NEW DATA

Application Note

Whole genome trio analysis

Application note on analyzing trio-based next-generation sequencing data using CLC Genomics Workbench

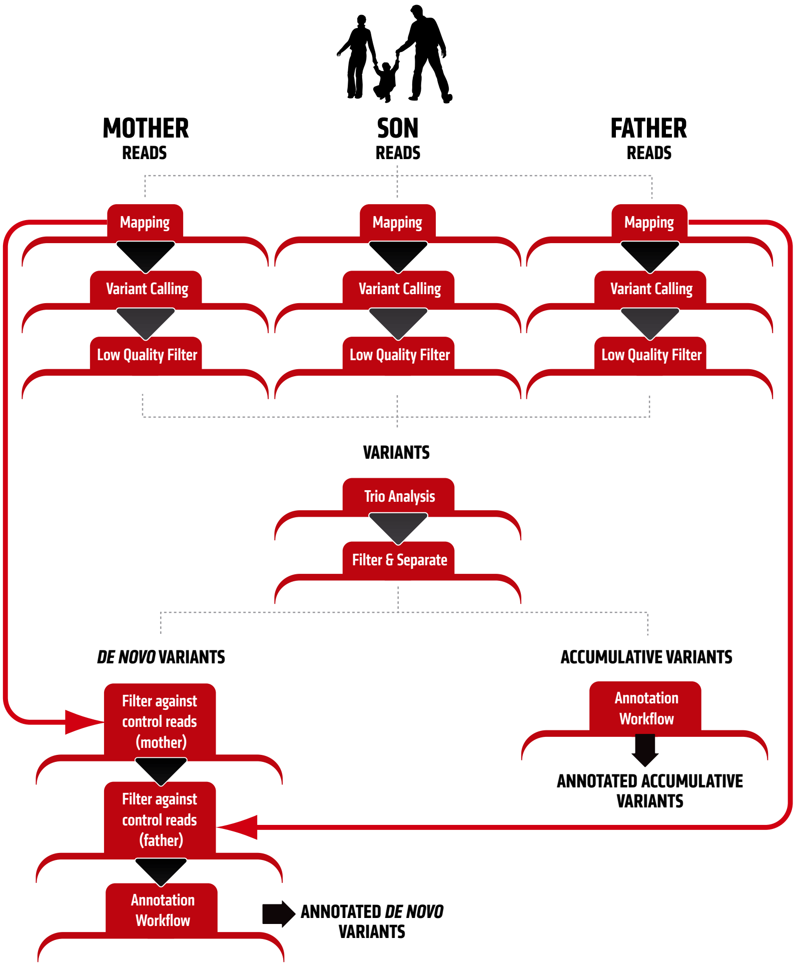
Whole genome trio analysis

This application note demonstrates how to analyze and compare data from a whole genome trio analysis using CLC Genomics Workbench. The goal is to identify inherited traits that have the potential to cause diseases in the offspring of two unaffected parents. To illustrate this, we use data from the Illumina Platinum CEU trio, which describes a mother, father and son of European ancestry.

Data

The Illumina Platinum CEU (Utah residents (CEPH) with Northern and Western European ancestry) trio contains whole genome sequencing results from a family composed of a mother (NA12878), father (NA12877) and son (NA12882). The data used has been generated on the Illumina 2000 platform and is accessed from the Illumina Platinum website (http://www.illumina.com/platinumgenomes/).

The members of the family sequenced are not known to exhibit any mendelian diseases, and thus no specific disease variations are anticipated to arise during the analysis. However, because disease and non-disease traits may be passed on through the same mechanisms, we can perform the same analysis to identify traits that do not confer disease status.



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 HapMap and 1000 Genomes Project. Furthermore, we download PhastCons conservation scores from UCSC. We also use the “Annotate from BIOBASE Genome Trax” tool to create a track for variants from the HGMD, dbNSFP, OMIM and disease relations database. Please note that access to BIOBASE Genome Trax is dependent on a current subscription to BIOBASE Genome Trax and installation of the plugin for CLC Genomics Workbench.

Please note that the downloads can take some time, due to the very large data volume.

Map sequence reads

Sequencing reads are mapped to the reference genome (hg19) using the “Mapping Reads to Reference”. The sequencing reads of our samples are mapped separately to the human reference sequence using default settings. The newly created tracks are added to the open track list by drag-and-drop.

DEFAULT SETTINGS

Mismatch cost: 2, insertion cost: 3, deletion cost: 3, length fraction:0.5, similarity fraction:0.8, non-specific match handling: map randomly.

Variant calling

We use the “Probabilistic Variant Detection”[1] to identify SNVs and small insertions and deletions in the mapped sequencing reads for each of the individuals in the trio.

- [1] Minimum coverage: 10
- 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
- Ignore broken pairs = No

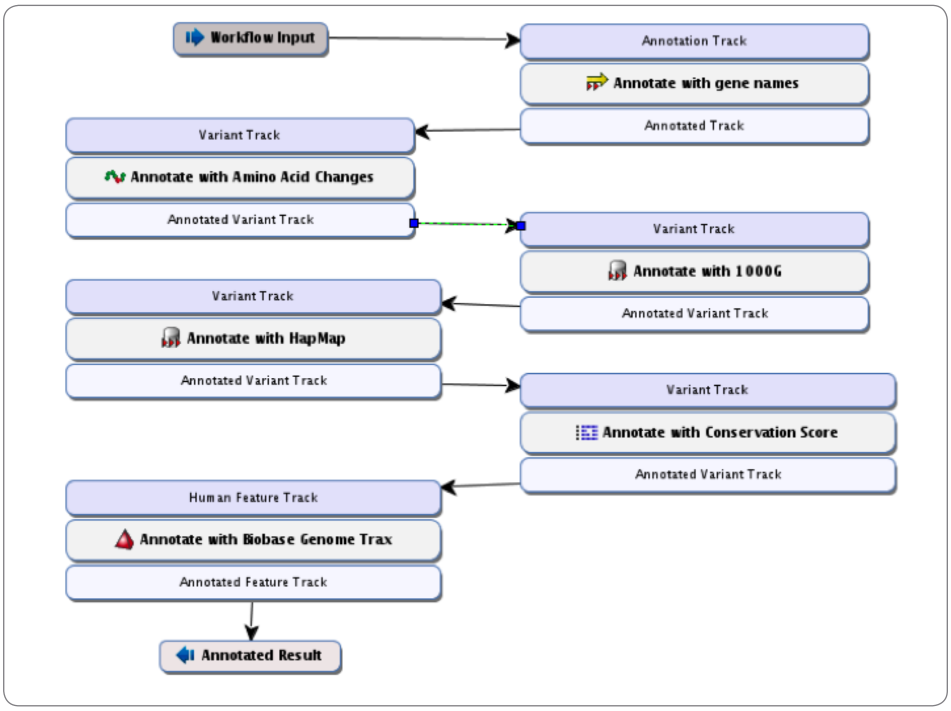


Figure 1: Annotation workflow

Trio analysis

The trio analysis step can be used to identify variations in the offspring that are present in one or both parents, yielding clues to the inheritance of variations that may have potential health consequences. For instance, diseases caused by inheriting a recessive variation from both parents (termed "accumulative" variations in CLC Genomics Workbench) can deprive the offspring from producing functional proteins from the affected genes. Additionally, *de novo* variations, which arise spontaneously in the offspring but are present in neither parent can play a role in non-hereditary diseases.

For the trio analysis performed here, the mapped and filtered data sets of the child and the two parents are used as input.

Separate *de novo* variations from accumulative variations

Once the trio analysis is completed, we gather the variations that are of interest for further analysis. Using the completed trio analysis, we filter on "Inheritance" and apply a filter containing "*de novo*". All the returned records are selected and a new track is created. We then change the filter on the original data set to filter on "Inheritance" = "Accumulative". Once again, we create a new track from the result.

Filtering *de novo* variations

Variations of the type "*de novo*" are called when the variations are not present in either parent. However, since we use a variation call to assess variations in the parent, it is possible that some variations were of poor quality or low sequencing depth to confidently be called a variation. Therefore, if there is any evidence that the variation is already present in a parent, we would like to discard this from the list of *de novo* variants.

Therefore, we use the "Filter Against Control Reads" tool to remove variations from the *de novo* results. This is done in two steps. First, using the *de novo* results and using the mapped reads from the first parent as the control set. Then using this new data set as input and using the second parent as the control set. This process removes the false positive *de novo* variations.

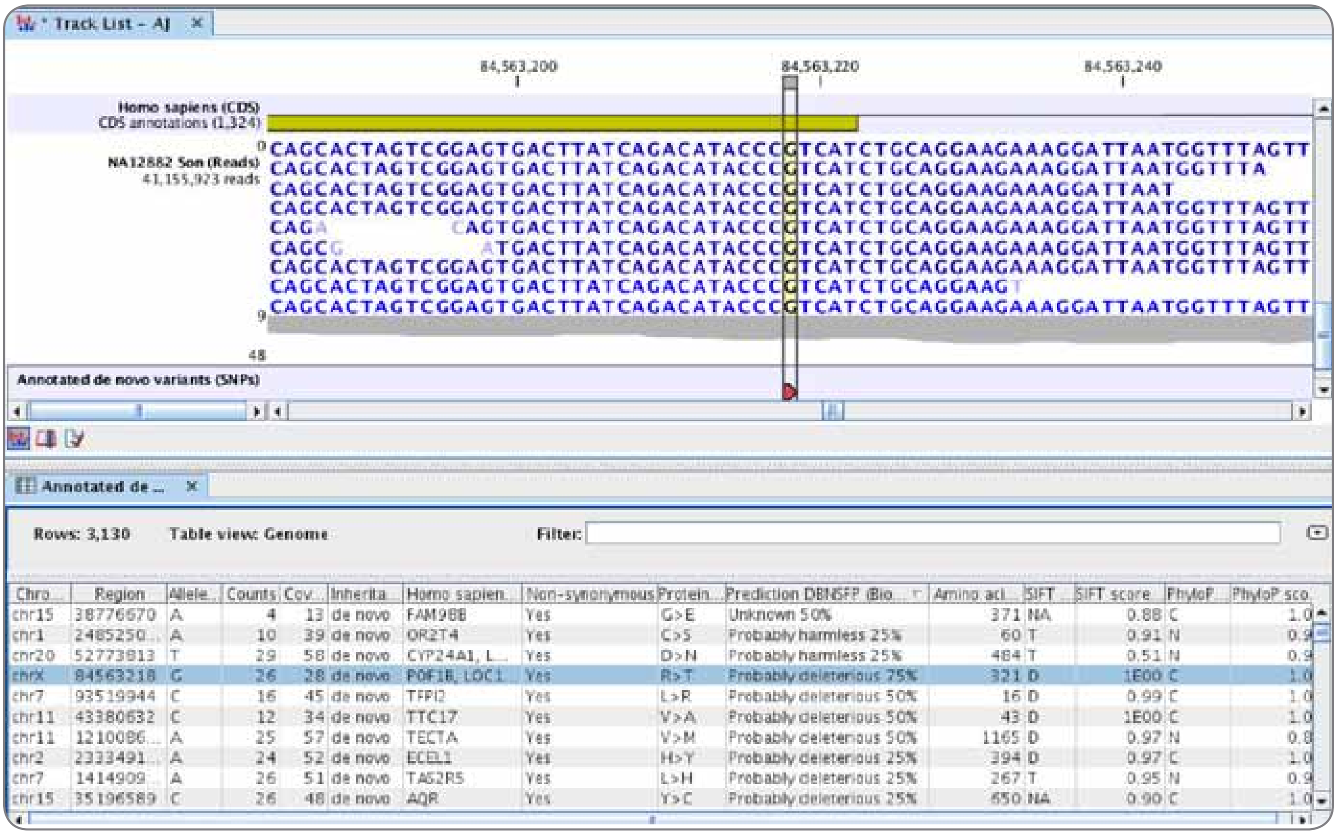
Filtering, annotation and comparison of the called variants

We annotate the combined variant calls. Filtering and annotation can be executed as an automatic workflow. In this case, the workflow used provides only the annotation step, allowing the filtering to be done manually.

Once the filtering is complete, the variants can now be prioritized or filtered by effects on splice sites, protein coding sequences, genes affected or other information provided by the annotation workflow.

Position	Gene	Variant	Associated disease (OMIM)	Biobase disease Annotation	Comments
Chr11:121008681	TECTA	G-A	Deafness, autosomal dominant		
ChrX:84563218	POF1B	C-G	Premature ovarian failure		
Chr2:233349186	ECEL1	G-A		Neuroblastoma	
Chr15:35196589	AQR	T-C			Intron-binding spliceosomal protein required to link pre-mRNA splicing and snoRNP (small nucleolarribonucleoprotein) biogenesis
Chr1:216158393	USH2A	A-G	Usher syndrome		Intron region, but well conserved
Chr7:141490961	TAS2R5	T-A			Taste receptor, well conserved
Chr11:43380632	TTC17	T-C			Well conserved
Chr11:76878005	USH1B		Usher syndrome		Intron region, but well conserved

Table 1: *De novo* variants with potential functional effects



```
CAGCACTAGTCGGAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAATGGTTTAGTT
CAGCACTAGTCGGAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAATGGTTTA
CAGCACTAGTCGGAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAAT
CAGCACTAGTCGGAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAATGGTTTAGTT
CAGA           CAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAATGGTTTAGTT
CAGCG             ATGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAATGGTTTAGTT
CAGCACTAGTCGGAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAATGGTTTAGTT
CAGCACTAGTCGGAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGT
CAGCACTAGTCGGAGTGACTTATCAGACATACCCGTCATCTGCAGGAAGAAAGGATTAATGGTTTAGTT
```

Rows: 3,130 Table view: Genome Filter:

Chro...	Region	Allele...	Counts	Cov...	Inherita...	Homo sapien...	Non-synonymous	Protein...	Prediction DBNSFP (Bio...	Amino aci...	SIFT...	SIFT score	PhyloP...	PhyloP sco...
chr15	38776670	A	4	13	de novo	FAM98B	Yes	G>E	Unknown 50%	371	NA	0.88	C	1.0
chr1	2485250...	A	10	39	de novo	OR2T4	Yes	C>S	Probably harmless 25%	60	T	0.91	N	0.9
chr20	52773813	T	29	58	de novo	CYP24A1, L...	Yes	D>N	Probably harmless 25%	484	T	0.51	N	0.9
chrX	84563218	G	26	28	de novo	POF1B, LOC1...	Yes	R>T	Probably deleterious 75%	321	D	1E00	C	1.0
chr7	93519944	C	16	45	de novo	TFPI2	Yes	L>R	Probably deleterious 50%	16	D	0.99	C	1.0
chr11	43380632	C	12	34	de novo	TTC17	Yes	V>A	Probably deleterious 50%	43	D	1E00	C	1.0
chr11	1210086...	A	25	57	de novo	TECTA	Yes	V>M	Probably deleterious 50%	1165	D	0.97	N	0.8
chr2	2333491...	A	24	52	de novo	ECEL1	Yes	H>Y	Probably deleterious 25%	394	D	0.97	C	1.0
chr7	1414909...	A	26	51	de novo	TAS2R5	Yes	L>H	Probably deleterious 25%	267	T	0.95	N	0.9
chr15	35196589	C	26	48	de novo	AQR	Yes	Y>C	Probably deleterious 25%	650	NA	0.90	C	1.0

Figure 2: Non-synonymous *de novo* variant in the POF1B gene, which is annotated in the DBNSFP database as probably deleterious

Results

De novo Variants

A total of 3130 variations SNVs that are found in the child are not detected in either parent. 21 SNVs are overlapping coding regions and according to dbNSFP 7 of these have been predicted as deleterious, 2 as harmless and 1 as unknown. 58 (52 in non-coding regions) variants have a conservation score above 0.9, which is a hint that these variants are likely overlapping regulatory sites.

"Accumulative" variants

In this dataset no accumulative variants could be detected. This is expected and shows the great data quality.

Variant analysis

Results are summarized in table 1.

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