Genomics 107 (2016) 163–169

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods Paper

Practicability of detecting somatic point mutation from RNA high throughput sequencing data

Quanhu Sheng [a,b], Shilin Zhao [a,b], Chung-I Li [c], Yu Shyr [a,b,d,*], Yan Guo [a,b,**]

[a] Vanderbilt Ingram Cancer Center, Center for Quantitative Sciences, Nashville, TN, USA
[b] Department of Cancer Biology, Vanderbilt University, Nashville, TN, USA
[c] Department of Statistics, National Cheng Kung University, Taiwan
[d] Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 11 February 2016
Received in revised form 29 March 2016
Accepted 30 March 2016
Available online 2 April 2016

Keywords:
Somatic mutation
RNAseq
Exome
Generalized linear model

ABSTRACT

Traditionally, somatic mutations are detected by examining DNA sequence. The maturity of sequencing technology has allowed researchers to screen for somatic mutations in the whole genome. Increasingly, researchers have become interested in identifying somatic mutations through RNAseq data. With this motivation, we evaluated the practicability of detecting somatic mutations from RNAseq data. Current somatic mutation calling tools were designed for DNA sequencing data. To increase performance on RNAseq data, we developed a somatic mutation caller GLMVC based on bias reduced generalized linear model for both DNA and RNA sequencing data. Through comparison with MuTect and Varscan we showed that GLMVC performed better for somatic mutation detection using exome sequencing or RNAseq data. GLMVC is freely available for download at the following website: https://github.com/shengqh/GLMVC/wiki.

1. Introduction

Traditionally, somatic mutations are detected using Sanger sequencing or real-time polymerase chain reaction (RT-PCR) by comparing paired tumor and normal samples. One obvious limitation of such methods is that the somatic mutation detection must be limited to a certain genomic region of interest. Now with high-throughput sequencing (HTS), whole exomes or genomes can be screened for somatic mutations at a reasonable cost (Fig. S1). There are two major next-generation sequencing (NGS) paradigms: RNA and DNA sequencing. Both RNA and DNA sequencing can be used to answer different sets of scientific questions important for biomedical research. RNAseq refers to the sequencing of the transcriptome. The two most common forms of DNAseq are exome and whole genome sequencing.

Due to the popularity of RNAseq technology for gene expression profiling over microarray technology [1–4], huge amounts of RNAseq data have been accumulated over the past few years. And the majority of these RNAseq data has been only studied for gene expression. More and more researchers have begun to ask the question of whether or not somatic mutations can be detected accurately through RNAseq data. Same as DNAseq, RNAseq is at single nucleotide resolution. Thus, single nucleotide variants (SNVs) can be detected. To date, many tools, such as Varscan [5] and MuTect [6], have been developed for the identification of somatic mutations through DNAseq data. Yet, less effort has been relatively spent on the detection of SNVs using RNAseq data. In contrast to using DNAseq data, identifying mutations using RNAseq data poses stronger challenges for the primary reason of RNAseq data having a much higher false positive rate for SNVs than DNAseq data [7,8]. The high false positive rate results from several issues, of which include cycle bias [9], strand bias [10] alignment complexity in the transcriptome, RNA editing, and random errors introduced during reverse transcription and PCR. Cycle bias happens in a heterozygous position when one of two alleles in the supporting reads lie heavily at the beginning or end of the reads [11,12]. Strand bias occurs when alternative allele detection heavily originates from one of the two strands (forward or reverse). Such bias indicates false positive mutation detection in RNAseq data [12]. Most advanced somatic mutation callers [5,6,13] have built-in strand bias quality control. Also, the alignment of RNAseq data proves more complicated than DNAseq data [14]. In mRNA, introns are removed by splicing, thus a read is likely to span the splicing junction, causing a higher probability for error. Similarly, processes such as RNA editing and polyadenylation introduce additional mismatches not found in DNAseq alignment. For conducting expression studies, minor mismatches in alignment do not affect expression value because the computation of expression value depends only on the count of reads mapped to a gene's genomic span and therefore do not require the examination of the RNAseq at single nucleotide resolution for gene expression. However, SNVs are detected by counting the number of mismatches in alignment against a reference. Thus, excessive mismatches due to errors described above will result in a high false positive

* Correspondence to: Y. Shyr, 2220 Pierce Ave, 571 PRB, Vanderbilt University, USA.
** Correspondence to: Y. Guo, 2220 Pierce Ave, 494 PRB, Vanderbilt University, USA.
E-mail addresses: Yu.shyr@vanderbilt.edu (Y. Shyr), Yan.guo@vanderbilt.edu (Y. Guo).

http://dx.doi.org/10.1016/j.ygeno.2016.03.006
0888-7543/

rate for SNV detection. False positives due to cycle bias may be filtered out through a quality control check that removes all reported mutations at the beginning or end of the reads that are disproportionate. This has been effectively demonstrated by Kleinman et al. [15]. False positives due to splicing locations are more difficult to distinguish from true variants. Thus, SNPs and somatic mutations identified near splicing sites should be removed or flagged for further review. Most RNAseq data specific variant detection tools, such as SNVQ [16] and SNPiR [14], focus on

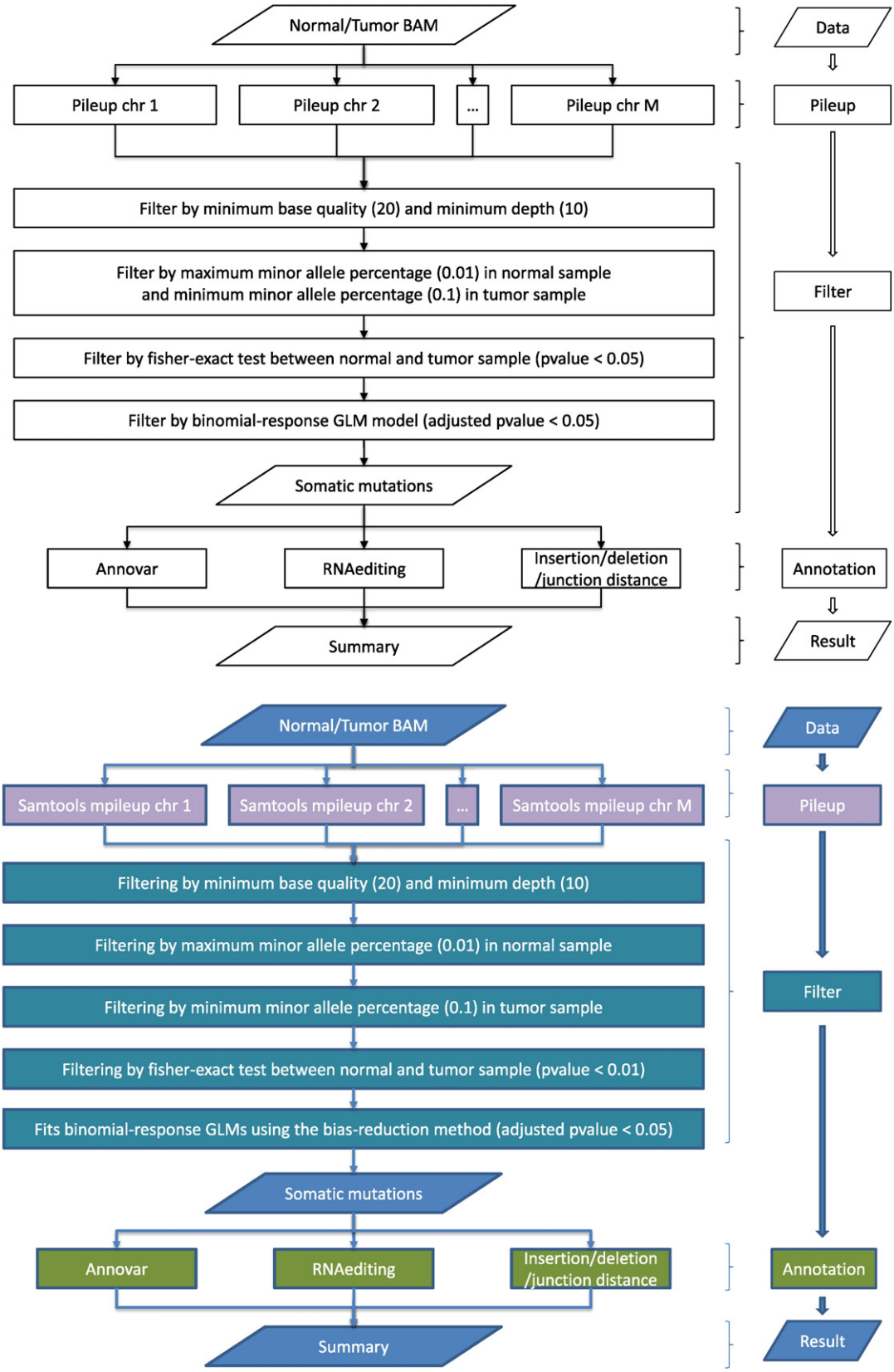

Fig. 1. GLMVC workflow.