Analysis pipeline for the detection of mutations causative of rare diseases on whole exome sequencing data

Antonio Rueda[1], Francisco Javier López[1], Javier Pérez[1], Pablo Arce[1], Luis Miguel Cruz[1], José Carbonell[2], Jorge Jiménez-Almazán[2], Enrique Vidal[2], Guillermo Antiñolo[1,3], Joaquín Dopazo[1,2], Javier Santoyo[1] ✉

[1]Andalusian Human Genome Sequencing Centre (CASEGH), Medical Genome Project (MGP), Sevillia, Spain
[2]Institute of Computational Genomics, Principe Felipe Research Centre (CIPF), Valencia, Spain
[3]Unidad de gestión clínica de genética, reproducción y medicina fetal. Instituto de Biomedicina de Sevilla (IBIS), Hospital Universitario Virgen del Rocío-CSIC-University of Seville, Sevilla, Spain

Motivation and Objectives

Recent advances in high-throughput sequencing technologies have made exome sequencing to be an outstanding tool for finding disease associated mutations at a relatively low cost. However, it is a non-trivial task to transform the vast amount of sequence data into meaningful variants to improve disease understanding. Several challenges arise when dealing with this approach, being critical checkpoints the raw read preprocessing, mapping procedure, variant calling and posterior variant selection. A number of computational algorithms and pipelines have been reported for variant analysis (Kumar *et al.,* 2009; Lam *et al.*, 2012; Li *et al.*, 2012; San Lucas *et al.,* 2012; Yandell *et al.,* 2011; Wang *et al.*, 2010) although none of them provide a complete strategy from raw data to mendelian analysis results. Here, we present a methodology that spans from SOLiD raw reads processing to mendelian analysis and variant selection, and its application over a set of samples from The Medical Genome Project, which proves the good performance of the applied methodology.

Methods

The input of the pipeline is an xsq file generated by Applied Biosystem SOLiD 5500 XL sequencers, while the output is the result of variant annotation and mendelian analysis, assuming samples to be derived from a group or a family. A brief description of the steps is provided below:

1. Fasta and qual files generation from xsq files.
2. Duplicated reads removal.
3. BLAT-like Fast Accurate Search Tool v0.7.0a (BFAST) (Homer *et al.,* 2009) for read mapping.
4. BAM cleaning: duplicated alignments and mismatched reads removal.
5. BAM realignment and SNV calling using the Genome Analysis Toolkit v1.4.14 (GATK) (DePristo *et al.,* 2011)
6. Variant quality filter based on GATK Best Practices V3 and depth filter.
7. Annotate Variation package (ANNOVAR) for variant annotation (Wang *et al.,* 2010); SIFT (Kumar *et al.,* 2009), Polyphen (Adzhubei *et al.,* 2010), 1000 genomes frequency (The 1000 genomes project consortium, 2010) and dbSNP (Sherry *et al.,* 2001) for assessment of variant frequency.
8. Mendelian filter of deleterious variants.

Results and Discussion

The Medical Genome Project (MGP) aims to characterize a large number of rare genetically-based diseases. As a proof of concept, we selected from the MGP a set of affected individuals by several hereditary rare diseases, their healthy relatives and a set of 50 control healthy individuals from Spanish population. The full methodology was run and the results reveal a number of deleterious haplotypes in several genes which could be directly associated with the diseases.

The validation of some of the predicted variants by the pipeline shows the good performance of our methodology analysis. Critical aspects to achieve such good performance are (i) BAM filtering, since an excessive number of mismatches are allowed by BFAST for short reads; (ii) the selection of variant filters and quality thresholds as recommended by GATK Best Practices V3 in combination with a depth threshold allowing high quality calls and (iii) the inclusion of control individuals in the analysis, which is essential since they remove population variants which can disturb the interpretation of the final variant set.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A et al. (2010) A method and server for predicting damaging missense mutations. *Nature methods* **7**(4), 248-249. doi:10.1038/nmeth0410-248

DePristo M, Banks E, Poplin R, Garimella KV, Maguire JR et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics,* **43**, 491-498. doi: 10.1038/ng.806

Homer N, Merriman B, Nelson SF (2009) BFAST: An Alignment Tool for Large Scale Genome Resequencing. *PLoS ONE* **4**(11). doi:10.1371/journal.pone.0007767

Kumar P, Henikoff S, Ng P.C. (2009): Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073-1081. doi:10.1038/nprot.2009.86

Lam HYK, Cuping P, Clark MJ, Lacroute P, Chen R et al. (2012) Detecting and annotating genetic variations using the HugeSeq pipeline. *Nature Biotechnology* **30,** 226-229. doi:10.1038/nbt.2134

Li MX, Gui HS, Kwan JSH, Bao SY and Sham PC (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. *Nucl. Acids Res.* **40**(7). doi: 10.1093/nar/gkr1257

San Lucas FA, Wang G, Scheet P and Peng B (2012) Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* **28**(3), 421-422. doi:10,1093/bioinformatics/btr667

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L et al. (2001) dbSNP: the NCBI database of genetic variation. *Nucl. Acids Res.* **29**, 308-311. doi: 10.1093/nar/29.1.308

The 1000 genomes project consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1051. doi:10.1038/nature09534

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl. Acids Res.* **38**(16). doi: 10.1093/nar/gkq603

Yandell M, Huff C, Hu H, Singleton M, Moore B et al. (2011) A probabilistic disease-gene finder for personal genomes. *Genome Research* **21**(9), 1529-1542. doi: 10.1101/gr.123158.11