Title

Evaluation of strategies for somatic mutation discovery in tumor specimens without matched germline: effect of tumor content, sequencing depth, and copy number alterations

Authors

Rebecca F. Halperin[1], John D. Carpten[2], Jessica Aldrich[1], Winnie S. Liang[1], Jonathan Keats[3], Megan Russell[1], Daniel Enriquez[1], Ana Claasen[1], Irene Cherni[3], Seungchan Kim[3], David W. Craig[1],

[1]Neurogenomics Division, Translational Genomics Research Institute, Phoenix, AZ
[2]University of Southern California, Department of Translational Genomics, Los Angeles, CA
[1]Integrated Cancer Division, Translational Genomics Research Institute, Phoenix, AZ

Introduction

Large-scale multiplexed identification of somatic alterations in cancer has become feasible with next generation sequencing (NGS). By definition, somatic alterations are those that are found in the tumor and not the germline sequence, so the standard approach to somatic variant detection involves comparing the tumor sequence to the germline sequence of the same individual. However, in some situations, such as with archival samples, blood or other constitutional tissue samples are not available to obtain germline sequence. In order to identify somatic variants in such tumor samples, the tumor is typically compared to a reference sample, and then the variants that are found public germline variant databases are filtered out. However, all individuals will have some private germline variants not found in any database. Differences in allele frequencies between somatic and germline variants in impure tumors can also help to differentiate somatic and germline variants. Here we will examine the extent to which leveraging allele frequencies can help to overcome false positives due to private germline variants in tumor only calling.

Results

We developed a Bayesian framework to integrate the population frequency and allele frequency information. At each position, we determine the prior probability of a germline or somatic based on 1000 Genomes or COSMIC frequencies, respectively. We also estimate copy number, minor allele copy number, and clonal sample fraction in order to calculate expected allele frequencies of somatic and germline variants at each position. As expected, the higher the clonal sample fraction, the closer the expected allele frequencies are for somatic and germline variants. We also find that there also other combinations of tumor content and copy number state where the expected allele frequencies of somatic and germline variants are very similar.

Applying this framework to simulated data, we estimate coverage required for different tumor content and copy number states. For example, to detect about 90% of the somatic variants in a diploid region of a 50% tumor sample, we would only need 200X mean target coverage, but we would need 800X mean target coverage to achieve the same sensitivity in a 75% tumor sample, or 1600X for an 85% tumor sample. We then apply the framework to a set of nine cancer samples. We find that the observed

sensitivity correlates well with the expected sensitivity based on the coverage, the clonal sample fractions, and the copy number alterations. In silico dilutions and downsampling experiments also confirm the expected relationships between coverage, tumor content, and sensitivity.

We find that the Bayesian tumor only caller is able to greatly reduce false positives due to private germline variants, with greater than 95% of true private germline variants correctly classified as germline. The calling precision is also significantly improved with Bayesian approach, which has an average positive predictive value of greater than 70% compared to 35% with database filtering alone. Overall the accuracy of the Bayesian tumor only caller is greater than 99.9%

Conclusions

Our Bayesian tumor only calling approach can eliminate most false positives due to private germline variants. However, the sensitivity of the approach is dependent tumor content, coverage, and copy number alterations. The data presented here can be used to design tumor only sequencing experiments with appropriate coverage based on the sample characteristics.