Single Nucleotide Variant Calling Tools for RNA-Seq

Yunqin Chen and Jia Wei
AstraZeneca, R&D Information, Shanghai, China
Email: {Yunqin.chen, jenny.wei} @astrazeneca.com

***Abstract*—Dissecting the transcriptome is essential for understanding the functional element of genome and molecular constituents of cells and tissues, and also important for revealing the cancer mechanism. High-throughput RNA sequencing (RNA-Seq) has enabled whole genome and transcriptome single nucleotide variant (SNV) discovery in cancer. In recent years, a number of SNV identification methods have been published from both public and commercial sources. Here we presented an overview and evaluation of these attempts on SNV calling. We defined a set of criteria and compared the performance of four tools (GATK, Samtools, VarScan and Array Studio) based on these criteria, and we further provided advices on lowering false positive mutation rate.**

***Index Terms*—RNA-Seq, SNV, transcriptome**

I. Introduction

Single nucleotide variants (SNVs) present as either germline or somatic point mutations are essential drivers of tumorigenesis in many human cancer types. Because the contribution of single germline alleles to the population burden of cancer is relatively low, the determination of tumorigenic mechanisms has focused on somatic mutations [1].

The somatic mutational landscape of cancer has to date largely been derived from small-scale or targeted sequencing approaches. RNA-Seq has emerged as a practical, high-throughput and low-cost sequencing method enabling the full and rapid interrogation of the genomes and transcriptomes of individual tumors for mutations. It is regarded as a revolutionary tool for transcriptomics [2] and helpful for cancer therapy [3]. One promising aspect of RNA-Seq is that it produces actual sequences of mRNA molecules and can be used for comparing tumor and normal samples for mutations in coding regions.

To fully enable RNA-Seq technology to detect SNVs in cancer, powerful computational tools are required. In the past two years, software applications for RNA-Seq analysis have been flooding the market from public domains as well as commercial organizations. Many tools identify SNVs with high false positive mutation rate. How to identify and use the suitable tools, and reduce false positive mutations becomes critical.

Here we focus on the evaulation of several computational methods for SNV calling by RNA-Seq. Using Google Scholar citation, we selected three popular analysis pipelines from public domains and one workflow from commercial products. We applied them to a human Gastric cancer RNA-Seq dataset consisting of 40 million paired-end 101-base reads.

II. Methodologies and Results

A. Datasets

A Gastric cancer sample was obtained from gastric cancer patient. mRNA was fragmented and plus- and minus- strand cDNA were synthesized for illumina pair-end sequencing. A 300-bp fragment size was selected by gel excision and the sample was sequenced twice to avoid technical variance.

There are totally 30,121,416 and 17,510,256 read pairs for each replica. Short paired reads (100bp) were assembled and mapped to the annotated human reference genome (human B37) using OSA [4]. We trimmed reads with quality score <=2 and no-unique mapping. The aligned bam files were used for the following variant detecting.

B. Methods

Three public SNV calling tools, GATK [5], Samtools [6], VarScan [7] are selected according to their average citation numbers per month (CPM) calculated by the total number of citations retrieved from Google Scholar divided by the number of months since their publication date (Table I).

Picard [6] which comprises java-based command-line utilities is used to remove duplicate reads and eliminate false positive variants.

These three tools have different strength on variant detection. GATK developed at Broad Institute has a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Unified Genotyper is its variant caller. Fig. 1 describes the GATK pipeline we constructed. SAMtools manipulates alignments in the SAM format with Bcftools as variant calling procedure. The SAMtools derived pipeline is shown in Fig. 2. VarScan is a variant detection tool suitable for massively parallel sequencing data. Fig. 3 shows the VarScan processing workflow we used.

All the above three tools are installed in the Linux environment: Red hat enterprise linux AS release 4

Manuscript received December 8, 2012; revised March 19, 2013;

doi: 10.12720/jomb.2.2.84-87

operation system, 2X Quad-Core AMD Opteron(tm) Processor 8360 (4 cores) cpu, 64G memory.

Array Studio is a suite of tools developed by OmicSoft (www.omicsoft.com) in which a number of RNA-Seq analysis workflows are provided. SNV calling can be performed directly after OSA alignment on Windows system. We used Array Studio version 5.0.

TABLE I. OPEN SOURCE TOOLS SELCTION CRITERIA

Tools	Reference	Citations	C.P.M	Availability	Version
GATK	Genome Res. *2010. 20: 1297-1303*	381	11.5	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	Release 1.4
Samtools	Bioinformatics *(2009) 25 (16): 2078-2079.*	1259	4.877	http://samtools.sourceforge.net/	0.1.17
VarScan	*Bioinformatics (2009) 25 (17): 2283-2285.*	130	4.877	http://varscan.sourceforge.net/	v2.2.8

C. Results

Among the three public tools running on Linux, VarScan consumed least time (168 minutes) compared to the other two tools (220 mins for GATK and 206 mins for SAMtools). And the three tools used same amount of memory (2G) when doing variant calling.

Here are the criteria for variation detection by all tools:

- Depth > 5 and variation frequency >0.2
- Map quality >=10 and base quality >=20
- Calling region : (exon start position -10bp, exon end position+10bp)

As shown in Table II, SNP transition to transversion ratio (Ti/Tv) was used to confirm SNP discovery. Recent human studies particularly from the 1000 genomes project have been showing that for whole human genome, a Ti/Tv of around 2-2.1 is generally correct. This is only when assessing the genome as a whole [8]. However, different specific genetic regions will display different Ti/Tv ratios. With regard to Human exomes, it appears that the ratio of Ti/Tv increases to about 3 [9]. The Ti/Tv ratio generated from all the tools we evaluated is consistent with these findings.

From Table II, we can see most SNVs detected by all tools can be found in DBSNP [10]. Pipeline using Samtools after pre-processing BAM files with GATK seems to detect higher novel SNVs rate than other methods, while VarScan seems to detecthigher known SNPs rate than other methods.

TABLE II. SNV RESULTS FROM DIFFERENT TOOLS

	No. of SNPs					Ti/Tv	
Tools	All	Known	Novel	dbSNP%	Concordant%	Known	Novel
GATK	29933	21169	8764	70.72	99.55	2.65	2.14
GATK+SAMtools (SAMtools)	31806 (36161)	21765 (24976)	10041 (11185)	61.48 (69.07)	99.25 (99.44)	2. 67 (2.70)	2.44 (2.12)
GATK+Varscan (VarScan)	23420 (22103)	18910 (18109)	4510 (3994)	80.74 (81.93)	-	2.67 (2.44)	1.79 (2.12)
Arraystudio	28703	20491	8212	71.39	-	2.68	3.28

TABLE III. NON- SYNONYMOUS SNV RESULTS FROM DIFFERENT TOOLS

Tools	Non-synonymous SNP	Associated genes with ns-SNP	Annotated by dbSNP	Novel ns-SNP
GATK	4963	2892	3449	1514
GATK+SAMtools (SAMtools)	4653 (5292)	2751 (3014)	3325 (3816)	1328 (1476)
GATK+VarScan (VarScan)	3802 (3486)	2340 (2180)	3057 (2531)	749 (955)
Arraystudio	4556	2737	3556	1200

SNP can be characterized as non-coding, synonymous which is in coding region with no protein sequence changes, or non-synonymous which falls in coding region with protein sequence changes. Non-synonymous SNPs are more interesting, so we used ANNOVAR [11] to further annotate SNPs and summarized non-synonymous SNP information from different tools in Table III. Our data show that GATK detects more novel non-synonymous SNPs than other tools, while VarScan detects less non-synonymous SNPs than other methods.

We also compared GATK and other tools in terms of overlapping SNVs detected. The results are summarized in Fig. 4. The data show that 93% of SNVs detected by GATK overlap with those detected by Samtools, 82% of SNVs detected by GATK overlap with those detected by Arraystudio and 76% of SNVs detected by GATK overlap with those detected by VarScan.

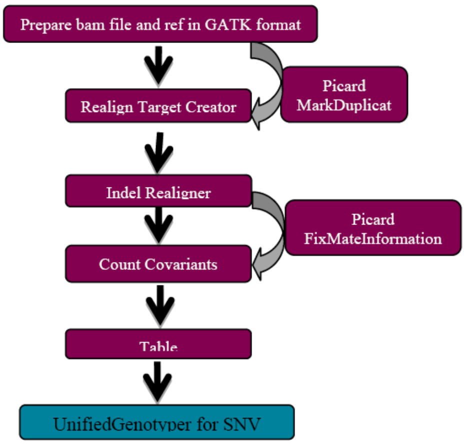

Figure 1. Pipeline for GATK calling SNV

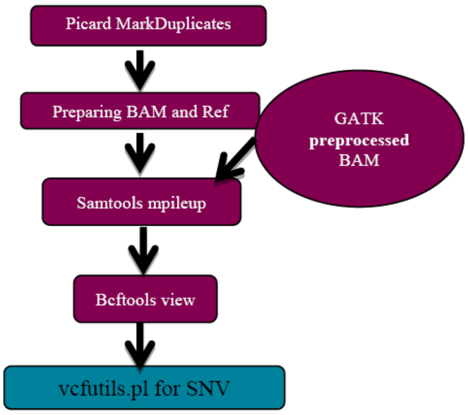

Figure 2. Pipeline for Samtools detecting SNV alone or in combination with GATK pre-processing bam file

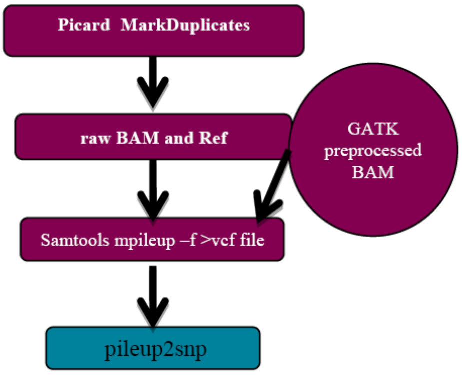

Figure 3. Pipeline for VarScan calling SNV alone or in combination with GATK pre-processing bam file

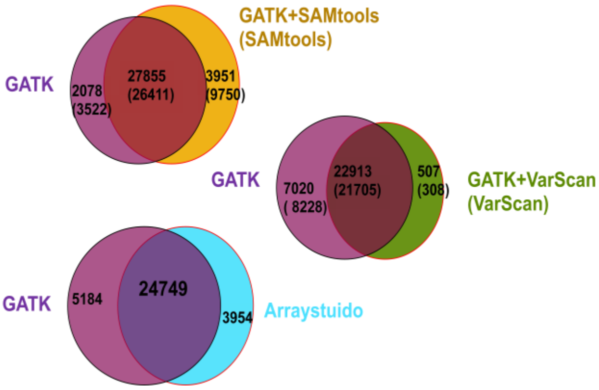

Figure 4. SNVs overlap between GATK and other tools

III. Conclusions

In this work, we have shown the evaluation results of a set of public and commercial tools for SNV calling by RNA-Seq. Among all the tools, GATK is more difficult to use due to its format restrictions, but it provides more metrics to evaluate SNVs and is more suitable for deep analysis. To reduce false positive rate of SNV calling, it is necessary to use Picard to mark duplicate reads and GATK to pre-process data. Based on our data, GATK and Samtools give similar SNV calling results, while SNV callings by VarScan have less overlap with those by GATK. Because of our alignment setting with no gap, no indels were detected in those tools. Because of rapid improvements in RNA-Seq data generation, more efforts need to be done in the areas of SNV detection to find driver mutations. New questions will continue to emerge and novel programs will evolve. The tool evaluation needs to keep up with the pace of these changes in order to apply RNA-Seq technologies to cancer discovery.

Acknowledgment

The authors wish to thank Ziliang Qian and Hua Dong for the helpful discussions on the results. This work was supported in part by a project from AstraZeneca.

References

[1] M. R. Stratton, *et al.*, “The cancer genome.” *Nature*, 458, 719–724, 2009.

[2] Z. Wang, M. Gerstein, *et al.*, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nat Rev Genet*, vol. 10, no. 1, pp. 57-63, 2009.

[3] D. Sinicropi, K. Qu, *et al*, “Whole transcriptome RNA-Seq analysis of breast cancer recurrence risk using formalin-fixed paraffin-embedded tumor tissue,” *PLoS One* vol. 7, no. 7, 2012.

[4] J. Hu, H. Ge, *et al.* “OSA: a fast and accurate alignment tool for RNA-Seq,” *Bioinformatics*, vol. 28, no. 14, pp. 1933-1934.

[5] A. McKenna, M. Hanna, *et al*, “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res*, vol. 20, no. 9, pp. 1297-303, 2010.

[6] H. Li, B. Handsaker, *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078-9, 2009.

[7] D. C. Koboldt , K. Chen, *et al.*, “VarScan: variant detection in massively parallel sequencing of individual and pooled samples,” *Bioinformatics*, vol. 25, no. 17, pp. 2283-5, 2009.

[8] H. Li, “Improving SNP discovery by base alignment quality,” *Bioinformatics (Oxford, England)*, vol. 27, no. 8, pp. 1157–8, 2011.

[9] A. J. Coffey, F. Kokocinski, M. S. Calafato, *et al.*, “The GENCODE exome: Sequencing the complete human exome.

European journal of human genetics," *EJHG*, vol. 19, no. 7, pp. 827–31, 2011.

[10] S. T. Sherry, "DBSNP: The NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[11] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, vol. 38, no. 16, pp. e164, 2010.

Yunqin Chen is an associate informatics scientist of R&D Information, AstraZeneca since 2012. She received master degree in bioinformatics from Shanghai TongJi University in 2011. She focuses on regulation of gene expression and function, next generation sequencing data analysis, and has experience in biological knowledge mining using informatics methods.

Jia Wei is the Head of R&D Information China, AstraZeneca global R&D.

Dr. Wei joined AZ Innovation Centre China as Principal Scientist in Informatics in 2008 and shortly was appointed to Head of Information Management in charge of informatics infrastructure and applications for AZ translational medicine efforts in China. In July 2011, she was promoted to Head of R&D Information China, leading R&D Information China Hub to deliver informatics solutions for AZ global R&D. Prior to AstraZeneca, she was a Principal Software Developer and Senior Product Architect at Merck Rosetta Biosoftware developing informatics solutions for genomics data management and data analysis. During her more than 14 years informatics/software careers, she also held positions in Merck Research Lab Department of Bioinformatics and AT&T Wireless.

Dr. Wei received her PhD in Pharmacology from University of Washington, USA in 1998 and her MS in Computer Science from Rutgers University, USA in 1999.