Application Note

Analysis and comparison of samples from patients with inherited hearing loss

Application note on disease-associated variant detection using CLC Genomics Workbench

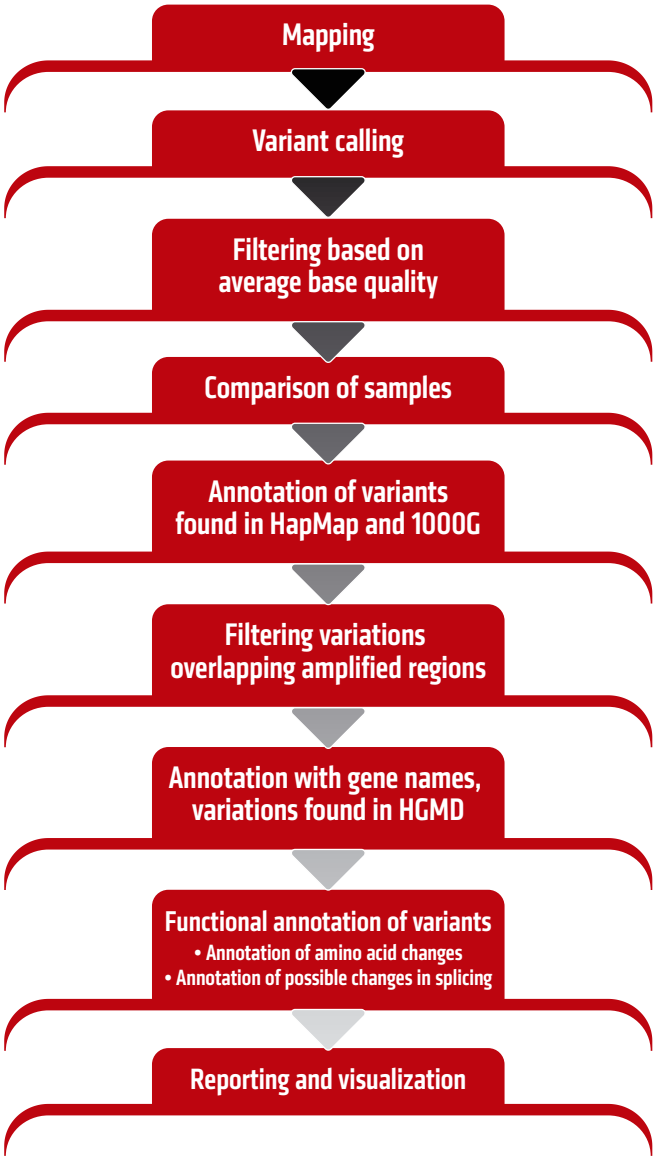
CLC bio

Analysis and comparison of samples from patients with inherited hearing loss

CLC Genomics Workbench is able to identify variants associated with Mendelian diseases, such as deafness, in a fast, easy, and accurate way. This application note provides some guidelines for analysis of targeted sequencing data and comparison of variants identified in a group of samples sharing the same Mendelian disease. The goal is to identify variants implicated in Mendelian hearing loss.

Data

We use 12 different samples from Jewish Israeli and Palestinian Arab probands with inherited hearing loss.



- Illumina GAIIx reads available at EBI Sequence Read Archive (ENA) with accession number ERP000823#

Exonic, untranslated regions and 40 flanking nucleotides in the introns of 246 genes, in total, are amplified using the Agilent Sure Select Targeted Enrichment Kit.

Analysis Workflow

Downloading and preparing human reference data

To prepare the human reference data for read mapping (human hg19), we download the sequence, gene annotations and variants from dbSNP, COSMIC, HapMap and 1000 Genomes Project. Please note that this can take some time due to data quantity.

Reads incorrectly mapped outside targeted regions can lead to false positives during variant calling. Incorrect mappings may be the result of sequencing errors, similar regions in the reference genome, or non-specific amplification. In order to filter out these artifacts, we import a .gff or .bed file with targeted regions using the "Import tracks from file" feature.

To show some of the newly created tracks in a combined view, we create a track list with the following tracks:

- Homo_sapiens.GRCh37.66.dna.toplevel (sequence)
- Homo_sapiens.GRCh37.66.gtf.gz_CDS
- Homo_sapiens.GRCh37.66.gtf.gz_Gene
- Targeted regions

You can also download HGMD annotations and use the "Annotate with BIOBASE Genome Trax" tool to annotate variants. Please note that access to BIOBASE Genome Trax is dependent on a current subscription to BIOBASE Genome Trax.

Once the track list has been created you can add or remove tracks by drag-and-drop. You can also rearrange the order of the tracks within the track list.

Map sequence reads

Sequencing reads are mapped to the reference genome (hg19) using the "Read Mapper". The sequencing reads of our samples are mapped separately to the human reference sequence using specific parameter settings (length

fraction: 0.8, similarity: 0.8). The newly created tracks are added to the open track list.

Variant calling

The "Probabilistic Variant Detection" is able to identify SNVs as well as small* insertions and deletions in a read mapping.

We run the variant detection[1] and use the "Filter Marginal Variant Calls" tool to remove variants with a low average quality (<19) from our called variant sets. The filtered variation tracks are dragged-and-dropped into the open track list.

*The maximum size of these will depend on the chosen mapper and the length of the mapped sequence reads

[1]Minimum coverage: 5
Variant present in forward and reverse reads: No
Maximum expected variations: 2
Variant Probability: 90.0
454/Ion Torrent correction: No
Use only specific matches: Yes

Filtering, annotating and comparing the called variants

We compare the variant calls from our samples using the "Compare Variants within Group" tool. For each variant allele, this counts and lists the samples in which the allele is found. We run the tool with a frequency threshold of 0% in order to get all variant alleles present in all samples.

We are not interested in common variants, but as some of them may have impact on inherited hearing loss, we use the "Annotate from Variant Database" tool to annotate the comparison result with annotations from HapMap and the 1000 Genomes Project.

Gene name	Found in samples	Allele	Amino acid change	Associated disease	Known in 1000G or HapMap?
DFNB59	49, 52	C/G	Arg265Gly	Progressive hearing loss, autosomal recessive	-
DFNB59	51, 56	C/T	Arg265Cys	Progressive hearing loss, autosomal recessive	-
WFS1	57	G/A	Glu864Lys	Optic atrophy, autosomal dominant, with hearing impairment	-
TMC1	53	T/C	Ser647Pro	Hearing loss	-
TMC1	53	C/T	Arg604*	Deafness	-
MYO15A	55	G/A	Glu1414Lys	Hearing loss	-
MYO15A	51	G/A	Arg2728His	Hearing loss	-
TRIOBP	50	C/T	Gly581*	Deafness, non-syndromic	-
CDH23	56	G/T	Val2635Phe	Deafness, autosomal recessive	-
CDH23	60	G/A	Ala366Thr	Usher syndrom 1d	-
TECTA	59	C/T	Thr1866Met	Deafness, autosomal dominant	-
GJB2	57	C/T	Val153Ile	Deafness	-
GJB3	60	C/T	Arg32Try	Deafness, non-syndromic, autosomal recessive	In 1000G
MYO7A	54	C/T	Thr1566Met	Usher syndrom 1b	In 1000G
MYO7A	59	A/G	Thr1719Cys	Usher syndrom 1b	In 1000G

Table 1: A selection of identified variants to be found in the HGMD database

In the subsequent steps, we annotate our candidate variants in order to prioritize them based on possible functional consequences and known Mendelian disease mutations.

Filtering and annotation of the combined variant calls can be executed as an automatic workflow using the following refiners:

- "Annotate from Overlapping Annotations" annotates each variant with the names of overlapping genes
- "Annotate from BIOBASE Genome Trax" annotates known variants with information from the HGMD database
- "Amino Acid Changes" investigates amino acid changes
- "Detect Splice effects" identifies possible splicing effects
- "Filter against Overlapping Annotations" filters variant alleles outside targeted regions

The resulting tracks can be added to the open track list.

Now, all variants are annotated with overlapping gene names, amino acid changes, possible splice effects, links to variant databases (as in our example - the HGMD and dbSNP databases), and information on dataset (origin) for the variants (Figure 1). Furthermore, it is annotated if the variant has been found in HapMap or in 1000 Genomes Project data.

Results

Variant analysis

A total of 142,988 variant alleles are found in all samples combined. As we are only interested in those overlapping targeted regions, this number is reduced to 3,710 alleles. Of these, 2,174 are present in the 1000 Genome Project and HapMap data which leaves us with 1,536 alleles likely to be rare variants.

Candidate variants

Further evaluation of the candidate variants shows that six known variant alleles associated with inherited deafness, as mentioned in the paper of *Bernstein et al.*[2] have been called.

Furthermore, one variant is associated with the Usher syndrom 1d, a genetic disease which is known to cause deafblindness. By not filtering out variants

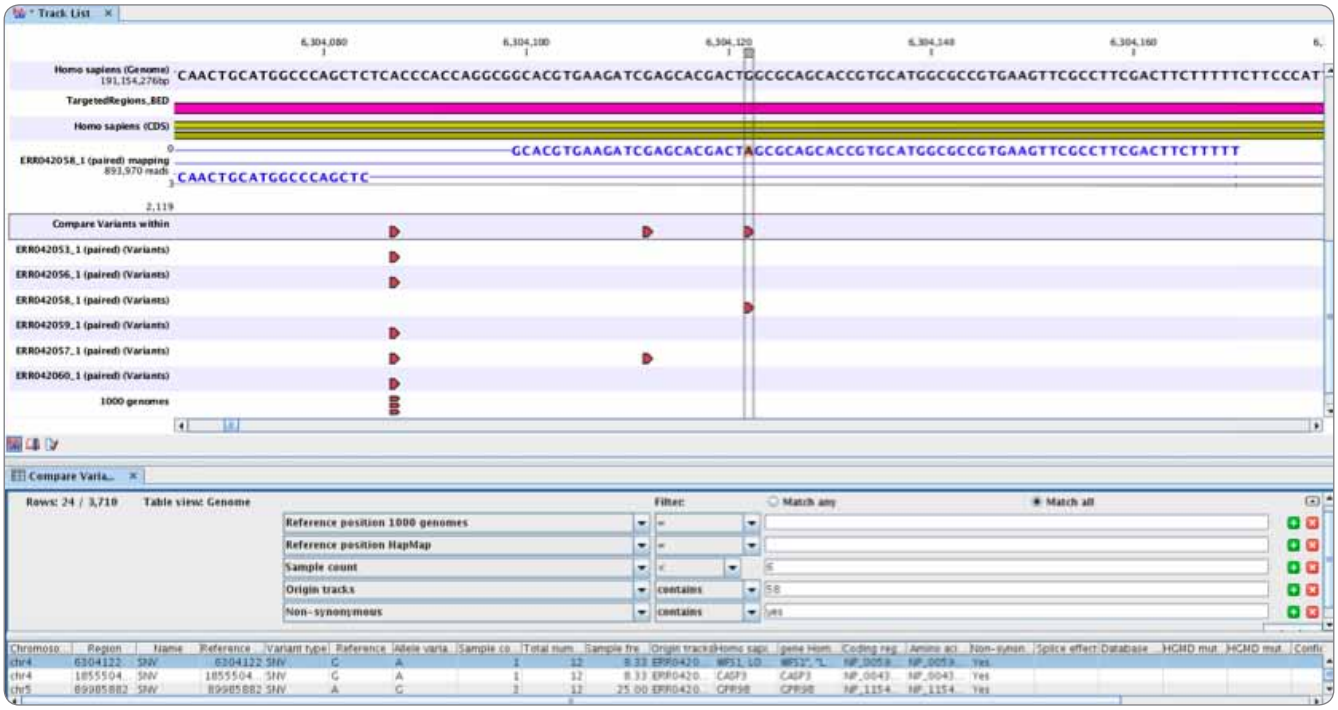


Figure 1: Filtering annotated comparison results to find candidate variants for sample 58

Gene name	Allele	Amino acid change
USH2A	C/T	Gly4838Arg
DFNB31	C/T	Arg244His and Arg627His
WFS1	G/A	Trp867*

Table 2: Potential unknown deafness-associated variants in sample 58

known from the 1000 Genomes Project, another two variants associated with Usher syndrom, and one variant associated with deafness are found (Table 1). Therefore, perhaps variants listed in this database should not be filtered out when identifying variants associated with a Mendelian disease.

Variants known to be associated with deafness can be identified in all samples, except for sample 58. However, opening the variant track in a tabular view and filtering for non-synonymous variants identified in sample 58, but not known in the 1000 Genomes Project or HapMap data (Figure 1), leaves 52 potential variants. Filtering out variants found in more than six samples, assuming that most of these are common variants in the Middle Eastern population, leaves us with 24 candidate variant alleles for this sample.

In summary, the candidate set of variants includes SNVs, MNVs, deletions, and insertions. A number of the variants are non-synonymous, some are known in the HGMD database and some are known to be directly or indirectly associated with inherited hearing loss.

[2] *Brownstein et al.* (2011): Targeted genomic capture and massively parallel sequencing to identify genes for hereditary hearing loss in middle eastern families. Genome Biology 12:R89.

CLC bio · EMEA
Finlandsgade 10-12
Katrinebjerg · DK-8200 Aarhus N
Denmark
Phone: +45 7022 5509

CLC bio · Americas
10 Rogers St # 101
Cambridge · MA 02142
USA
Phone: +1 (617) 945 0178

CLC bio · AsiaPac
69 · Lane 77 · Xin Ai Road · 7th fl.
Neihu District · Taipei · Taiwan 114
Taiwan
Phone: +886 2 2790 0799