Comparison of variant calling methods in exome sequencing of matched tumor-normal sample pairs

Sara Monzon[1,2], Javier Alonso[2], Gonzalo Gómez[3], David Gonzalez-Pisano[3], Isabel Cuesta[1] ✉
[1]Bioinformatic Unit, National Centre of Microbiology, Instituto de Salud Carlos III (ISCIII), Majadahonda, Madrid, Spain
[2]Chilhood Solid Tumor Unit, Instituto de Investigación de Enfermedades Raras, ISCIII, Majadahonda, Madrid, Spain
[3]Bioinformatic Unit, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain

Motivation and Objectives

Next Generation Sequencing techniques are allowing the determination of new somatic mutations involved in cancer development. One of the Bioinformatics' challenges is to develop computational tools able to distinguish in a reliable way the germline polymorphisms present in healthy tissue from the somatically acquired mutations in tumor cells. There has been described two families of somatic variant calling approaches, in earlier one somatic variants have been detected by independently genotyping both samples and subtracting the results (i.e Samtools, Unified Genotyper), in contrast to new one which make simultaneous analysis of tumor and normal datasets from the same individual (i.e. Strelka, JointSNVMix). In our knowledge, just a few reliable studies compare their results. The aim of this work is to compare these two different somatic variant calling approaches analyzing sequenced exome of tumor-matched normal sample pairs.

Methods

A new workflow that allows comparison of different variant calling methods (Samtools v. 0,1,16 (Li *et al.*, 2009), Unified Genotyper v. 1,6 (DePristo *et al.,* 2011), Strelka v. 0,4,6 (Saunders *et al.,* 2012) and JointSNVMix v. 0,8 (Roth *et al.,* 2012)) has been developed. This workflow has been tested using two exome paired-end matched tumor-normal data sets obtained from pediatric cancers: dataset A (Illumina GAIIx, two patients – two tumors) and dataset B (Illumina HiSeq, 11 patients – 13 tumors). The threshold used for Samtools was selected by default, for Unified Genotyper was the recommended configuration in the GATK best practice guidelines, for Strelka was the recommended configuration with a quality score ≥15, and for JointSNVMix was P(somatic) ≥ 0,8. Somatic variants reported by all methods were manually curated using IGV (Integrative Genomics Viewer; Robinson *et al.,* 2011).

Results and Discussion

The results obtained analyzing Dataset A are showed in Table 1. The simultaneous analysis of tumor–normal paired sequence used by Strelka or JointSNVMix, gives a lower false positive variant number than independent analysis of the tumor and normal data, approach followed by software like Samtools and Unified Genotyper, both commonly used in variant calling workflows.

Preliminary results prove that the determination of somatic mutations in tumors requires that the specific algorithms are able to analyze, in a combined way, the information provided by tumor DNA and constitutional DNA, and thus enabling better precise distinction between germinal and somatic variants.

Table 1. Number of somatic variants validated by manual curation with IGV. FP: False positive, TP: True Positive.

Methods	Variant number	FP	TP
Samtools - pileup	71	63	8
Unified Genotyper	14	13	1
Strelka	13	0	13
JointSNVMix	7	5	2

Acknowledgements

We acknowledge ASION (Chilhood Cancer Association of The Community of Madrid), special Program "Hucha de Tomás", to support this work. Chilhood Cancer Genome grant TVP 1278/21.

References

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078-9. doi: 10.1093/bioinformatics/btp352

DePristo M, Banks E, Poplin R, Garimella K, Maguire J 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.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES et al. Integrative Genomics Viewer (2011). *Nature Biotechnology* **29**, 24–26. doi:10.1093/bib/bbs017

Roth A, Ding J, Morin R, Crisan A, Ha G, et al., (2012). JointSNVMix: A Probabilistic Model For Accurate Detection Of Somatic Mutations In Normal/Tumour Paired Next Generation Sequencing Data. *Bioinformatics* **28** (7), 907–913. doi:10.1093/bioinformatics/bts053

Saunders CT, Wong W, Swam S, Becq J et al. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28** (14), 1811-1817. doi:10.1093/bioinformatics/bts271