De Novo Variant Caller using Google Genomics API

Subhodeep Moitra
Carnegie Mellon University
subhodee@cs.cmu.edu

September 5, 2014

Abstract

This tool identifies de novo genetic mutations in a family trio (child, mother and father). De novo genetic mutations are defined as those mutations which occur in the child but do not occur in the genomes of the parents. This tool queries the Google Genomics API in order to search and retrieve genetic variants and reads. It implements a Bayesian inference algorithm for identifying de novo mutations from the retrieved data. This document describes theory, design and early experiments for the de novo caller.

1 Background

De novo genetic mutations are a set of rare genetic mutations that occur in the genomes of a child but do not occur in the genome of the child's parents. These mutations are being increasingly implicated to play an important role in a variety of diseases such as Autism [2] and Schizophrenia. With the advent of Whole Genome Sequencing, a large number of these mutations are being discovered at both the exonic and the intronic regions.

Detecting de novo mutations is a challenging task. Some of the issues that affect de novo mutation detection are :

- **Rare event**: A de novo mutation is a very rare event ($\sim$1 in 10^8 bp).
- **Sequencing errors**: Sequencing errors occur at the rate of 1 in 100 bp. This makes it much more likely to observe a sequencing error rather than a de novo mutation.
- **Read Depth**: Since the reads mapped to a particular position are essentially random, it is possible that the denovo mutation is completely skipped by inadequate coverage at that position.
- **Misalignment**: Variant detection pipelines assume a high quality alignment of reads. In areas of poor alignment, variant detection suffers.
- **Expensive Validation**: Candidates have to be verified through expensive Sanger sequencing.
- **Structural Variation**: Complications arise through structural variations in the chromosome through insertions, deletions, inversions, copy number variations, translocations, etc.

The Google Genomics API [5] provides services to import and query variants and reads data. This tool assumes that variants and reads are available for the trio under study and have been imported into the Genomics API. The tool provides two callers (1) Variant Caller (2) Read Caller. The variant caller walks over the trio of genomes and examines SNPs avaiable through the `variants` API. After filtering for variants based on quality cutoffs and mendelian inheritance rules, a set of candidate calls are generated and passed onto the Read Caller for finer discrimination. The Read Caller makes requests to the `reads` API to fetch mapped reads for these candidate position and makes a finer call using a bayesian inference algorithm. Validation experiments against NA12878 trio from Platinum genomes [6] were also run and compared against experimentally validated de novo and somatic mutations [4] ; results are reported in section 3.

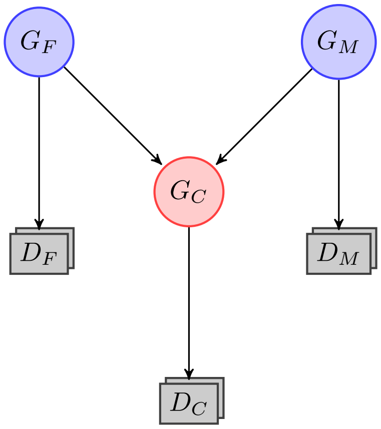

Figure 1: The hidden variables G_F, G_M and G_C correspond to the genotypes of the father, mother and child respectively. The genotypes take one among 10 values from $AA, AC, AT, AG, CC, CT, CG, GG, GT$ and TT. The trio of genotypes values can thus have 10^3 possible values. The observed variables (D_F, D_M and D_C) are the bases present at the candidate position for a particular read

1.1 Bayes Net

We employ a probabilistic inference framework using bayes nets for de novo variant calling. A bayes net belongs to the class of mathematical models known as graphical models which are a marriage between graph theory and probability theory [3]. A Bayes Net is a directed acyclic graph which defines conditional independence relationships between random variables in a probability distribution. A Bayes net can be used for sampling, marginal inference and decoding of the hidden variables.

We follow the bayes net implementation for de novo variant calling as in Li et al [1]. They model a bayes net to encode the diploid Mendelian relationship between the members of the trio. See Figure 1 for an illustration of the bayes net. The hidden variables G_F, G_M and G_C correspond to the genotypes of the father, mother and child respectively. The genotypes take one among 10 values from $AA, AC, AT, AG, CC, CT, CG, GG, GT$ and TT. The trio of genotypes values can thus have 10^3 possible values. The observed variables (D_F, D_M and D_C) are the reads obtained at a candidate position. We fix the baseline sequencing error rate ϵ_{seq} and the baseline de novo mutation rate ϵ_{denovo}. In equation 1, we apply Bayes rule to calculate the probability of a particular genotype trio given observed reads.

$$P(G_F, G_M, G_C | D_F, D_M, D_C) = \frac{P(D_F, D_M, D_C | G_F, G_M, G_C) P(G_F, G_M, G_C)}{\sum_{G_F, G_M, G_C} P(D_F, D_M, D_C | G_F, G_M, G_C) P(G_F, G_M, G_C)} \tag{1}$$

The term $P(D_F, D_M, D_C | G_F, G_M, G_C)$ is the data likelihood and is calculated as in equation 2.

$$P(D_F, D_M, D_C | G_F, G_M, G_C) = P(D_F | G_F) P(D_M | G_M) P(D_C | G_M) \tag{2}$$

$$= \left(\prod_{f=1}^{|D_F|} P(R_f | G_F) \right) \left(\prod_{m=1}^{|D_M|} P(R_m | G_M) \right) \left(\prod_{c=1}^{|D_C|} P(R_c | G_C) \right) \tag{3}$$

The term $P(G_F, G_M, G_C)$ encodes the mendelian inheritance relationship in the trio and can be factorized further according to equation 4. Furthermore, this likelihood can be expressed in terms of ϵ_{seq} and ϵ_{denovo} as in equation 5 where N_{denovo} corresponds to the number of G_C cases that are de novo given the parents ; similarly $N_{mendelian}$ corresponds to the number of G_C cases that are mendelian given the parents.

$$P(G_F, G_M, G_C) = P(G_C|G_F, G_M)P(G_F)P(G_M) \tag{4}$$

$$P(G_C|G_F, G_M) = \begin{cases} \epsilon_{denovo}/N_{denovo} & \text{if } G_C \text{ is a denovo mutation} \\ (1-\epsilon_{denovo})/N_{mendelian} & \text{otherwise} \end{cases} \tag{5}$$

The base likelihood $P(R|G)$ conditioned on the genotype of the sample is calculated according to equation 6.

$$P(R|G) = \begin{cases} 1-\epsilon_{seq} & \text{if } G \text{ is homozygous and } R \in G \\ \epsilon_{seq}/3 & \text{if } G \text{ is homozygous and } R \notin G \\ 0.5(1-2\epsilon_{seq}/3) & \text{if } G \text{ is heterozygous and } R \in G \\ \epsilon_{seq}/3 & \text{if } G \text{ is heterozygous and } R \notin G \end{cases} \tag{6}$$

1.2 Bayesian Inference Strategies

In the previous section we defined the bayes net model. In this section we will discuss the inference strategies for determining if a candidate position is a de novo variant. We define the function $I_{denovo}(G_F, G_M, G_C)$ as an indicator function when the child genotype is a de novo mutation. We also calculate the likelihoods of all the de novo and the mendelian inheritance cases as described in equation 7.

$$L_{denovo} = \sum_{G_F, G_M, G_C} P(D_F, D_M, D_C|G_F, G_M, G_C)P(G_F, G_M, G_C)I_{denovo}(G_F, G_M, G_C) \tag{7}$$

$$L_{mendelian} = \sum_{G_F, G_M, G_C} P(D_F, D_M, D_C|G_F, G_M, G_C)P(G_F, G_M, G_C)I_{mendelian}(G_F, G_M, G_C) \tag{8}$$

1.2.1 Maximum A Posteriori (MAP)

Call a de novo variant if condition 9 holds.

$$I_{denovo}(\underset{G_F, G_M, G_C}{\arg\max}\ P(G_F, G_M, G_C|D_F, D_M, D_C)) \tag{9}$$

1.2.2 Bayesian Risk Minimizer (Bayes)

Call a de novo variant if condition 10 holds. This inference rule minimizes the empirical bayesian risk.

$$\frac{L_{denovo}}{L_{denovo} + L_{mendelian}} > 0.5 \tag{10}$$

1.2.3 Likelihood Ratio Test (LRT)

Call a de novo variant if condition 11 holds where $\mathcal{T}$ is a user specified likelihood ratio threshold. Note that LRT is the same as Bayes if $\mathcal{T}$ is equal to 1.

$$\frac{L_{denovo}}{L_{mendelian}} > \mathcal{T} \tag{11}$$

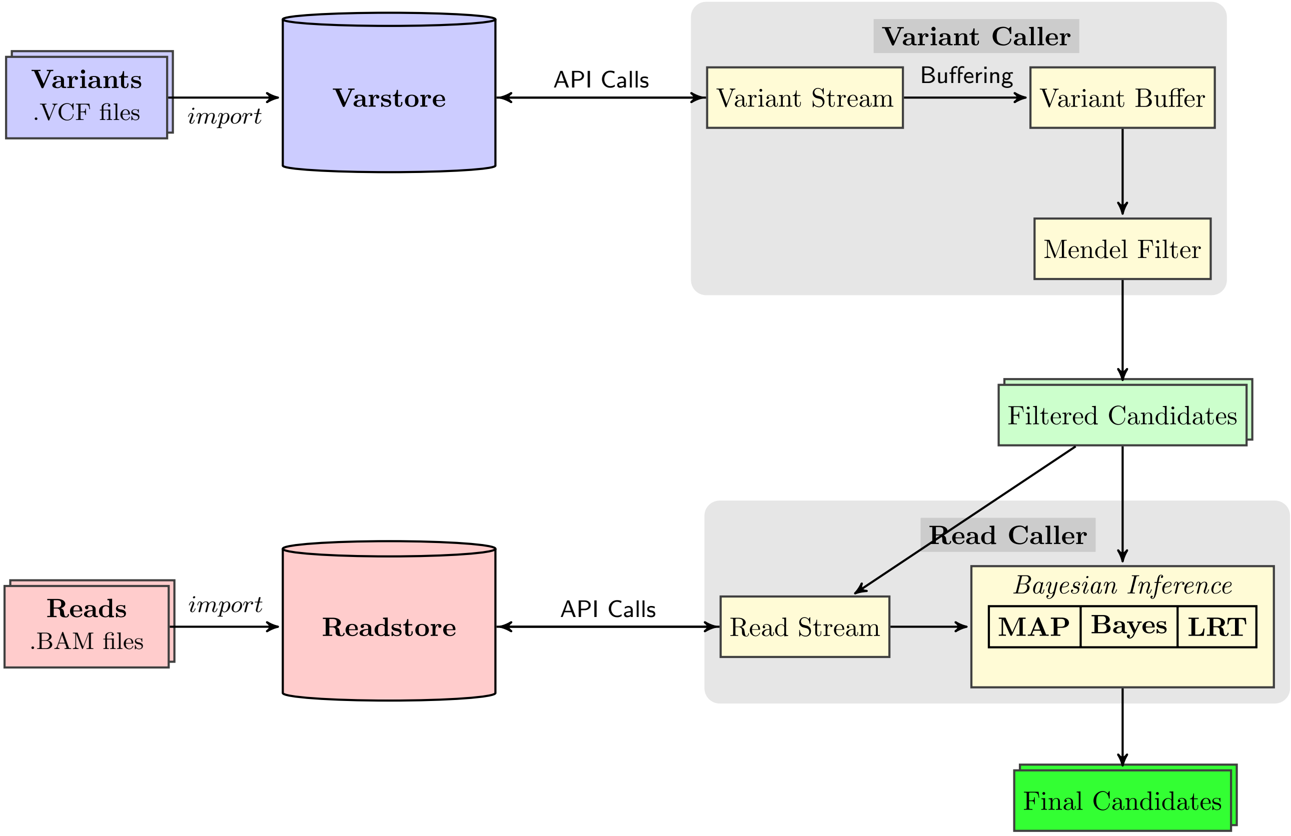

Figure 2: De novo variant caller design : There are two main components (1) Variant Caller interacting with Varstore (2) Read Caller interacting with Readstore. The Variant caller makes API calls to Varstore using a Variant Stream object that interfaces with a variants buffer. The buffer manages overlapping variants from the trio. Once a set of suitable variants have been obtained it is filtered using mendelian inheritance checks. These candidates are then passed onto the the Read Caller which obtains reads for the candidate position from readstore. The candidates are filtered using one of three Bayesian inference procedures (1) MAP - Maximum a posteriori (2) Bayes Rule (3) LRT - Likelihood Ratio test. Successful candidates are written onto disk.

2 Design and Implementation

This tool is designed as a standalone command line tool for performing inference. This tool assumes that variants and reads are available for the trio under study and have been imported into the Genomics API.

There are two main components (1) Variant Caller interacting with Varstore (2) Read Caller interacting with Readstore. See Figure 2 for an illustrative diagram. The Variant caller makes API calls to Varstore using a Variant Stream object that interfaces with a variants buffer. The buffer manages overlapping variants from the trio. Once a set of suitable variants have been obtained it is filtered using mendelian inheritance checks. These candidates are then passed onto the the Read Caller which obtains reads for the candidate position from readstore. The candidates are filtered using one of three Bayesian inference procedures (1) MAP - Maximum a posteriori (2) Bayes Rule (3) LRT - Likelihood Ratio test. Successful candidates are written onto disk.

2.1 Concurrency / Multi-threading

Profiling indicated that a lot of time was spent in network I/O delays. To speed up the execution, requests to the `variants` API and the `reads` API can be parallelized using multiple threads. Performance testing showed that multi-threading achieves a linear speedup with the number of threads(Graphs not shown). For this purpose the Variant Caller and the Read Caller implements a helper class `SimpleDenovoRunnable` and `BayesDenovoRunnable` respectively. Their behavior can be controlled by the `--num_threads` option.

2.2 Commandline usage

This tool is predominantly meant to be used from the commandline.

Typical commandline use

See below for typical commandline usage :

```
java -jar target/denovo-variant-caller-0.1.jar --caller full \
--client_secrets_filename ${HOME}/Downloads/client_secrets.json \
--dataset_id 14004469326575082626 \
--dad_callset_name NA12891 \
--mom_callset_name NA12892 \
--child_callset_name NA12878 \
--chromosome chr1 \
--start_position 1 \
--end_position 14000000 \
--log_level debug \
--num_threads 25 \
--output_file NA12878_full.calls
```

All options

See below for all available options:

```
Usage: DenovoMain [flags...]
 --caller [VARIANT | READ | FULL]       : The caller mode
 --child_callset_name <name>            : Child's callset name e.g. NA12879
 --chromosome <name>                    : specify the chromosomes to search
                                          (specify multiple times for multiple
                                          chromsomes)
```

```
--client_secrets_filename <file>       : Path to client_secrets.json
--dad_callset_name <name>              : Dad's callset name e.g. NA12877
--dataset_id <id>                      : Dataset id
--denovo_mut_rate <rate>               : Specify the denovo mutation rate
                                         (default 1e-8)
--end_position <position>              : end position ( usually set
                                         automatically )
--inference_method [MAP | BAYES | LRT] : Inference method (map | bayes | lrt)
--input_calls_file <file>              : File to read from
--log_file <file>                      : specify the log file
--log_level [ERROR | INFO | DEBUG]     : specify the logging level
--lrt_threshold <sig_level>            : likelihood ratio test significance
                                         level (default 1. ;higher the
                                         stricter)
--max_api_retries <num>                : max api retry count (default 5)
--max_variant_results <num>            : max variants returned per request
                                         (default 10000)
--mom_callset_name <name>              : Mom's callset name e.g. NA12878
--num_threads <num>                    : Specify the number of threads
                                         (default 1 ; 1 to 50 suggested)
--output_dir <dir>                     : File to write results
--output_file <file>                   : File to write results
--seq_err_rate <rate>                  : Specify the sequence error rate
                                         (default 1e-2)
--start_position <position>            : start position ( usually 1 )
```

3 Experiments

We ran validation experiments against the NA12878 trio (NA12878 child, NA12891 father, NA12892 mother) from Platinum genomes [6] and compared them against experimentally validated de novo and somatic mutations [4]. The results are shown in Table 3. The tool has high recall $\sim 96\%$. The precision is less than desirable ; however it is only a lower bound since the NA12878 trio is derived from cell lines and hence has accumulated a lot of somatic cell mutations. A blood line dataset may reveal more accurate numbers. The compute time is favorable ; 3 hours to go over 3 Billion base pairs.

Figure 3 shows a false de novo mutation candidate at chr10:23,662,774 in NA12878 trio. The top two tracks correspond to parents (NA12891 and NA12892) while the bottom track (NA12878) corresponds to the child genome. Father Genotype from `variants` API is AA, mother genotype is AA while child is AG. Variant is called by variant caller but correctly rejected by read caller due to insufficient evidence from reads. The Bayesian inference reads caller observes that there have been occurrences of some G bases in the parents which diminishes the likelihood of a denovo mutation.

Figure 4 shows a true experimentally validated germline mutation at chr1:75,884343 in NA12878 trio. The top two tracks correspond to parents (NA12891 and NA12892) while the bottom track (NA12878) corresponds to the child genome. Inferred father Genotype is TT, mother genotype is TT while child is CT. De novo mutation is called correctly by both variant and read caller.

4 Future Work and Conclusion

This tool identifies de novo genetic mutations in a family trio (child, mother and father) by exercising the Google Genomics API to fetch variants and reads which is subsequently . It provides concurrency options

Filter	Num Calls Filter	Precision	Recall	Time	Throughput
Variant	4090	22.12%*	96.69%	∼3 hours (20 threads)	∼3 mbps(20 threads)
Variant + BayesNet	2207	40.55%*	95.62%	∼30 mins(20 threads)	∼ 1.5 mbps (20 threads)

Table 1: Results on NA12878 trio (NA12878 child, NA12891 father, NA12892 mother) from Platinum genomes [6] compared against experimentally validated de novo and somatic mutations [4]. The tool has high recall ∼ 96%. ** The precision is inconclusive ; most likely is a lower bound since the NA12878 trio is derived from somatic cell lines. A follow up blood line experiment will likely reveal more accurate numbers.* The compute time is favorable; 3 hours to go over 3 Billion base pairs.

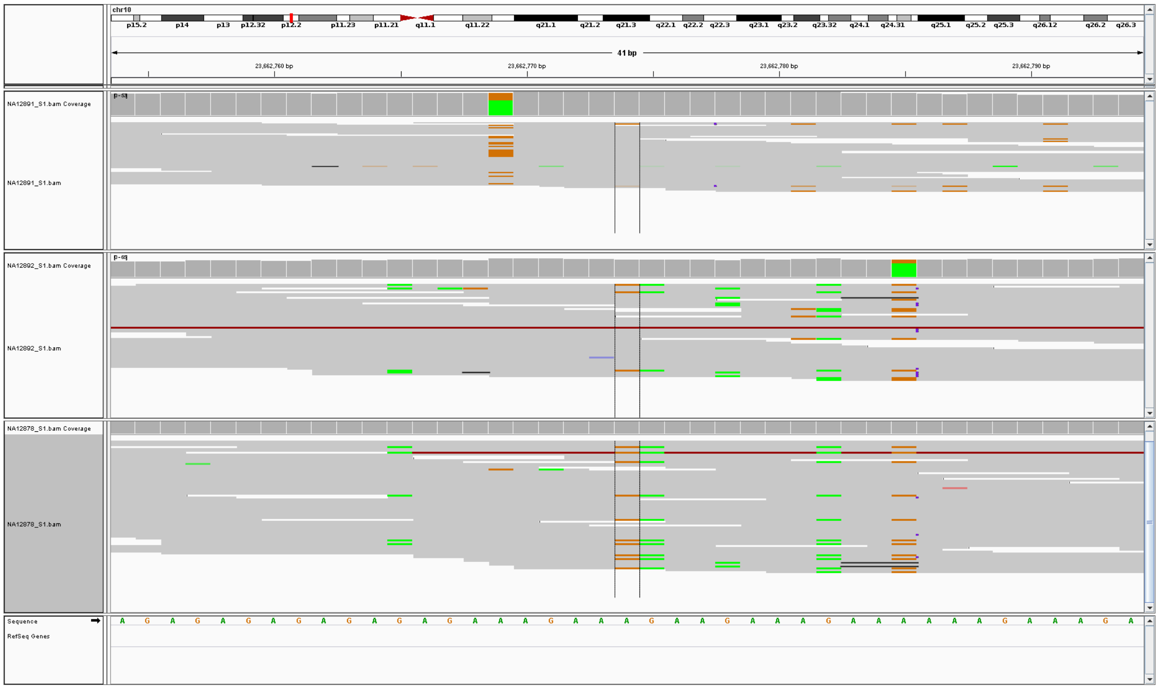

Figure 3: Shows a false candidate for a denovo mutation at chr10:23662774 in NA12878 trio. The top two tracks correspond to parents (NA12891 and NA12892) while the bottom track corresponds to the child genome (NA12878). Father Genotype variant is AA, mother genotype is AA and child is AG. Variant is called by variant caller but correctly rejected by read caller due to insufficient evidence from reads. The bayesian inference reads caller observes that there have been occurrences of some G bases (brown lines) in the parents which diminishes the likelihood of a denovo mutation.

to speed up execution. Experiments were run against the NA12878 trio which showed high recall ∼ 96% against experimentally validated germline and cell line mutations. Precision scores were inconclusive and will need follow up experiments against blood derived genomes.

References

[1] Li B et. al, *A likelihood-based framework for variant calling and de novo mutation detection in families*, PLoS Genetics, 2012 Volume 8, Number 10, Pages e1002944 1.1

[2] Michaelson et al, *Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation*, Cell, 2012, Volume 151, Number 7, Pages 1431 - 1442 1

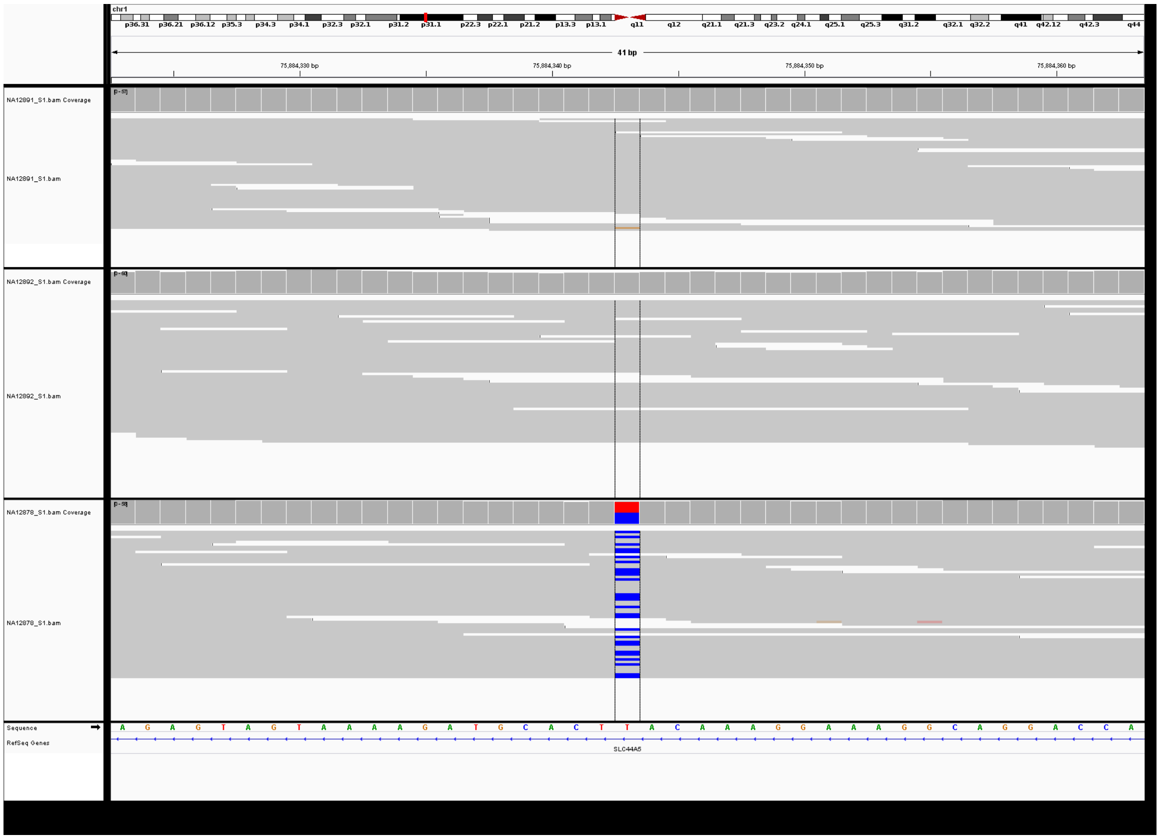

Figure 4: Shows a true experimentally validated germline mutation at chr1:75,884343 in NA12878 trio. The top two tracks correspond to parents (NA12891 and NA12892) while the bottom track (NA12878) corresponds to the child genome. Inferred father Genotype is TT, mother genotype is TT while child is CT. Variant is called by correctly by both variant and read caller

[3] Wainwright, Jordan, *Graphical Models, Exponential Families, and Variational Inference*, Foundations and Trends in Machine Learning, 2008, Volume 1, Issue 1-2, pp 1-305 1.1

[4] Conrad et al., *Variation in genome-wide mutation rates within and between human families*, Nature Genetics, July 2011 Volume 43, Number 7, pp 712-715 1, 3, 1

[5] *Google Genomics API*, `https://developers.google.com/genomics/` 1

[6] *Platinum Genomes - Illumina Sequencing*, `http://www.illumina.com/platinumgenomes/` 1, 3, 1