Building an optimized pipeline for whole-exome sequencing

M. D'Antonio[1], P. D'Onorio De Meo[2], B .Elmi[1], N. Sanna[2], G. Pesole[3,4], T. Castrignanò[2]✉

[1]Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli Studi di Bari, Bari, Italy
[2]CASPUR, Consorzio interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy
[3]Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche, Università degli Studi di Bari, Bari, Italy
[4]Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Itlay

Motivations

Managing the huge amount of data produced by NGS platforms requires non trivial IT skills. Furthermore the wide list of freely available analytical tools for NGS data analysis makes difficult to choose easily the pipeline components. An additional layer of complexity is due to the need of integrate all steps in a single analysis: the best tool for a specific purpose could be incompatible with other tools of the pipeline, i.e. a tool performing a statistical calculation when raw data are required for the next step. In case of a whole exome analysis building effective pipelines that relate variants to their samples and controls, annotate them from multiple sources requires a large customization effort.

Methods

Our proposed pipeline walks through several steps to perform a full analysis. 1) Before mapping the short-reads against the reference genome, a pre-process is necessary. FastQ files should be checked for integrity and cleaned up from any unwanted symbols that can alter any NGS tool behavior. Quality checks can be pursued with tools like FastQC to ensure that sequences provided reach the minimum level of mean quality necessary for a complete analysis. 2) Alignment is usually performed with BWA [1], which is capable of finding gaps. It results to be a good compromise between speed and accuracy. When there are known problems in the sequence provided, e.g. FastQC outlines a poor quality in the last or first bases sequenced, other tools can perform a more sensible alignment at a lower speed. 3) BWA provides mapping results in SAM format [2]. This is the most widely used format for alignment output. This text-based format should be converted into its binary equivalent BAM format through the SAMtools; BAM can be indexed and sorted to enable faster operations at subsequent steps. 4) Before searching any variant in mapping binary data, some other editing are required to prevent artifacts in results. Quality recalibration is required to refine some oddness caused by sequencing and mapping on quality scores. Duplicates are in most of the case result of PCR amplification and should be avoided as they lead to false positives. A re-alignment around known indels position should be also carried on to delete other artifacts. 5) Single Nucleotide Polymorphism (SNP, a single nucleotide occurring in one member is replaced by another nucleotide in the other member) and Deletion-Insertion Polymorphism (DIP, refers to the fact that a short nucleotide sequence in one member is omitted in the other member) can be now called from the mapping data obtained from the previous 4 steps. 6) SNP and DIP obtained have various score to consider to ensure a minimum depth coverage and quality score in order to remove any false positive in the list. 7) When dealing with multiple WES data lanes, the usual scenario is a combination of affected/unaffected tissue samples. In this case a critical information is about the haplotype phasing, which allow discovering complex heterozygous or homologous mutations. 8) The last critical aspect of variants calling is to associate as many annotation as possible to the variant list i.e. annotation stored in database like dbSNP, 1000genomes, etc. After these steps data can be saved into custom databases to allow cross-linking and intersections, statistics and much more.

Results

We have tested different freely available algorithms used at the alignment and post alignment stage and integrated them with custom-build scripts to provide the most suitable and complete combination to create significant whole exome dataset results. Hence, we have customized the whole-exome data analysis pipeline to preferentially held true variants by minimizing the incidence of false positives and providing the benchmarks for the best choice of right analytical tools.

References

1. Li, H, Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60
2. Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G, Durbin, R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9