Application Note

Identification and comparison of BRCA1 and BRCA2 variants

Application note on analyzing amplicon sequencing data and identifying cancer relevant variants using CLC Genomics Workbench

CLC bio

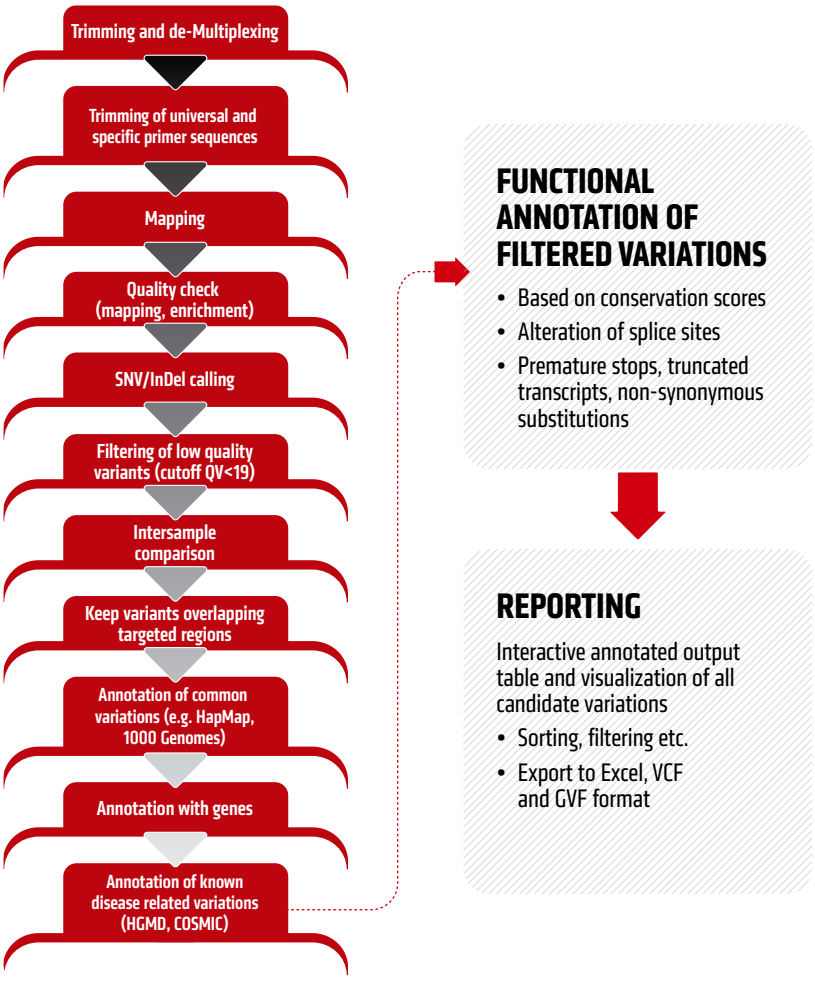
Identification and comparison of BRCA1 and BRCA2 variants

Combined use of Mutiplicom's BRCA MASTR assay with Ion Torrent's sequencing technology and data analysis using CLC Genomics Workbench results in high sensitive and specific detection and annotation of cancer relevant mutations in BRCA1 and BRCA2. This application note shows how to analyze amplicon sequencing data and identify cancer relevant variants in these two highly-studied tumor-suppressor genes using CLC Genomics Workbench.

Data

- Ion Torrent data in sff format that includes multiplexed sequencing reads from 14 DNA samples
- GFF file with targeted region coordinates according to Hg19[1]
- Excel sheet with barcodes and the corresponding names of the samples[2]
- Universal Primer sequences[3]
- Specific PCR primer sequences[4]

All BRCA1 and BRCA2 coding exons are amplified with Multiplicom's BRCA MASTR v2.1 assay from 11 DNA samples and 3 control DNA samples. Shearing, barcoding and sequencing is performed according to the Ion Torrent protocols.*



[1] available on request from customerservice@multiplicom.com
[2] Ion XpressTM Barcode Adapters kits 4471250, 4474009
[3] Multiplicom
[4] Multiplicom

*Performed at LGTC, Leiden, The Netherlands by H. Buermans and his team.

Analysis workflow

Genomic DNA is extracted from blood samples[5]. All coding regions of BRCA1 and BRCA2 are amplified and universal tag sequences are incorporated with the BRCA MASTR v2.1 Assay according to manufacturer's protocol[6]. After a universal PCR (Short Read Amplification kit, Multiplicom), the resulting amplicon libraries are pooled per individual. Libraries are fragmented by enzymatic shearing[7] and purified[8]. Subsequently, barcode adaptors are ligated[9] and the different barcode ligated libraries are pooled. Size selection for fragments of 100bp on gel is performed[10], and amplified for 5 cycles. Libraries are subsequently processed according to the manufacturer's protocol using the Ion OneTouch™ System Template kit (100 bp) and sequenced on the Ion Torrent™ Personal Genome Machine™ (PGM™) using an Ion 316™ chip and the Ion Sequencing Kit (100 bp chemistry).

[5] QiAamp DNA Blood kit, Qiagen
[6] Multiplicom
[7] Ion XpressTM Plus Fragment Library Kit, Life Technologies
[8] AMPure® XP, Agencourt
[9] Ion XpressTM Barcode Adaptors 1-16 Kit, Life Technologies
[10] MinElute Gel Extraction kit, Qiagen

Download and prepare human reference data

Reference data can be prepared using the "Download Genome" tool. In this example, we select Homo sapiens (hg19) and download the sequence, gene annotations and variants from dbSNP, COSMIC, HapMap and 1000 Genomes Project. Please note that the download may take a while due to the volume of the data.

Incorrectly mapped reads outside the targeted regions can lead to false positives during variant calling. Reads may be mapped incorrectly due to sequencing errors, similar regions in the reference genome or even non-specific amplification.

To have the option to filter out these artifacts at some point during the analysis, targeted regions should be annotated. For this purpose, we import a file with targeted regions.

Tracks can be viewed together by creating a track list. We can create a track list of 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

Tracks can be removed, added and rearranged using drag-and-drop.

Figure 1: Four-base deletion in BRCA2 found in barcode 6 sample known to be associated with breast cancer (HGMD).

Name	Sequence
Universal primer Tag 1	AAGACTCGGCAGCATCTCCA
Universal primer Tag2:	GCGATCGTCACTGTTCTCCA
Target-specific primers	Available on request by email at customerservice@multiplicom.com

Table 1: Subset of some of the identified unknown somatic variants

Figure 2: Composition of the adaptor and barcode sequences as part of the Ion Express library preparation kit.

Trimming of reads & de-multiplexing

To enable a better mapping of reads to the human reference genome, all reads should be trimmed prior to mapping. For this approach we import all relevant adapters and primers into the workbench.

The two library-specific adapter sequences flank the sample-specific barcodes. These adaptors are designated as "linkers" during the de-multiplexing step, thus they are automatically trimmed during the parsing process.

We use the "Process Tagged Sequences" tool together with the file including the corresponding barcodes and the name of the samples to de-multiplex the reads.

The results are 14 separate folders named according to each sample and including all reads identified by the specified barcode. Next, the reads in each folder are trimmed for the Universal Tag primers and the target-specific primers.

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 (insertion costs=2, deletion costs=2, length fraction=0.5, similarity=0.8). The mapping is run in batch mode in order to analyze all samples simultaneously. Each of the newly created tracks should be added to the open track list.

Enrichment

The "Targeted Regions Statistics" tool is used for statistics about enrichment specificity and coverage.

Variant calling

We run the Probabilistic Variant Detection[11] to identify SNVs as well as small insertions and deletions in the mapped sequencing reads of all samples. The resulting tracks are filtered with an average base quality of <19 and for comparison of the results with an average base quality of <16. Resulting tracks are dragged-and-dropped into the open track list.

[11]
- 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

Filtering, annotating and comparing the called variants

Variant calls from different samples can be compared using the "Compare Variants within Group" tool. This will, for each variant allele, count and list the samples in which the allele is found. We run this tool with a frequency threshold of 0% in order to get all variant alleles present in all samples.

Filtering and annotation can be executed as an automatic workflow. The workflow in our case uses the following refiners:

- "Filter against Overlapping Annotations" extracts all variant alleles overlapping targeted regions
- "Annotate from Variant Database" filters against variant databases (e.g. HapMap and 1000 Genomes Project) to identify candidate variants
- "Annotate from Overlapping Annotations" annotates the candidate variants and prioritizes them based on possible functional consequences and known mutations

Sample	# Reads after demultiplexing	Avg. length	# Reads after trimming	Avg. length	% Reads mapped	% Specificity	Avg. coverage	# Variants > QV19	% Concordance	# Variants >QV16	% Concordance
Barcode 1	127,360	142	95,460	116.5	89.7	99.6	339	43	0.97	49	1.0
Control 1	280,229	138	213,803	115	83.22	99.67	620	30	1.0	36	1.0
Control 2	162,518	137	122,559	115	83.6	99.7	360	15	1.0	21	1.0
Control 4	212,396	137	161,301	114.6	83.55	99.67	471.8	42	1.0	45	1.0

Table 2: Overview of summary statistics from selected samples[14]

- "Annotate from Variant Database" annotates known mutations with information from variant databases (e.g. COSMIC[12] and HGMD[13]- please note that a license is required to access the HGMD database)
- "Amino Acid Changes" investigates amino acid changes

Results can be added to the open track list.

[12] Forbes *et al.* (2011) Nucl. Acids Res. 39 (suppl 1): D945-D950
[13] Stensen *et al.* (2009) Genome Med. 1:13

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

Comparison with validated variants

Some of the variants from three control samples as well as one cancer sample have previously been validated by Sanger sequencing. Our variant calls are compared to these datasets to evaluate the sensitivity of the analysis.

Results

Enrichment

Approximately 90% of all trimmed reads can be mapped to the human reference genome (hg19). 99.6% of aligned reads partially or completely overlap the amplified regions and the average coverage is between 330 and 620. 99.9 - 100% of all bases in the targeted region have a minimum of 100x coverage.

Variant comparison

Omitting the homopolymer error correction and a quality cut-off of 19, we find 87 variant alleles overlap the targeted region in all samples together. 34 variants are known in the 1000 Genomes Project data and 31 variants are known in HapMap. Five variants are found in COSMIC and of these, three are also found in the 1000 Genomes Project data and could be cancer associated variants. Sixteen of the variants are present in HGMD, 14 of these associated with cancer. 42 of the variants (InDels, MNVs and SNVs) are not found in the databases above nor are they detected by the reference method.

These variants may be false positives but they also should be considered as candidates for further validation. However, nine of them partly overlap variants in the databases and there is a high probability that these are actually true positives.

The 454/Ion Torrent homopolymer error correction fails to detect another four-base deletion and an insertion associated with breast cancer found in HGMD and COSMIC.

Running a fisher-exact test to compare for each variant allele the significance that it is cancer specific using three control samples (Control 1, Control 2 and Control 4) and the remaining samples as cases, we get a return of no variant alleles with a cut-off > 0.05. However, 36 variant alleles are not found in any of the control samples and 16 of these are not known in any of the annotated databases.

Comparison with validated variants

Table 2 shows that CLC Genomics Workbench is able to achieve a 100% sensitivity by omitting the homopolymer error correction as well as using a quality threshold of 16. Using a quality threshold of 19 instead results in a single false negative but the number of variants called in homopolymer regions has increased.

[14]
- % Reads Mapped - # reads mapped/# reads total after trim
- Specificity - mapped reads on target vs. amplicon coordinates
- Avg. Coverage - average # of times a position or region is covered by at least one read
- % Concordance - # true positives detected/# true positives validated by Sanger sequencing

Acknowledgements:

We thank the following people for providing us with the data and collaboration on this project:

Giuseppe Giannini, Valeria Colicchia & Amelia Buffone from the Department of Experimental Medicine at the University La Sapienza, Rome, Italy

Dirk Goosens, Jurgen Del-Favero & Annelies Rothier from Multiplicom

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) 444 8765

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