Quality measures to improve variant calling of Ion Torrent data

Simona De Summa[1]*, Giovanni Malerba[2]*, Antonio Mori[2], Vladan Mijatovic[2], Rosamaria Pinto[1], Stefania Tommasi[1]

[1] Molecular Genetics Laboratory, IRCCS-Istituto Tumori "Giovanni Paolo II", Bari-ITALY [2]Department of Life and Reproduction Sciences, University of Verona, Verona, Italy

* Equal contributors

Introduction

- In the last decade, sequencing technologies, the so-called next generation sequencing (NGS), have delivered a step change in the ability to sequence genome. The introduction of benchtop NGS has offered a powerful alternative for mutation detection.
- In 2011, the distribution of Ion Torrent Personal Genome Machine (PGM) has began. A number of studies has used the PGM to detect genetic variation associated to human diseases.
- In general, the PGM performed well in accurately detecting single nucleotide variations (SNVs) but the overall specificity remained low due to the high false positive rate for indel detection. Due to the nature of the sequencing chemistry of PGM, any homopolymeric region will have a higher indel calling error rate The bioinformatic question on NGS data and, in particular, Ion Torrent data from targeted sequencing requires a lot of efforts in order to correctly identify the best analysis pipeline.
- **The aim of the present study was to compare results from the "gold standard" GATK and Suite Variant Caller analyses regarding a custom panel including 11 genes involved in the response to Vemurafenib, a target therapy for melanoma patients.**

Methods

- *Dataset.* A dataset of 26 metastastic melanoma formalin-fixed paraffin embedded (FFPE) samples was used. Exonic regions of a panel of 11 genes were sequenced with Ion Torrent PGM.
- *GATK pipeline.* We followed Toolkit for Genome Analysis (GATK) recommendations of DNAseq best practices for calling variants. The following software were used: BWA-mem for sequence alignment and GATK 3.4 software for the later steps. In detail, sequences after having been aligned to the human genome reference (version hg19), were not marked for duplicates but underwent realignment around *indels*, base recalibration and variant discovery (using the haplotype caller function in ERC mode). Different filters for SNPs and *Indels* were applied.
- *Torrent Suite Variant Caller (TVC).* TMAP 4.0.6 was used by Torrent Suite as mapper to align reads. VCF files generated by Torrent Variant Caller (TVC 4.2-18), setting "somatic" and "high stringency" as parameters.
- *Simulated Dataset*. One hundred (100) random replicates of 26 samples each with samples presenting a genetic profile according to the genotipic frequencies observed in the real samples were generated and analysed. Fastq files were then simulated though ART [10]. In order to identify the POQ parameters to be used for hard filtering, an in-house script was created to explore 200,000 different combinations.

STEP 1
REAL DATASET

- **TMAP 4.0.6/TVC 4.2-18**
- *Parameters:* «somatic» and «high stringency
- **BWA/GATK 3.4**
- *SNP filters*: QD>2.0, FS< 60.0, MQ> 40.0, MQRankSum> -12.5 and ReadPosRankSum > -8.0.
- *Indels filters*:QD>2.0, FS>200, ReadPosRankSum > -20.

STEP 2
SIMULATED DATASET

- 100 random replicates of 26 samples
- Fastq files were then simulated though ART
- **BWA/GATK 3.4**
- in-house script was created to explore 200,000 different combinations of parameters for hard filtering

Results and Conclusions

- **Real dataset**. TVC called 399 variants in the entire dataset, 73 of which were shared with GATK that detected 83 SNVs. Then we performed comparison on VCF files, which are the output files both of TVC and GATK, focusing on some Parameters Of sequencing Quality (POQ). In particular, QD (Quality By depth), PD (Coverage) and AF (Allele Frequency) tags were shared by VCF outputs. The 73 SNVs called by both TVC and GATK showed higher mean QD and DP values. This observation could suggest that shared SNVs are true positive calls, having higher coverage and quality by depth values. It is noteworthy that mean POQ values of the variants called only by GATK were similar to TVC calls. SNVs identified only by TVC showed lower mean QD, DP and AF values than shared variations. Such results suggest that SNVs called by TVC may be enriched of many false positives. (Table 1).
- **Simulated data.** Statistical analyses on the 200.000 different combinations of POQ are still ongoing. We show results of one replicates from 100 replicates considering only minimum and maximum values of the considered range for POQ parameters. In a very preliminary way, the more interesting observation is that more stringent filters are able to increase the number of combinations of POQ with errors (Table 2).
- **In conclusion, even if our analyses on simulated dataset have not been completed, the correct setting of hard filters could be useful in improving GATK calling of Ion Torrent PGM data.**

	Mean QD	Mean PD	Mean AF
TVC only calls	1.52	663.75	0.08
GATK only calls	8.06	1795.33	0.16
Shared calls	12.67	4155.86	0.18

Table 1. Mean POQ values for calls of GATK, TVC and both.

Table 2. Number of correct calls on 412 total alterations present in each replicates and number of errors, considering also the number of combinations

	QD		FS		MQRankSum		MQ		ReadPosRankSum	
	<2	**<5**	**<30**	**<290**	**<-4**	**<-20**	**<30**	**<50**	**<-6**	**<-30**
Correct calls/n combinations	395/4	394/4	395/7	395/6	395/7	393/7	395/7	395/6	393/7	395/8
Errors/n combinations	3/15	3/15	3/24	3/21	3/24	3/21	3/21	3/24	3/21	3/24