Analysis workflow for the identification of allelic variant associated with a complex disease using NGS approach

V. Maselli✉, D. Cittaro, E. Stupka

Center for Translational Genomics and Bioinformatics, San Raffaele Scientific Institute, Milan, Italy

Motivations

Recent advances in sequencing technologies allowed for unprecedented possibilities and applications for clinical data analysis. In particular, Whole Genome sequencing at decent coverage has turned into cost-effective technology to characterize the genetic framework of rare diseases. We here present an example of a complex disease. Our purpose is to identify allelic variant associated with the described syndrome using a Next-Generation Sequencing approach.

Methods

We performed a 100 bp paired-end sequencing of a single human genomic sample using a Illumina HiSeq 2000 sequencer. Read tags were aligned to the hg19 reference genome using BWA [1]. The Genome Analysis ToolKit was used to pipeline the downstream analysis: local realignment around indels, quality score recalibration, SNP and indel calling and, most important, Variant Quality Score Recalibration. We filtered SNV with a confidence lower than 0.1%. We used the Seattle SNP Annotation tools [2] in order to annotate the SNP on the reference genome. We performed some preliminary statistics in order to identify a threshold that allowed us to identify a reliable subset of SNPs to use for our purpose. Using the sub set of known SNPs as guide we identify the value of the SNP quality, for which we are confident regarding our data. In a first step we used a single lane data to validate the homozigosity analysis procedure. Using a so defined quality threshold of 50 we identified a subset of SNPs that we analysed with the HomozygosityMapper web tools [3]. Analysis on a double lane is work in progress.

Results

We sequenced almost 400 million of read pairs achieving a 20x average coverage. We filtered out 2 million of SNPs with a read depth of at least 5 and max 250. The average quality above 400k SNPs is 313 (max 9371, min 30). We found about 100k novel SNPs (10% homozygous). We had a prior knowledge about the pathology: in particular we are interested in large stretches of homozygous SNV calls (Genome-wide linkage analysis performed under the assumption of recessive inheritance identified a common homozygous haplotype for this condition.). Accordingly, we found 6 stretches of homozigosity, two of which on the same chromosomes (6 and 16) described in a previous study. We think that this workflow could be easily automatized and used for different type of re-sequencing projects and that would lead to a strong interaction between clinical and molecular data, which is the purpose of the translational genomics.

References

1. Li and Durbin (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26 (5), 589-95
2. Seattle SNP Annotation tools: http://snp.gs.washington.edu/SeattleSeqAnnotation134/
3. Seelow et al. (2009) HomozygosityMapper an interactive approach to homozygosity mapping. Nucleic Acids Res. 37 (suppl 2), W593-W599