BIOINFORMATICS APPLICATIONS NOTE
Vol. 29 no. 15 2013, pages 1888–1889
doi:10.1093/bioinformatics/btt293

Genome analysis
Advance Access publication June 6, 2013

PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data

Nicholas B. Larson[1,*] and Brooke L. Fridley[2]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55901, USA and [2]Department of Biostatistics, University of Kansas Medical Center, Kansas City, KS 66160, USA

Associate Editor: Michael Brudno

ABSTRACT

Summary: We have developed a novel Bayesian method, PurBayes, to estimate tumor purity and detect intratumor heterogeneity based on next-generation sequencing data of paired tumor-normal tissue samples, which uses finite mixture modeling methods. We demonstrate our approach using simulated data and discuss its performance under varying conditions.

Availability: PurBayes is implemented as an R package, and source code is available for download through CRAN at http://cran.r-project.org/package=PurBayes.

Contact: larson.nicholas@mayo.edu

Supplementary information: Supplementary data are available online at *Bioinformatics* online.

Received on January 11, 2013; revised on May 16, 2013; accepted on May 17, 2013

1 INTRODUCTION

With advances in high-throughput next-generation sequencing (NGS) technologies, sequencing of tumor-normal tissue pairs is becoming commonplace in cancer studies. Often, the sampled tumor tissue is contaminated with stromal cells, resulting in a mixture of tumor and normal sequence data in the tumor sample. There has been a recent interest in accurate estimation of tumor purity levels in tumor data analysis (Carter *et al.*, 2012; Song *et al.*, 2012), including methods specific to NGS data such as PurityEst (Su *et al.*, 2012). However, a subset of the observed somatic mutations may be subclonal because of intratumor heterogeneity (Michor and Polyak, 2010). Unlike clonal mutations, which are observed tumor-wide, subclonal mutations will be observed at cellularities less than the tumor purity level and subsequently bias purity estimates under an assumption of tumor tissue homogeneity. By modeling this heterogeneity, it may also be possible to make inferences about tumor evolution and founder events. To date there are no methods that aim to both quantify tumor purity and detect intratumor heterogeneity using NGS data.

In this article, we present a Bayesian mixture modeling approach, PurBayes, toward estimating tumor purity and subclonality using NGS data, resulting in posterior distributions of tumor cellularities from which credible intervals (CI) can be derived. To illustrate its implementation, we conduct a simulation study under a variety of conditions and discuss the performance of PurBayes on synthetic data.

2 METHODS

For a set of S observed heterozygous loci because of somatically acquired single-nucleotide variants (SNVs) for a given tumor sequencing sample, each SNV can be represented by respective normal and mutant allele read counts X_i and Y_i. The total number of sample reads $N_i = X_i + Y_i$ can in turn be decomposed into respective tumor and normal tissue read counts N_i^t and N_i^w, such that $N_i = N_i^w + N_i^t$. As it cannot be directly determined which cell type each individual read was derived, N_i^t and N_i^w are latent variables. If we assume N_i^t to be binomially distributed, such that $N_i^t \sim \mathrm{Bin}(N_i, \lambda)$ and λ indicates tumor sample purity, and $Y_i|N_i^t \sim \mathrm{Bin}(N_i^t, 0.50)$, then Y_i follows a binomial–binomial hierarchical mixture model with marginal distribution $\mathrm{Bin}(N_i, \lambda/2)$ (Villa and Escobar, 2006).

Consider a tumor that exhibits intratumor heterogeneity. If we assume subclonal mutations cluster into an *a priori* finite number of *J-1* subclonal populations, $\mathbf{Y}$ can be modeled under a Bayesian finite mixture model. Let κ_j denote to the probability a mutation corresponds to variant population j with respective cellularity λ_j, for $j=1,\ldots,J$, such that $\sum_j \kappa_j = 1$, $\lambda_1 < \ldots < \lambda_{J-1} < \lambda_J$, and $\lambda_J \cong \lambda$, with uniform priors on λ_j. To obtain a data-driven value for J, PurBayes generates model fits iteratively by initially assuming tumor homogeneity and then increasing the subclonal population count by one until an optimal model fit is achieved under a penalized expected deviance (PED) criterion (Plummer, 2008).

Mapping bias can result in non-reference alleles in heterozygous loci being mapped at rates <0.50 (Degner *et al.*, 2009), which would impact tumor purity estimation. PurBayes can accommodate this bias by estimating it from additional reference and alternate allele counts in heterozygous normal tissue variant calls.

PurBayes is implemented in the statistical programming language R and uses the MCMC software JAGS (Plummer, 2003). The only inputs required for PurBayes are the tumor tissue read counts (**N** and **Y**) for a set of high-confidence SNVs, which can easily be derived from most variant calling software output file formats on NGS data.

3 SIMULATIONS

To illustrate the performance of PurBayes under a variety of conditions, we conducted simulation studies based on real sequencing data from the 1000 Genomes Project (Abecasis *et al.*, 2010) (details in Supplementary Materials). We first simulated read count data for homogenous tumors ranging in purity from 20–80%, with $S=100$ and average sequencing depth at 50× and 100×. We ran 100 replications of each unique set of conditions and examined the PurBayes posterior median estimates. We ran similar simulations for heterogeneous tumor data with $J=2$ at 100× for various values of κ_j and λ_j to determine how well PurBayes can detect intratumor heterogeneity and estimate tumor purity. For each application, we also simulated read count data from 100 additional germ line variant calls to account for mapping bias. For purposes of comparison, we also applied the PurityEst algorithm to each simulation replicate.

*To whom correspondence should be addressed.

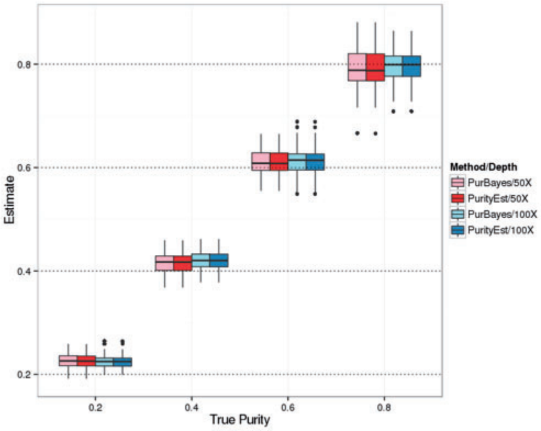

Fig. 1. Side-by-side boxplots of tumor purity estimates by method for true values of $\lambda = 0.2$, 0.4, 0.6 and 0.8

Table 1. Results for heterogeneous ($J = 2$) tumor simulations at 100×, which includes the mean and mean absolute error (MAE) of the posterior median purity estimates under various values of (λ_1, λ_2) and $\kappa_1 = 1 - \kappa_2$. Proportion of replications in which correct heterogeneity was detected (Het) for PurBayes is also reported

(λ_1,λ_2)	κ_1	PurityEst		PurBayes		
		Mean	MAE	Het	Mean	MAE
(0.4,0.8)	0.25	0.701	0.099	0.18	0.790	0.125
	0.50	0.615	0.185	0.43	0.838	0.162
	0.75	0.508	0.292	0.33	0.673	0.230
(0.3,0.6)	0.25	0.534	0.067	0.00	0.534	0.066
	0.50	0.469	0.131	0.05	0.486	0.131
	0.75	0.391	0.209	0.10	0.430	0.199
(0.2,0.8)	0.25	0.656	0.144	0.43	0.923	0.130
	0.50	0.521	0.279	0.55	0.904	0.109
	0.75	0.365	0.435	0.75	0.853	0.075

For each application of PurBayes, the first 50 000 iterations of the optimal MCMC model fit were discarded as a burn-in before posterior sampling of 10 000 iterations. Mean per-sample execution time was ~2 min on a workstation equipped with an Intel® Core$^{\mathrm{TM}}$ i5 3.10 Ghz processor and 4 GB of random access memory.

4 RESULTS AND DISCUSSION

For the homogenous tumor simulations, PurBayes correctly identified tumor homogeneity in all replications. Distributions of the posterior median estimates of tumor purity for each value of λ and method are displayed in Figure 1. Estimates from PurBayes and PurityEst were nearly identical, with a Pearson correlation of 0.9997. Both methods were accurate, tending toward overestimation at lower values of λ. When we applied PurBayes to heterogeneous data, the ability to detect heterogeneity was highly dependent on the disparity between cellularities (Table 1). The proportion of clonal variants also affected detection, with larger values of κ_1 leading to higher mean absolute error (MAE) of the posterior median purity estimates. Although PurityEst performed comparably under certain conditions, the ability for PurBayes to detect heterogeneity generally resulted in greater estimate accuracy.

Our simulation results highlight the potential bias of tumor purity estimates in the presence of unaccounted intratumor heterogeneity. By simultaneously estimating tumor purity and subclonality, PurBayes may also provide additional advantages, such as facilitating inference regarding the tumor composition and evolution as well as isolation of potential founder events. As a Bayesian approach, measures of uncertainty are directly derived from the posterior distribution of λ_J in the form of CIs.

One possible issue in the application of PurBayes is if it estimates J to be larger than the true value because of outlier observations, which leads to a positively biased tumor purity estimate. This can be especially problematic with the existence of copy number variation (CNV) and structural rearrangements. Given that regions of CNV will result in multiplicative impact on the number of mapped reads and SNVs contained within such regions will not truly reflect heterozygosity at a proportion of 0.50, such SNVs would highly influence estimation of λ_J. As such, we anticipate PurityEst to perform better in instances in which CNVs are present and unaccounted for in purity estimation because of its robust estimation procedures. It is thus highly recommended that regions indicated to be CNVs by parallel analyses be filtered from the estimation procedure.

We foresee a variety of extensions to the concepts in PurBayes. For example, the mixture model could be alternatively formulated to characterize tumor cellularity as a continuous distribution using semi-parametric approaches. Integration of CNV and ploidy information will also make PurBayes a more effective estimator.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for the constructive commentary and suggestions.

Funding: This work is supported by the Mayo Foundation and the US National Institute of Health (R21 GM86689; P30 CA168524; P20 GM103418).

Conflict of Interest: none declared.

REFERENCES

Abecasis,G.R. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Degner,J.F. *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

Michor,F. and Polyak,K. (2010) The origins and implications of intratumor heterogeneity. *Cancer Prev. Res. (Phila)*, **3**, 1361–1364.

Plummer,M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria. 2003.

Plummer,M. (2008) Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9**, 523–539.

Song,S. *et al.* (2012) qpure: a tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. *PLoS One*, **7**, e45835.

Su,X. *et al.* (2012) PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, **28**, 2265–2266.

Villa,E.R. and Escobar,L.A. (2006) Using moment generating functions to derive mixture distributions. *Am. Stat.*, **60**, 75–80.