Genome Medicine

METHOD Open Access

A simple consensus approach improves somatic mutation prediction accuracy

David L Goode[1,5*†], Sally M Hunter[2†], Maria A Doyle[3], Tao Ma[3,6], Simone M Rowley[2], David Choong[2], Georgina L Ryland[2,4] and Ian G Campbell[2,5,7]

Abstract

Differentiating true somatic mutations from artifacts in massively parallel sequencing data is an immense challenge. To develop methods for optimal somatic mutation detection and to identify factors influencing somatic mutation prediction accuracy, we validated predictions from three somatic mutation detection algorithms, MuTect, JointSNVMix2 and SomaticSniper, by Sanger sequencing. Full consensus predictions had a validation rate of >98%, but some partial consensus predictions validated too. In cases of partial consensus, read depth and mapping quality data, along with additional prediction methods, aided in removing inaccurate predictions. Our consensus approach is fast, flexible and provides a high-confidence list of putative somatic mutations.

Background

Massively parallel sequencing (MPS) of cancer exomes is becoming a commonplace technique, and has led to the identification of genes underlying the pathogenesis of a number of cancer types [1-6]. In response to the volume of data generated by these genome-scale studies, a host of software tools has been developed to aid in distinguishing genuine somatic mutations from germline variation, alignment artifacts, and inherent MPS errors [7-11]. The rarity and diversity of somatic events that occur on a background of tumor heterogeneity, normal contamination, technical artifacts, and genomic complexity makes this task particularly challenging [1,12].

Although the methodology applied by somatic mutation algorithms varies somewhat, the aim of each program is to identify tumor-specific variants by comparing sequence data from a tumor with that generated from a normal tissue (representing the germline) from the same patient (that is, matched normal DNA). The most common application is the identification of point mutations. The germline sample is usually assumed to be free of genetic material from the tumor, although this assumption can be tested and corrected for [12,13]. At every site where there are reads that differ from the reference genome, the probability that these reads contain legitimate genetic variants and not sequencing errors or technical artifacts is calculated. The probabilities for the tumor and germline data are compared, and a prediction about whether the site harbors a somatic mutation is made [7-11]. From this, a list of putative somatic mutations and associated confidence values is produced, which can be used in downstream analyses.

The choice of somatic mutation detection algorithm may have an important influence on the outcome of a tumor exome-sequencing study. Incorporating more information from a sequencing run (such as site-specific mapping and base qualities) improves the performance of variant detection over that of *ad hoc* metrics based on read counts alone [7,10,14,15]. Thus, it would be expected that predictions from different algorithms, which weigh different properties of the data in unique ways, may differ significantly. A conservative algorithm with high specificity may make very few incorrect predictions, but may miss many legitimate somatic mutations because of its low sensitivity. Similarly, a high-confidence set of somatic mutation predictions with a low false-positive (FP) rate is very useful in a clinical setting, but in a discovery-based research setting, it could limit the power to identify novel mutated genes and pathways [16]. This is important given the small number of recurrently mutated tumor driver genes, and the long list of infrequently mutated, yet biologically

* Correspondence: david.goode@petermac.org
†Equal contributors
[1]Peter MacCallum Cancer Centre, Sarcoma Genetics and Genomics Laboratory, St. Andrew's Place, East Melbourne, Victoria, Australia
[5]Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Victoria, Australia
Full list of author information is available at the end of the article