On the impact of short-reads quality on variants detection

S. Castellana✉, T. Mazza

IRCCS Casa Sollievo della Sofferenza - Mendel, Italy

Motivations

Next Generation Sequencing technologies have greatly improved our ability to detect genetic variations within organisms. In particular, they have been recently applied with great success to the discovery of mutations linked to rare mendelian diseases. The goodness of the results has been showed to be tightly dependent on two as crucial as systematic family of errors: those introduced by the particular sequencing chemistry and those due to the library preparation procedure. As a consequence, it is rather important to always assess the quality of the sequencing products before proceeding to carry out any downstream analysis, especially when errors, even if rare, might be decisive to the fulfillment of the overall objective. In the context of variants discovery, errors are deleterious. We have then implemented an accurate bioinformatics pipeline, which particularly applies to SOLiD raw data, and that aims to reduce the occurrences of false positive mutation calls. It consists of several sequential steps: SOLiD short reads quality assessment, dataset filtering, mapping, exome coverage analysis and variants discovery.

Methods

Four exon-enriched genomic libraries, belonging to 3 unrelated patients with a severe neurogenetic disease and one control sample, were sequenced by the SOLiD 4 sequencing platform (Applied Biosystems, part of Life Technology, Foster, USA). Site-specific and global reads quality were analyzed by the QV _ assessment tool [1] and by means of custom scripts. Then, because of the assessed qualities, color-space sequence files were filtered according to different threshold values. Afterwards, filtered and original sequence libraries were mapped, using the LifeScope Genome Analysis Software. Aligned reads were recalibrated through the GATK (The Genome Analisys ToolKit). Performances of mapping were calculated by custom scripts, while depth of coverage for each target site and target exon was calculated by means of the Bedtools Package and some R custom scripts. Variants were detected by the GATK and, later, cross-checked by Samtools and diBayes. Therefore, they were annotated by ANNOVAR. We looked for novel candidate mutations by searching our variants in the dbSNP (ver. 1.35) repository and setting a minimum allele frequency threshold to 0.02. For two out of the four libraries, we could check the resulting variants against those of as many SNP arrays. We could then assess the performance of our filtering strategies and do some statistics on the number of false/true positive variants call.

Results

The quality of these SOLiD 4 raw data has been found to be generally low. Accuracy of color calls decreased gradually along the 50 bp fragments length, and resulted to be under 20 already around the 20th-25th read position. As a confirmation, even the application of the most relaxed filtering strategy caused a dramatic reduction of mappable reads: indeed, the four raw datasets ranged from 50 to 80 Mb, but only the 35-65% of sequence data passed the filters. Reason for this seems to be linked to SOLiD 4 sequencing chemistry rather than to the library preparation procedure. For filtered and unfiltered libraries, the median site-specific coverage ranged from 12 to 32, with the 50%-80% of sites having a depth of coverage higher than 10. Regarding the entire set of target exons (over than 200000), from 4% to 10% of them were systematically skipped or poorly covered by the sequencing process. Regarding the variant calling task, a large number of low quality SNPs and Indels has been detected on the four raw datasets, while from 5000 to 15000 variations have been found among the filter-passing datasets (going from the more stringent to the more relaxed filtering criteria). Because of the recessive mode of inheritance of the mendelian diseases under examination, we searched for homozygous mutations in exons and splice sites, finding different pools of non-synonymous, stop-gain, stop loss and splicing mutations within each filtered library. In addition, low ratios of transition vs. transversion and non-synonymous vs. synonymous variants clearly

confirmed the general low sequencing quality of the four experiments. Comparison with SNP array data finally helped to determine the best filtering configuration and the variants set with minimal impact of false positive calls. As a side effect, the presented pipeline produced some detailed reports about the SOLiD read quality, mapping accuracy, coverage of target regions and variants detection. It implemented some well known best practices and custom methodologies and, even in case of bad sequencing outputs (as the above-mentioned cases), tried to extract reliable information and present them to biologists and clinicians.

References

1. Sasson A, Michael TP (2010) Filtering error from SOLiD Output, Bioinformatics 26, 849-850.