Bioinformatic Analysis of Full Exome trio experiments

The full exome trio family analysis (father, mother and affected son or daughter) is a very powerful approach to identify potentially pathogenic 'de novo' mutations in the proband under exam. By sequencing the patient and his/her parents, variants can be filtered according to their consistency/inconsistency evaluated on the basis of the Mendelian Heredity laws and prioritized on the basis of technical parameters such as call quality, population frequency and relevance of the gene for the disease under study.

The global number of variants which can be identified in a whole exome 'trio' experiment can be estimated to be around 150.000 (mostly single nucleotide variants). The potentially pathogenic variants at the end of this procedure are, however, some tens, simplifying a lot the subsequent validation work. The trio exome analysis starts from the standard full exome analysis and continues with an analytical procedure developed 'ad hoc'.

Full Exome Analysis

The bioinformatic analysis of Full Exome experiments applied to trio is composed by two consequential steps: reference genome mapping and variation analysis, with the associated statistics.

Reference genome mapping

The analysis starts with the mapping of sequences generated by the sequencer on the reference genome of the most recent version. This step is performed with standard software, using parameters that are optimized for the Ion Torrent technology, in order to correctly manage the particular features of this sequencing technology (e.g. for short indels).

The global mapping and enrichment exome summary statistics are reported to the customer in the final report. Alignments in the binary ".bam" format (with the associated .bai indexes) are also given back to the customer, in order to allow the direct visualization of alignments on the reference genome with the software Integrated Genome Viewer (.bam files and associated .bai indexes: http://www.broadinstitute.org/igv/).

In this step we also generate a table for each analysed sample reporting the metrics of minimal, maximum and mean sequence coverage for every capture target identified by the genome coordinates and associated HUGO and RefSeq gene IDs.

Variation analysis and associated statistics

The second part of the analysis is performed using standard procedures and parameters optimized for the Ion Torrent technology, and consists in the analytical identification ('call') of sequence polymorphisms and of small insertions and deletions (minor or equal to 20 nt), always with respect to the reference genome. All variants are then annotated by comparison with standard databases such as dbSNP, Cosmic and Clinvar.

The variations identified in this step, both SNPs and INDELs, are reported for each sample in a result file in the standard ".vcf" format (Variant Call Format). These are given to the customer and can be analysed in autonomy with a range of different analytical tools, which can also be web-based. The customer also receives files in Excel format containing the variant annotation for all the evaluated samples. These annotations include sequencing depth (total sequence coverage) in the boundary of the position under exam; the classification in homo- or heterozygotes; the gene associated with the mutation and its genome localization; the count of the reads including the nucleotide corresponding to the variant and of those corresponding to the reference genome; the associated quality values; the relative frequency of the variant allele; the predicted effect of the variant on the amino acid sequence.

The analysis of small insertions and deletions (INDELs) in the range of 20 nucleotides represents a very important and unexplored resource for the identification of potentially causative variations. The result tables of these variations also report the allele calls, the sequence in the context of insertion and deletion and the evaluation of homozygosity/heterozygosity/hemizygosity of the identified variants.

Trio analysis and identification of putative pathogenic variants

The main points which can drive in variant prioritization, and hence in their numeric variation, in the trio exome analyses are:

⇨ The family structure (the pedigree is always considered in the analysis).

- ⇨ The selection criteria of always including variants with a functional effect on the gene product of the gene targeted by the mutation. Evidences such as mutations in the promoter and in the splicing signals are not prioritized in a first instance.
- ⇨ The extensive usage of quality control criteria such as the call quality and the coverage depth.

The exome 'in trio' has the big advantage of the easy reconstruction of the variant phases using the pedigree information. The three kinds of mutations which are interesting as potentially pathogenic in the proband (hence in the affected son or daughter of the family in exam) are:

- ⇨ De Novo Variants;
- ⇨ Compounded Heterozygote Variants;
- ⇨ Recessive Homozygote Variants.

The *De Novo Variants (neomutations)* are mutations not shared with none of the two parents. The main features of this kind of mutations are:

- ⇨ Correspond to a so-called "Mendelian Error", hence they do not respect the dynamic of monogenic variations since they are not inherited;
- ⇨ They can be originated by DNA replication errors, spontaneous genetic lesions, transposable genetic elements and so on;
- ⇨ Occur in early steps of human development;
- ⇨ They can be unique for each patient;
- ⇨ They are rare variants, with a MAF < 1% in comparison with the data derived from the 1000 Genome Project sequencing results;
- ⇨ Have a clear and strong effect on the gene product;
- ⇨ The proband (son/daughter) is frequently heterozygous (hemizygous);
- ⇨ The quality measures that can be used for selection/prioritization are the allelic depth for the reference and alternate allele (number of WT alleles/number of ALT alleles); the sequencing depth for the given polymorphic site; high call quality values.

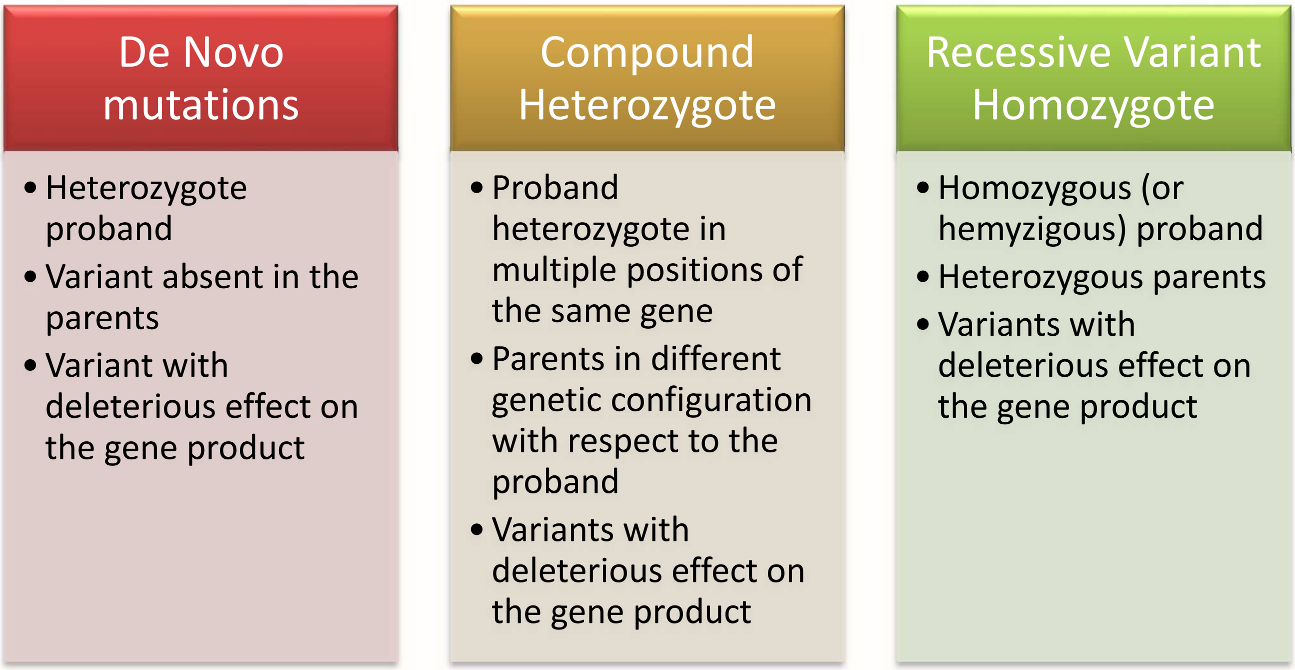
The second class of polymorphisms considered in this analysis is the *compound heterozygote* of the proband. The filters to be used in this case are:

- ⇨ Both parents must be heterozygous for different polymorphisms in the same gene;
- ⇨ The proband (son/daughter) must be heterozygous for (at least) two SNPs in the same gene, one inherited from each parent;
- ⇨ The polymorphism must alter the amino acidic sequence of the gene product;
- ⇨ They must be rare variants, with MAF < 1% with respect to the exome sequence data included in dbSNP and associated with the 1000 Exomes Sequencing Project;
- ⇨ Additional filters on the depth of the genotype call, on the quality score, on the heterozygosity index of the alternative allele are applied.

The third class of polymorphisms to be considered in this analysis are *rare variants in homozygosis***.** The criteria for the filters associated with this condition are:

- ⇨ The parents must be heterozygous for the same polymorphism;
- ⇨ The proband must be homozygous;
- ⇨ They are rare variants, with MAF < 1% with reference to the exomes belonging to the exome sequence data included in dbSNP and associated with the 1000 Exomes Sequencing Project;
- ⇨ The polymorphism must alter the amino acidic sequence of the gene product;
- ⇨ Additional filters on the depth of the genotype call, on the quality score, on the heterozygosity index of the alternative allele in the parents are applied.

Figure 1 summarizes the most relevant features for these three categories of mutations



De Novo mutations
- Heterozygote proband
- Variant absent in the parents
- Variant with deleterious effect on the gene product

Compound Heterozygote
- Proband heterozygote in multiple positions of the same gene
- Parents in different genetic configuration with respect to the proband
- Variants with deleterious effect on the gene product

Recessive Variant Homozygote
- Homozygous (or hemyzigous) proband
- Heterozygous parents
- Variants with deleterious effect on the gene product

Figure 1 – categories of mutations which can be identified with a trio analysis

The variants produced by the first part of the Full Exome analysis procedure are then subjected to a series of comparisons and specific filters, through standard procedures for Ion Torrent sequence data, for the identification of de novo mutations; of the variants in compound heterozygosis; of the recessive variants in homozygosis, according to the conceptual schema illustrated above. The results are summarized in Excel tables, and also the complete lists of all the globally identified variations are included in the results for the researcher.

The results of the complete procedure are reported, specifically highlighting the variants identified in the three categories: De Novo; Compound Heterozygotes; Recessive Homozygote. The variants are then provided with the following annotations:

- Variation type (SNV/Indel);
- Category (De Novo; Compound Heterozygote; Recessive Homozygote);
- The variant allele and the read assigned to the variant allele and reference allele;
- The ID and Gene Name;
- Annotations related to standard databases such as dbSNP, COSMIC, Clinvar
- The consequences on the gene product of the variant: position on cDNA, on CDS, in the protein; the reference amino acids and the variants with associated codons;
- The predicted effect of the variant on protein functionality, according to SIFT and Polyphen;
- The status of the proband, of the mother and of the father for the variation under exam (Homozygote; Heterozygote; Corresponding to the Reference Genome);
- Prioritization of the variant according to the relevance for the disease, also through candidate gene lists provided by the customer.

Ordering information

Item	Catalog N.
Bioinformatic Analysis III: DNA (full exome in trio)	DNA-BF03