

Finding Causative Mutation Candidates in Rare Disease Studies using NextGENe's Variant Comparison Tool

August 2011

John McGuigan, Megan Manion, Shouyong Ni, Sean Liu, C.S. Jonathan Liu

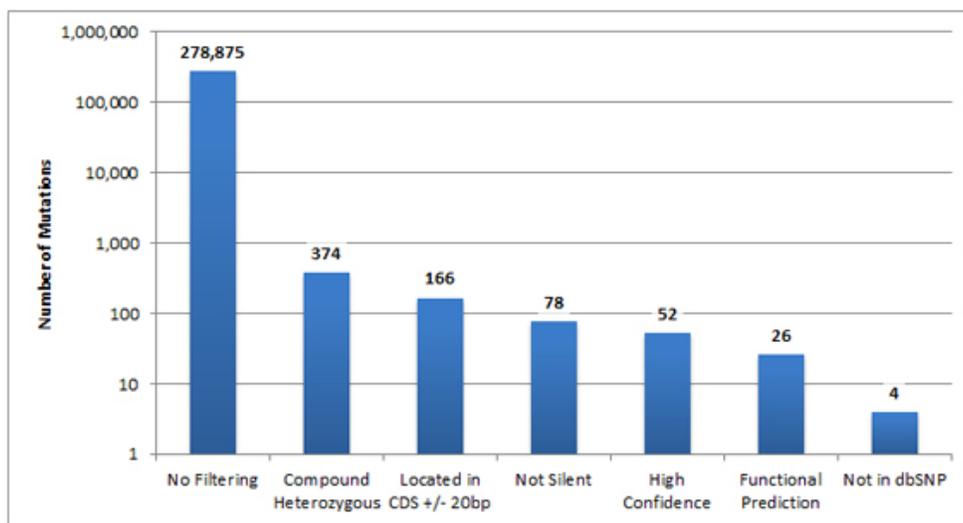


Figure 1 – Number of Mutations Left After Each Filtering Step

Introduction

As the cost of sequencing continues to decline, large sequencing projects are no longer limited to large labs with dedicated bioinformaticians. Increased sequencing output requires tools that can handle more advanced applications, while remaining easy to use for biologists. In collaboration with researchers at the NIH, SoftGenetics has developed a biologist and geneticist-friendly tool designed to speed up and simplify the discovery of causative mutations. Sequencing the exomes of a small family or several individuals with similar phenotypes can result in a list of hundreds of thousands of possible mutations. NextGENe's Variant Comparison tool makes it easy to narrow down this list based on a number of criteria:

- Expected genotype - including control samples and inheritance patterns
- Annotated gene information - exons, coding sequences, splice sites, etc
- Database information – dbSNP
- Functional Prediction - dbNSFP (including 1000 genomes frequency and PhyloP)
- Type of mutation – Indels or Substitutions (Silent, Missense, or Nonsense)
- Inclusive and Exclusive region of interest (ROI) filtering
- Mutation confidence score filtering

Procedure

After aligning reads in NextGENe to take advantage of its superior indel detection or using the BAM reader tool to convert pre-aligned reads to NextGENe format, the projects are loaded into the Variant Comparison tool. There are six options for comparing the samples (figure 2):

1. Show all mutations
2. Show shared or different mutations
 - a. Minimum coverage (both samples) and minimum difference can be specified
 - b. "Ignore 0% Mutations" will always treat positions with no mutant alleles as different even if they are in the expected range. For example, a position with no mutant alleles in one sample and 15% mutant alleles in another will be treated as different even if the range is set to 30%.
3. Show mutations with low coverage in at least one sample
4. Manually specify expected mutation types Homozygous, Heterozygous, Negative, Undetermined, or some combination
5. Specify relationship (Father, Mother, Son, Daughter) and phenotype (Affected or Unaffected) for each sample and select an inheritance template to automatically adjust expected mutation types
6. Perform compound heterozygous filtering
 - a. Requires sequence data for at least one affected offspring
 - b. Additional samples can be used for more specific filtering

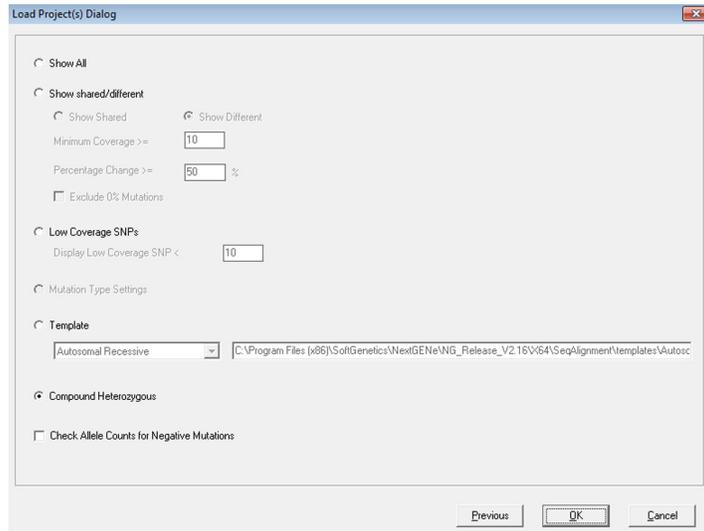


Figure 2 – Mutation Comparison Selection

“Check Allele Counts for Negative Mutations” will look up and display actual allele counts for negative samples and display them instead of ‘N/A’. This takes extra processing time.

Functional prediction, conservation, and 1000 genomes information can be imported using the “Convert dbNSFP files” tool. The tool includes a link to the website where the database can be downloaded. The archived files should be downloaded and extracted and then imported using the tool. NextGENe will index the files and store them in a specified location. From there, the database can be queried from within the viewer manually. New human genome alignment projects will query the database automatically. More information about the scores can be found in the paper published in Human Mutation (1), which is linked in the “About Prediction Scores” dialog from the Help menu in the viewer.

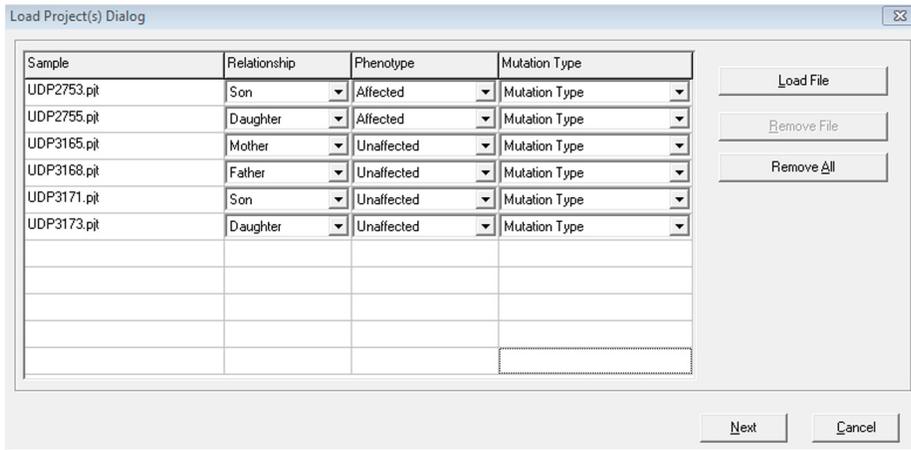


Figure 3 – The sample projects are loaded in the Variant Comparison Tool

In this example six sequenced exomes were examined to find potential causative mutations that follow a compound heterozygous inheritance pattern causing a form of epilepsy (figure 3). There were two parents, two affected children, and two unaffected children. For each sample mutations were called if they occurred in regions with at least 5x coverage and occurred in at least 3 reads or 20% of reads aligned at a position. If any sample had below the mutation coverage cutoff (5x) at a position, it was excluded from the final report.

Results

There were a total of 278,875 mutation calls in at least one of the six exomes. There are advantages and drawbacks to using many samples in a comparison. Using more samples allows more specific filtering, but may cause false negatives due to some samples not having adequate coverage. The compound het report lists all possible pairs of mutation calls where both are in the same gene and one is inherited from each parent. Using only compound heterozygous filtering of the initial 278,875 mutation calls reduced the list to 374 calls allowing for 497 possible pairs. There are many filters available for narrowing down the list even further (figure 4). Additional filtering was applied and the results are summarized in table 1.

Each dbNSFP score is passed or failed for each call independently and filtering is performed based on the minimum number of passing scores (between 1 and 5). By default every score will pass- all classifications are checked and the score filter is set to 0. Scores can be intentionally failed in order to avoid counting them towards the total number of passing scores for a given mutation call.

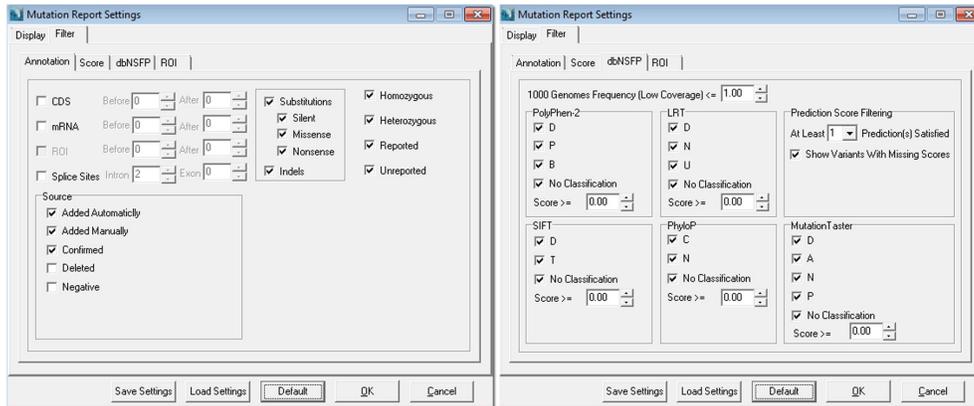


Figure 4 – Some of the filters available in the Variant Comparison Tool

Filtering	Number of Mutations	Number of Mutation Pairs
1. Compound Heterozygous	374	497
2. CDS +/- 20 bp	166	183
3. Hide Silent Mutations	78	71
4. Mutation Confidence Score >= 8	52	39
5. Hide dbSNP Mutations	6	3

Table 1 – Mutation Filtering Results

At each step the new filtering was applied first (such as hiding silent mutations) and the resulting lists of mutation calls for each sample were compared for compound heterozygous inheritance. The final 2 pairs of variants included a pair in the gene known to cause this disorder. Figure 5 shows a comparison for one of those variants across all of the samples. Both of these mutations occurred at positions predicted to be conserved by PhyloP. One was predicted to be damaging according to PolyPhen-2.

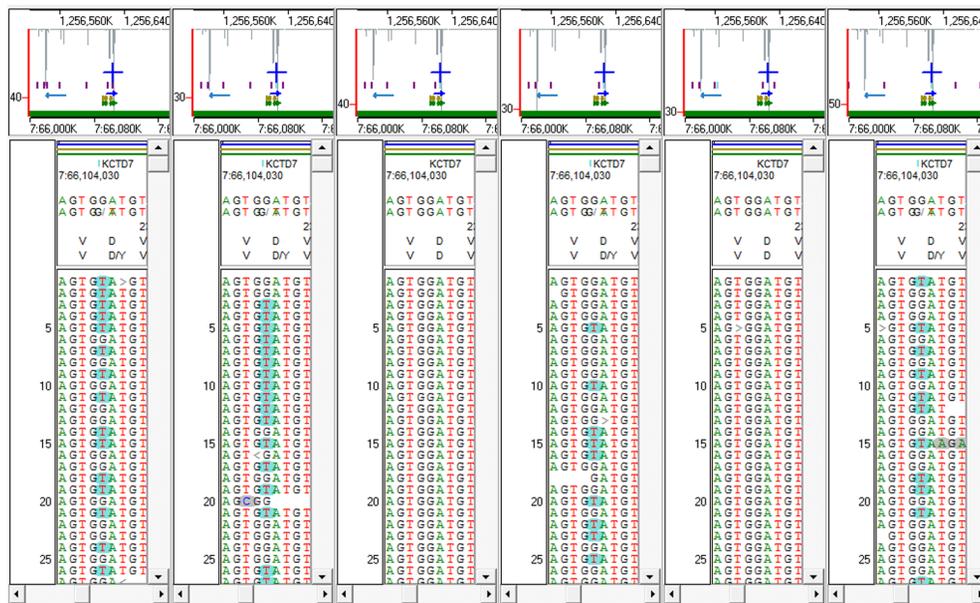


Figure 5 – One of two causative mutations found as compared across all six projects. The projects (left to right) are the two affected children, the mother, the father, and the two unaffected children. The second unaffected child has this mutation but does not have the other mutation in the same gene.

Discussion

NextGENe's improved Variant Comparison tool makes it very simple to quickly sort through thousands of mutation calls in order to find a few dozen (or even fewer) candidate mutations that can be confirmed with Sanger sequencing and assessed for their phenotypic impact. NextGENe is able to import 1000 genomes frequencies and several functional prediction scores from the dbNSFP database including PolyPhen-2, SIFT, LRT, and MutationTaster. Hiding reported (dbSNP) variations may be useful in some cases, but dbSNP does contain some clinical variants and some compound heterozygous pairs may involve a known SNP and a novel mutation. Knowledge about the disease can be used to include or exclude certain genes or regions in order to improve filtering even more (figure 6).

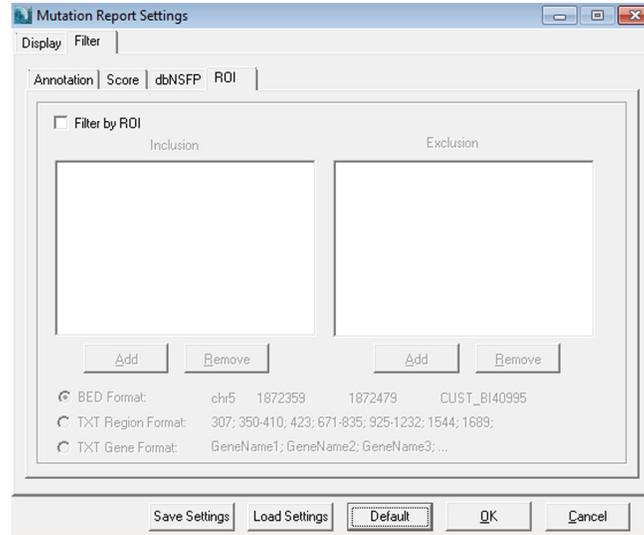


Figure 6 – Multiple ROI files (BED files or text files specifying positions or gene names) can be loaded and used for filtering in an inclusive or exclusive manner.

Depth of coverage is very important when comparing multiple projects. Ideally all samples will have 30x or more coverage throughout the targeted regions so that confident mutation calls can be made. Missing data in any one sample may cause false negatives because that project cannot be used for filtering. In this test case all samples had more than 25x coverage at the potential causative mutation positions. Each whole-exome sequencing project consisted of 32 to 67 million aligned 100 bp paired-end Illumina reads.

References

[1] Liu, X., Jian, X. and Boerwinkle, E. (2011), dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation*, 32: 894–899. doi: 10.1002/humu.21517

Acknowledgements

We would like to thank Drs. William A. Gahl, MD, PhD, David Adams, MD, PhD, Thomas Markello, MD, PhD, Murat Sincan, MD and Cornelius Boerkoel, MD, PhD, (Undiagnosed Diseases Program, National Institutes of Health) for providing the sample data and collaborating in the design of the tool.