A Bayesian Network Algorithm for Somatic Mutation and Germline Variant Identification from Tumor Molecular Profiling of Cancer Patients by High-Throughput Sequencing

Francisco M. De La Vega,[1,2] Sean Irvine,[3] David Ware[3], Kurt Gaastra[3], Yosr Bouhlai[1], Daniel Mendoza[1], Anna Vilborg[1], Yannick Pouliot[1], Federico Goodsaid[1], Austin So[1], and Len Trigg.[3]
[1]TOMA Biosciences, Foster City, CA 94404, USA, [2]Stanford University School of Medicine, Stanford, CA, USA, and [3]Real Time Genomics, Hamilton, New Zealand.

Background

Cancer tumor profiling by targeted resequencing of actionable cancer genes is rapidly becoming the standard approach for selecting targeted therapies in refractory cancer patients. In this scenario, DNA from a tumor FFPE sample is sequenced deeply by targeted next-generation sequencing (NGS) to uncover actionable somatic mutations in relevant cancer genes. Currently, clinical labs preforming such tests under the CLIA regulation, largely utilize analysis pipelines based in academic tools developed as part of the TCGA or ICGC projects, where tumor and germline specimens from cancer patients are sequenced in parallel to facilitate the identification of cancer somatic mutations *vs* germline variants. A major challenge that arises in the clinical scenario is the need to analyzing tumor-derived data in the absence of normal/germline tissue data, as the current standard of care only requires pathologists to obtain a biopsy of the tumor tissue[1]. This makes very difficult to distinguish between somatic and germline variants, leaving clinicians to resort to crude heuristic filtering procedures with unknown performance. Furthermore, recent benchmarking of somatic calling methods have shown poor performance and significant inconsistencies in the major published algorithms, even when provided with both tumor and normal tissue data[2].

Results

Here we present Bayesian network variant caller to identify both SNV and indel somatic mutations and germline variants from targeted resequencing data from tumor tissue samples. Our approach models the distribution of reads harboring germline and somatic mutations in cancer cells, estimates the contamination from normal tissue in tumor specimens, scores putative somatic mutation, and imputes germline variants present in the genome of cancer cells and contaminating normal cells, without matching normal tissue data. Our "tumor-only" caller can also utilize site- and allele-specific prior information to calculate the scores of somatic mutations, from sources such as databases of *bona fide* somatic mutations (e.g. COSMIC), catalogs of germline variation in populations (e.g. 1000 Genomes Project), and data from a panel of normal samples analyzed with the same assay platform to reduce systematic technology artifacts. This method has been developed in Java on top of the libraries of a previously developed variant caller.[3]

We validated our method by analyzing data obtained with the TOMA OS-Seq targeted enrichment assay for 130 cancer genes and then sequencing with the Illumina platform. Firstly, we obtained data from a gold standard sample for which a ground truth is available, the cell-line NA12878, upon which we simulated about 1,800 somatic mutations at variant allele fractions (VAF) ranging from 0.1 to 0.4, using the `bamsurgeon` software [4]. Secondly, we analyzed data from experiments where varying proportions of a reference sample (e.g. NA12878) is mixed with a constant amount of one of its parents, to simulate the behavior of tumor somatic mutations. Finally, we also analyzed data from cancer patient case triads, where normal, tumor and plasma cell free-DNA have been sequenced and we are able to compare the the results from the tumor-only caller vs the paired tumor/normal analysis also implemented in the software.

The ability to compare our results to a ground truth dataset permits us to evaluate our performance via Receiver Operator Characteristic (ROC) curves, where we can measure performance with the area under the curve, or true positive rates at a fixed FDR. Our initial evaluation of the caller showed that we can improve the AUC by providing priors for a database of somatic mutations, but the major benefit comes from utilizing a panel of normal samples. We can recover over 99% of true positives at a FDR of 1.6% when simulating mutations at a VAF of 0.4. As we reduce the VAF the separation in the improvements obtained by either of those methods decrease, as expected. As we evaluate the performance of our caller, it is important to compare to other commonly used algorithms in cancer tumor profiling. We thus compared our results to the output from `FreeBayes`. We found that we can achieve >90% True Positive Rate (TPR) at 1.5% FDR while `FreeBayes` achieves only 15% TPR. At a 2% FDR, we achieve >99% TPR, while `FreeBayes` only achieves less than 80% TPR. While this is a work in progress and are un the process of evaluating additional datasets through our method and adjusting priors, we observe that our caller performs significantly better than other methods, and highlights the challenges of somatic mutation identification at low VAF.

Conclusions

We show that a Bayesian network approach is a very powerful method to infer somatic mutation calls from NGS data of mixed samples, such as tumor specimens, with the ability to decompose the mixture returning both somatic and germline variants calls, and leverage prior information in a natural and principled fashion. The Bayesian network approach allows not only to call somatic mutations, but to impute the germline genome to a considerable accuracy from the tumor sample. This is important information, as inherited susceptibility variants exist in cancer patients and this information should be used to both inform therapy and provide family counseling. Our method and ensuing software implementation provides a robust solution for a very common use case in clinical applications of NGS, where material form tumor biopsies from patients are analyzed to identify actionable somatic aberrations in the lack of normal sample. While we can strive to change the standard of practice by requiring a sample of the normal tissue to be sequenced in parallel to the tumor sample as done in research protocols for a paired tumor/normal analysis, these changes take many years[1]. In addition, even if these changes occur, this use case is still important to leverage the large scale biobanks of FFPE blocks that medical centers have accumulated for years together with clinical information and that are being sequenced to correlate molecular profiles, therapies, and outcomes retrospectively.

References

1. Topol, E. J. From Dissecting Cadavers to Dissecting Genomes. *Sci Transl Med* **5,** 202ed15–202ed15 (2013).
2. Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications* **6,** 1–13 (2015).
3. Cleary, J. G. *et al.* Joint Variant and De NovoMutation Identification on Pedigrees from High-Throughput Sequencing Data. *Journal of Computational Biology* **21,** 405–419 (2014).
4. Boutros, P. C. *et al.* correspondence. *Nat. Genet.* **46,** 318–319 (2014).