Genome Medicine

RESEARCH Open Access

Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers

Qingguo Wang[1], Peilin Jia[1,2], Fei Li[3], Haiquan Chen[4,5], Hongbin Ji[3], Donald Hucks[6], Kimberly Brown Dahlman[6,7], William Pao[6,8*] and Zhongming Zhao[1,2,7,9*]

Abstract

Background: Driven by high throughput next generation sequencing technologies and the pressing need to decipher cancer genomes, computational approaches for detecting somatic single nucleotide variants (sSNVs) have undergone dramatic improvements during the past 2 years. The recently developed tools typically compare a tumor sample directly with a matched normal sample at each variant locus in order to increase the accuracy of sSNV calling. These programs also address the detection of sSNVs at low allele frequencies, allowing for the study of tumor heterogeneity, cancer subclones, and mutation evolution in cancer development.

Methods: We used whole genome sequencing (Illumina Genome Analyzer IIx platform) of a melanoma sample and matched blood, whole exome sequencing (Illumina HiSeq 2000 platform) of 18 lung tumor-normal pairs and seven lung cancer cell lines to evaluate six tools for sSNV detection: EBCall, JointSNVMix, MuTect, SomaticSniper, Strelka, and VarScan 2, with a focus on MuTect and VarScan 2, two widely used publicly available software tools. Default/suggested parameters were used to run these tools. The missense sSNVs detected in these samples were validated through PCR and direct sequencing of genomic DNA from the samples. We also simulated 10 tumor-normal pairs to explore the ability of these programs to detect low allelic-frequency sSNVs.

Results: Out of the 237 sSNVs successfully validated in our cancer samples, VarScan 2 and MuTect detected the most of any tools (that is, 204 and 192, respectively). MuTect identified 11 more low-coverage validated sSNVs than VarScan 2, but missed 11 more sSNVs with alternate alleles in normal samples than VarScan 2. When examining the false calls of each tool using 169 invalidated sSNVs, we observed >63% false calls detected in the lung cancer cell lines had alternate alleles in normal samples. Additionally, from our simulation data, VarScan 2 identified more sSNVs than other tools, while MuTect characterized most low allelic-fraction sSNVs.

Conclusions: Our study explored the typical false-positive and false-negative detections that arise from the use of sSNV-calling tools. Our results suggest that despite recent progress, these tools have significant room for improvement, especially in the discrimination of low coverage/allelic-frequency sSNVs and sSNVs with alternate alleles in normal samples.

* Correspondence: william.pao@vanderbilt.edu; zhongming.zhao@vanderbilt.edu
[6]Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA
[1]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA
Full list of author information is available at the end of the article