

GENOMICS AND BIOINFORMATICS

Article title

Brian K. Mannakee^{1,*}, Ryan N. Gutenkunst² and Second Co-Author^{2*}

¹University of Arizona Mel and Enid Zuckerman College of Public Health and ²University of Arizona Department of Molecular and Cellular Biology

Received January 1, 2009; Revised February 1, 2009; Accepted March 1, 2009

ABSTRACT

[illegible]

INTRODUCTION

Cancer develops as the result of the accumulation of somatic mutations and clonal selection of cells with mutations that confer a selective advantage on the cell. Understanding the forces that shaped the evolutionary history of a tumor, the mutations that are responsible for its growth, the rate at which mutations are occurring, or how much genetic diversity is likely present in the tumor, requires accurate variant calling, particularly at low variant allele frequency [Williams2016; Bozic2016; Williams2018]. Accurate variant identification is also critical in optimizing the treatment regime for an individual patients disease [Ding2012; Mardis2012; Chen2013; Borad2014; Findlay]. Low frequency mutations present a significant problem for

current mutation calling methods because their signature in the data is difficult to distinguish from the noise introduced by Next Generation Sequencing (NGS), and this problem increases as sequencing depth increases.

Methods for identifying true somatic mutations - i.e. variant calling - from NGS data are an active area of research in bioinformatics. The earliest widely used somatic variant callers aimed specifically at tumors, Mutect1 and VarScan2, used a combination of heuristic filtering and a model of sequencing errors to identify and score potential variants, setting a threshold for that score designed to balance sensitivity and specificity [Koboldt2012; Cibulskis2013]. Subsequent research gave rise to a number of alternate variant calling strategies including haplotype based callers [Garrison2012], joint genotype analysis (SomaticSniper, JointSNVMix2, Seurat, and CaVEMan, MuClone)(????), allele frequency based analysis (Strelka, MuTect, LoFreq, EBCall, deepSNV, LoLoPicker, and MuSE)[Saunders2012; Wilm2012; Shiraishi2013b; Gerstung2012; Zhang2017; Fan2016], and a mixture of ensemble and deep learning methods (MutationSeq, SomaticSeq, SNooPer, and BAYSIC). All of these methods have varying levels of complexity, and some are focused on specific types of data. The one thing they all have in common is that they either implicitly or explicitly assume that the probability of a mutation occurring at a given site is proportional to the overall mutation rate, and the same at every site in the genome.

Single nucleotide substitutions, i.e. simple mutations, arise in tumors at a rate and at genomic locations driven by two main processes. The first is the spontaneous accumulation of mutations that occurs in all dividing tissues, and has a characteristic mutation signature that describes the probability of mutation in a given genomic context [Nik-Zainal2012a; Alexandrov2015; Lee-Six2018]. The second, and far more complex, process is the accumulation of mutations through exposure to mutagens or degradation - via mutation or deletion - of cellular machinery responsible for the identification and repair of damage or replication errors. Many mutagens and DNA repair mechanism defects also have highly specific mutation signatures, such that they can be identified by observing the mutations in the tumor [Alexandrov2013a; Helleday2014a; Nik-Zainal2016; Kandoth2013; Alexandrov2016].

Here we present an empirical bayes method for estimating the prior probability of mutation at a given site using the observed mutation spectrum of the tumor, and show that the addition of this prior to the MuTect variant calling model produces a superior variant classifier in both simulated and real tumor data. We then extend the method

*To whom correspondence should be addressed. Email: rgutenk@email.arizona.edu

with an application of the local false discovery rate by computing the probability that a site is non-null under an assumption of clonal expansion with either early or small selective differences between clones. We provide a simple implementation in R that takes MuTect caller output as input, and returns the posterior probability that a site is variant for every site observed by MuTect.

MATERIALS AND METHODS

Materials subsection one

[illegible]

Materials subsection one. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text:

$$\text{LD}^r = \frac{\text{LD}}{A_{\text{iso}}} = 1.5S \left(3\cos^2\alpha_i - 1 \right) \quad (1)$$

[illegible]

Materials subsection two

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text (see Figure ??).

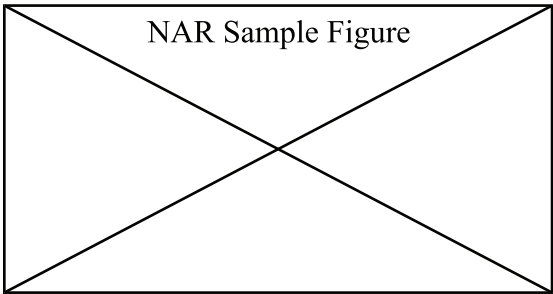


Figure 1. Caption for figure within column.

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text.

$$\text{LD}(t) = \sum_i a_i \exp\left(\frac{-t}{\tau_i}\right)$$

[illegible]

RESULTS

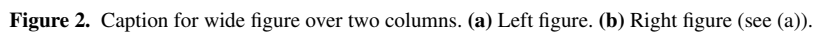
Results subsection one

[illegible]

Table 1. This is a table caption

Col. head 1	Col. head 2 (%)	Col. head 3 (s ⁻¹)	Col. head 4 (%)	Col. head 5 (s ⁻¹)
Row 1	Row 1	Row 1	—	—
Row 2	Row 2	Row 2	Row 2	Row 2

This is a table footnote

[illegible]

Discussion subsection three

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

CONCLUSION

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.

ACKNOWLEDGEMENTS

Text. Text. Text. Text. Text. Text. Text. Text. Text. Text.
Text. Text. Text. Text.

Conflict of interest statement. None declared.