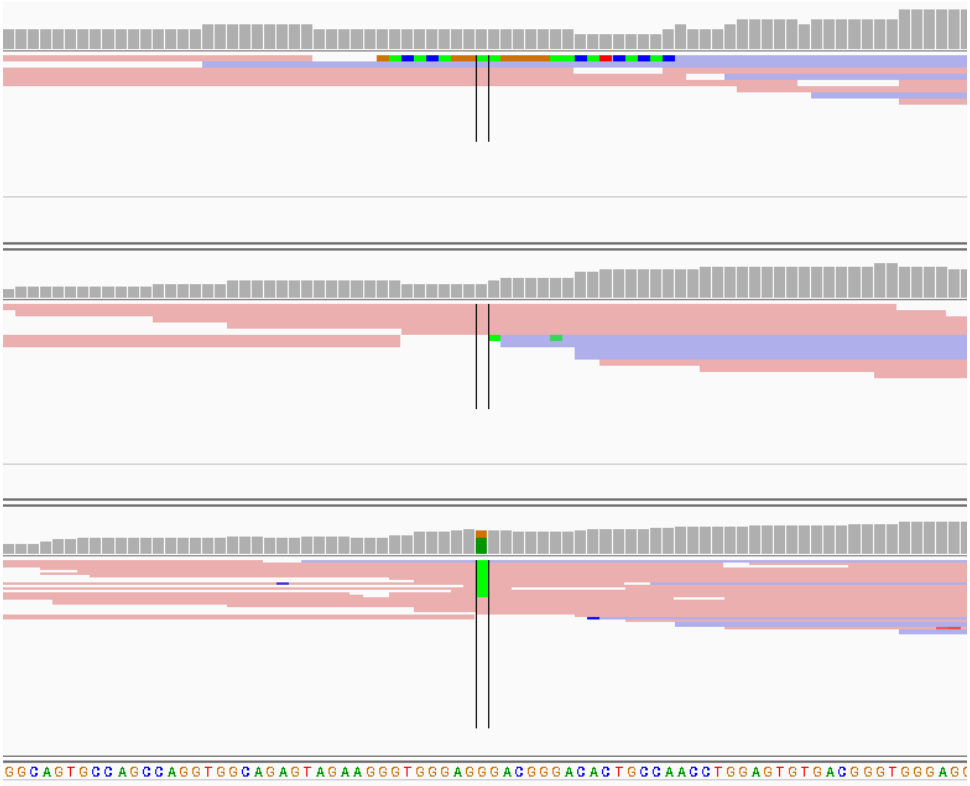

Figure 6. Call missed by DeNovoGear, chromosome 1.

False calls by DeNovoGear

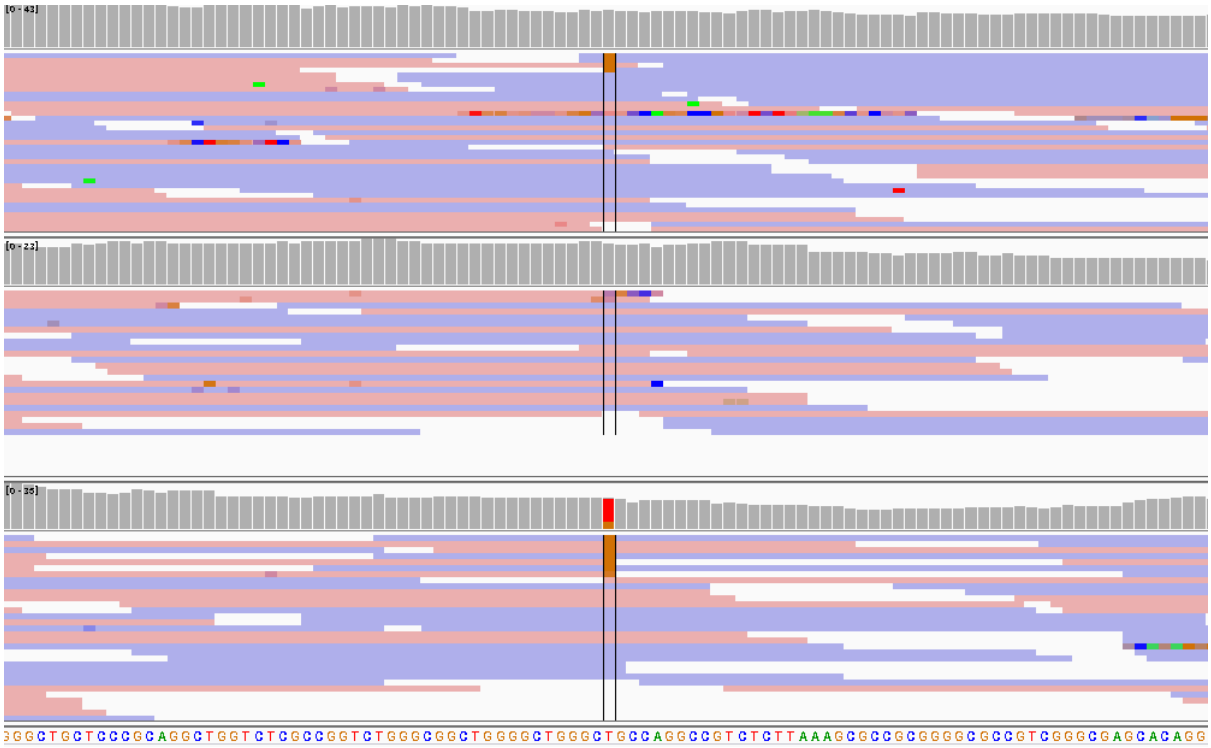

Figure 7. This is likely a false call caused by mapping artifact, the alternate allele appears in both parents. Can't exclude parental mosaicity.

Cases missed by BCFtools/trio-dnm

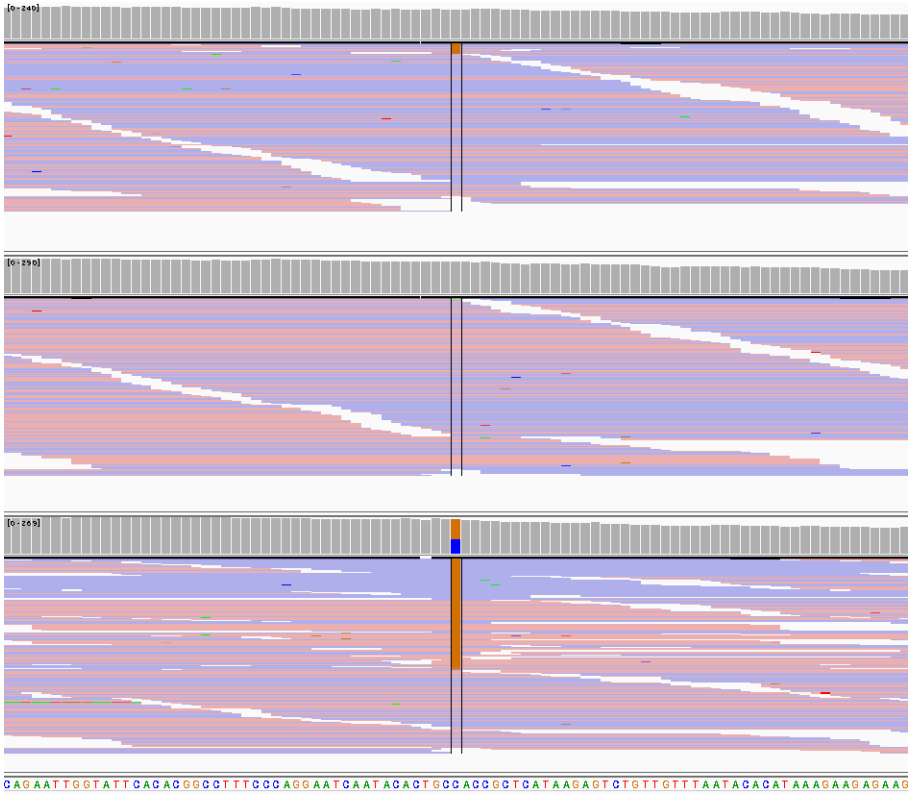

Figure 8. Arguably, this might be a valid biological case with low-level mosaicity in the father. However, it could be also argued that the program correctly determined that the alternate allele is not *de novo* in the proband and a different calling model should be used. Note that BCFtools/trio-dnm has a parameter that allows to adjust noise tolerance and effectively call cases like this at the cost of decreased specificity.

Uncertain cases - should these be called or not?

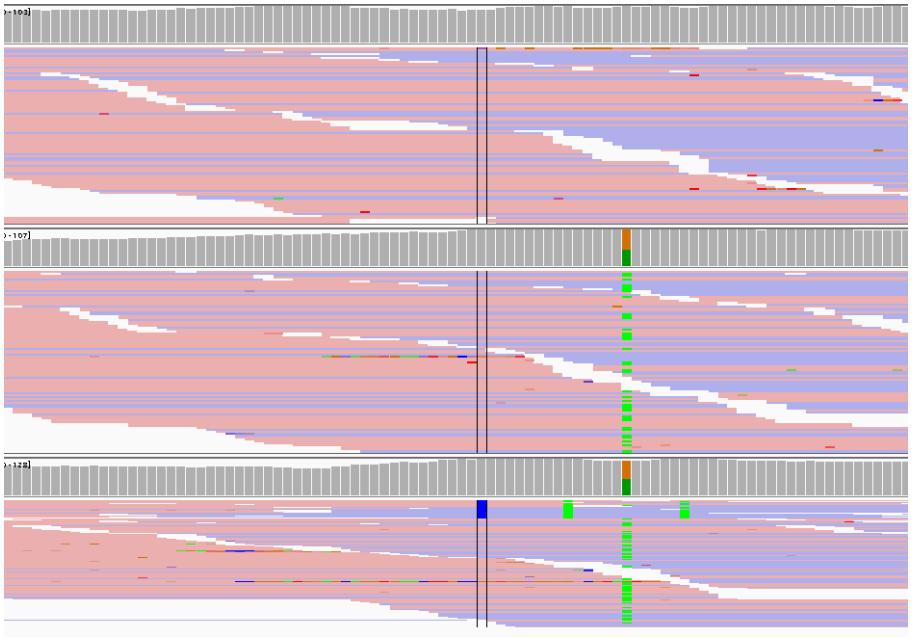

Figure 9. Mosaic variant in the child or an artifact?

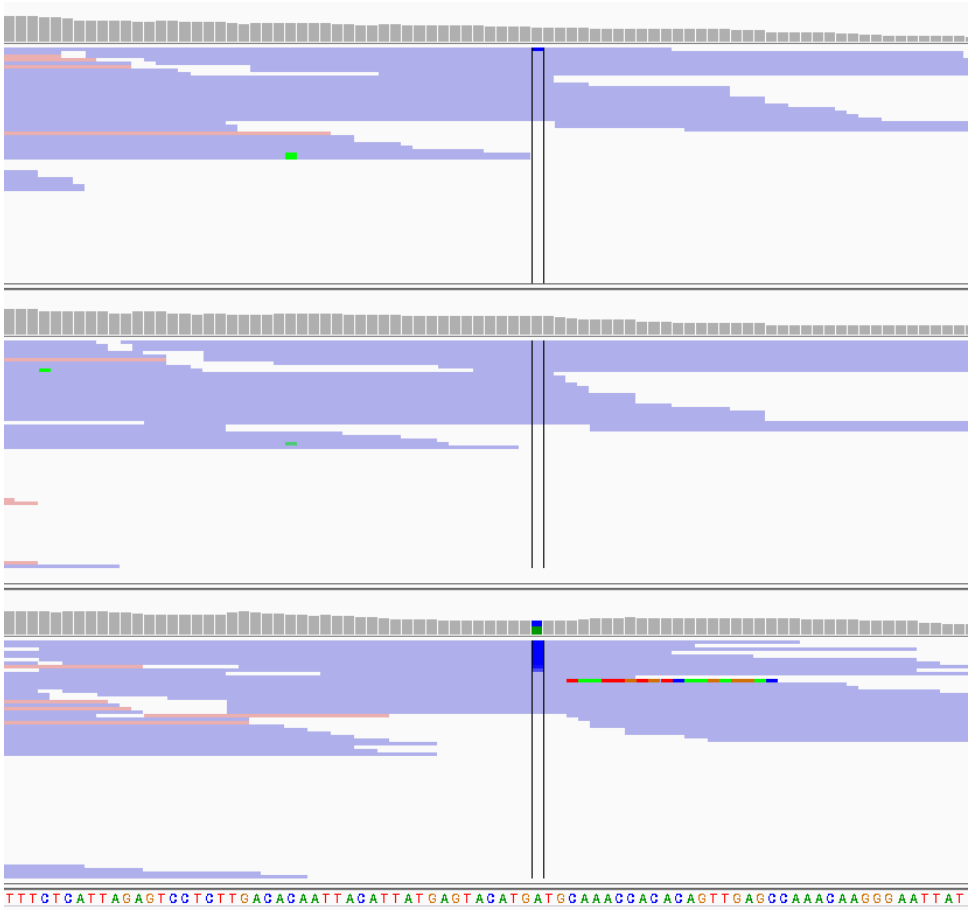

Figure 10. Is this a valid DNM? Note the single alternate read in the father.

References

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

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

[Li, 2010] Li H, The revised MAQ model, (2010)
http://samtools.github.io/bcftools/samtools.pdf