454 Variant Detector:  
a New Blast-based Pipeline  
for Low Frequency INDEL Detections  
on 454 Amplicon Data

ZARA I (1), ALBIERO A (1, 2), SIMIONATI B (2), DAL PERO F (2), LOVISA F (3), BONVINI P (3, 4), ROSOLEN A (3), VALLE G (1)

(1) CRIBI University of Padua  
(2) BMR Genomics srl  
(3) Hematology and Oncology Clinic of Pediatrics, Azienda Ospedaliera University of Padua  
(4) Fondazione Città della Speranza, Vicenza

Motivation  
The Second-generation sequencers have opened new perspectives in biological and medical research. 454 Amplicons approach enables ultra-deep genetic analysis at levels of sensitivity not previously possible. Discover new somatic mutations can reveal the biological bases of cancer. The Amplicon approach allows to focus the analyses on limited genomic regions which are then sequenced with 454. These regions are generally key regions for cancer evolution and their variations can be at the base of transformation. The sequencing output is generally composed by hundreds of thousands 300 base long reads and the bioinformatic analyses can be very complex. Major part of dedicated software (Roche AVA Software, VIP, Mutation taster...) are able to fast analyse all the reads reporting SNPs and small INDELs but they often fail in large INDELs detections. SNPs analyses have to take into considerations two factors: the sequencing errors and the PCR errors. These two factors limit the variant identification on 0.5% of frequency. Below this limit it is not possible to distinguish between error and true SNP. The large INDELs on the contrary can not be treated like SNPs. PCR errors or sequencing errors can not introduce large deletions or large insertions; at the same time moreover, alignment algorithms are often not able to open large GAPs with these small reads. In this scenario, we developed a new blast based pipeline specifically designed for large INDELs detections that can help classical software in variants discovery.

Methods  
Our pipeline, named 454 Variant Detector, is a blast-based alignment where each read is aligned against the reference. Using blast, reads with large deletions or large insertions are generally splitted in 2 or more blocks. These blocks are then analysed to refine the alignments on the boundaries that are often not well defined. In this phase our algorithm is able to find the best solution performing a global alignment.
**Results**

454 Variant Detector was tested on a 454 Amplicon run performed in our lab. The amplicon was designed on a cancer involved gene, and sequenced using 1 lane of the PTP, giving 41,519 good quality reads. Mean reads length was 356 base pairs. The amplicon length was 356 base pairs and the reads were sequenced from A and from B. Large INDELs discovery was performed using different software (AVA, VIP, Mutation Taster, CLC) and the unique INDEL detected was a 130 base gap (a known alternative splicing) found by Roche AVA Software. 454 Variant Detector, on the contrary, found a large number of INDELs. We found 90 INDELs (74 deletions and 16 insertions). After a first filtering, based on frequencies (we selected all variants represented by at least 5 reads), we focused our attention on 23 INDELs (18 deletions and 5 Insertions) and we noted that some solutions could be merged in a unique variant. Major part of INDELs had a very low frequency (<0.01%) but their reliability has been confirmed by other data. A 84 base deletion represented by only 6 reads (freq << 0.01) was confirmed by cloning and sequencing with Sanger. Major parts of deletions was compatible with exon skipping whereas insertions were compatible with alternative splicing, showing typical donor and acceptor splicing sites (GT-AG) at the boundaries. Computation performance of 454 Variant Detector are quite good. It takes 40 min for analysing 40,000 reads on a desktop pc. For the future, 454 Variant Detector will be improved for SNPs detection and we are now working for reducing the computational time. We are also developing a pipeline for visualizing the output results and alignments on gbrowse, pointing up the alternative splicing and the effects of the found variations.

Contact email
alessandro.albiero@unipd.it