TNER: A Novel Bayesian Background Error Suppression Method for Mutation Detection in Circulating Tumor DNA

Shibing Deng[1], Maruja Lira[2], Stephen Huang[2], Kai Wang[2], Crystal Valdez[2], Jennifer Kinong[2], Paul A Rejto[2], Jadwiga Bienkowska[2], James Hardwick[2] and Tao Xie[2*]

**contact: xietao2000@gmail.com*

[1]Pfizer Early Clinical Development Biostatistics, [2]Pfizer Early Oncology Development and Clinical Research, 10770 Science Center Drive San Diego, CA, USA

Abstract. The use of ultra-deep, next generation sequencing of circulating tumor DNA (ctDNA) holds great promise for early detection of cancer as well as a tool for monitoring disease progression and therapeutic responses. However, the low abundance of ctDNA in the bloodstream coupled with technical errors introduced during library construction and sequencing, complicates mutation detection. To achieve high accuracy of variant calling via better distinguishing low frequency ctDNA mutations from technical errors, we introduce TNER (Tri-Nucleotide Error Reducer), a novel Bayesian background error suppression method that provides a robust estimation of background noise to enhance the specificity for ctDNA mutation detection without sacrificing sensitivity. Results on both simulated and real healthy subjects' data demonstrate that the proposed algorithm consistently outperforms current position-specific models, particularly when the sample of is small. TNER is publicly available at https://github.com/ctDNA/TNER.

1. Introduction

Cancer is a genetic disease that is driven by changes to genes that control cellular function [1]. Characterizing the disease at the molecular level is essential to early detection, personalized therapy based on tumor genomic profile, monitoring tumor progression and response to treatment as well as identification of resistant mechanisms [2]. This typically requires tumor tissue biopsies to obtain samples for genotyping or other molecular analyses for solid tumors. Biopsy procedures are usually invasive, and come with additional risk to patient's health. In many cases, tumor tissue biopsy is contraindicated medically and the tissue samples are often insufficient or unsuitable for molecular profiling [3]. In addition, cancer is a heterogeneous disease with different subclones within the same primary tumor and between the primary tumor and metastatic lesions that can lead to variations in tumor sampling [4,5].

Both cancer and normal cells shed DNA as a result of apoptosis and other biological processes, and release DNA fragments into the blood stream to become cell-free DNA (cfDNA)[6-8]. cfDNA derived from tumor cells is called circulating tumor DNA (ctDNA), and provides a real time genomic snapshot of cancer cells due to the relatively short half-life of cfDNA (~1-2 hours)[2,9]. ctDNA is a form of "liquid biopsy" that provides a non-invasive alternative to tissue biopsy for cancer diagnosis and monitoring [10-12]. Moreover, ctDNA generally comes from all tumor lesions and is pooled in the circulatory system, therefore it can reduce sampling variation associated with tumor heterogeneity than a single tissue biopsy [13].

The fraction of ctDNA in the total cfDNA in plasma, however, can be extremely low in many cancer patients [2,9]. Recently established techniques such as droplet-digital PCR (ddPCR) enable detection and quantification of low abundance ctDNA, but with limitation to cover only a small

number of known "hotspot" mutations [9,14]. Advances in DNA sequencing technology have made it possible to identify ctDNA mutations at comparable sensitivity to ddPCR [15,16] when the sequence coverage is sufficient (>10,000 sequence reads per base). One of the most significant challenges in detecting ctDNA mutations is suppressing technical errors introduced during library preparation, PCR amplification or the sequencing itself [17]. While errors arising during PCR amplification can be removed effectively using molecular barcodes [17], other technical errors are more universal and also need to be removed before mutation calling [3,18]. Newman *et al* [19] recently proposed a creative integrated digital error suppression (iDES) method that included both a molecular barcoding system to reduce PCR errors and a background polishing model with an improved estimation of background mutation error rate (BMER) compared to the previous computational method used in CAPP-Seq [20]. Specifically, the BMER was estimated using a model of Gaussian or Weibull distribution on the mutation data from a collection of healthy subjects [19]. The method increased the accuracy for variant detection significantly, but the relatively small sample size (n=12) and the nature of the data (small discrete count) made it difficult to robustly estimate the background. As a result, false positive errors are still observed in a substantial number of bases [19].

To provide a more robust estimation of background error to further reduce the false positive rate of variant calling, we developed a novel background polishing method called TNER (Tri-Nucleotide Error Reducer) with a Bayesian approach to overcome the small sample size of healthy subjects. TNER is based on tri-nucleotide context data and uses a binomial distribution for the mutation error to estimate the background from healthy subjects. Tri-nucleotide context, or to be more exact, the 96 tri-nucleotide contexts consisting of the 6 distinguishable single nucleotide mutations and the 16 possible combinations of immediately preceding and following bases, has been extensively studied in cancer genetics to construct mutation signatures as a response to carcinogens, to compare the mutational spectra of trunk and branch mutations, or to predict the clinical implications of called mutations [5,21-23]. Given that the pattern of low frequency technical errors from next-generation sequencing (NGS) should be similar in normal control samples and patient samples, we argue that local sequence context could better model noise for small sample size by leveraging information from other bases with a shared tri-nucleotide context. In this paper, we describe the Bayesian background polishing model based on tri-nucleotide contexts and demonstrate that TNER reduces errors more effectively than the position specific error model when there are a limited number of healthy donor samples for modeling the background error. The TNER methodology proposed here, to the best of our knowledge, is novel in this area. TNER is publicly available at https://github.com/ctDNA/TNER.

2. Methods

2.1. *NGS data for analysis*

To demonstrate the performance of the error suppression model in detecting single nucleotide variations, we analyzed targeted sequencing data of plasma cfDNA of healthy subjects using a panel of 87 cancer genes described previously (Lira et al. AACR 2017, #2749). The barcoded target-enriched DNA library (147Kb) was sequenced on an Illumina HiSeq 4000 platform generating ultra-deep coverage with an average coverage per base of ~12,000.

2.2. *Tri-nucleotide error reduction model*

The detection of ctDNA is typically achieved through detecting signature mutations associated with tumors in cfDNA. Sequence data from cfDNA has many stereotypical errors or other background mutation errors that are not of tumor origin [24]. In order to call a mutation in ctDNA, the distribution of the BMER needs to be characterized at each nucleotide base position to reduce the false positive error, for example, by modeling cfDNA data on the same NGS panel from healthy subjects [19]. The mutation rates from healthy subjects are assumed to be background mutation noise associated with both technical and biological sources. One challenge in characterizing the individual nucleotide BMER from healthy subjects is the relatively small cohort size. The iDES method used 12 healthy subjects [19]; we used a comparably sized set of 14 healthy subjects. These small sample sizes do not allow building a reliable estimate of the background error distribution for individual nucleotides. Bayesian method with prior information can help to overcome this limitation.

To better estimate the BMER distribution, we propose a background error model based on a hierarchical Bayesian method that utilizes the distribution of mutation error rate in a tri-nucleotide context, which consists of the mutated nucleotide and the combinations of immediately preceding and following nucleotides. Mutation signatures characterized by tri-nucleotide context have been used frequently in cancergenetics [5,21,22] . There are 96 independent tri-nucleotides contexts. For a nucleotide in tri-nucleotide group i (i=1, …, 96) at base position j (j=1,…J), the number of reads X_{ij} observed for a given coverage N_{j} is assumed to follow a binomial distribution

$$X_{\mathrm{ij}} \sim \mathrm{Binom}(N_{\mathrm{j}}, \pi_{\mathrm{ij}}) \tag{1}$$

with position specific mutation rate parameter π_{ij}. J is the total number of bases in the panel (147k). With a large N (>1000 typically) and a small π(<1%), X can also be modeled as a Poisson distribution

$$X_{\mathrm{ij}} \sim \mathrm{Pois}(N_{\mathrm{j}} * \pi_{\mathrm{ij}}) \tag{2}$$

with rate parameter $N_j * \pi_{ij}$.

The BMER at position j can be estimated using the average mutation rate of the jth nucleotide from the 14 healthy subjects, $\hat{\pi}_{\mathrm{ij}}$. This position specific parameter will be poorly estimated with the small sample size. To improve the estimate of π, we propose a Bayesian framework and assume that π follows a beta distribution within a tri-nucleotide

$$\pi \sim \mathrm{Beta}(\alpha, \beta) \tag{3}$$

For convenience, we re-parametrize the beta distribution using its mean as a parameter.

$$\pi \sim \mathrm{Beta}(\mu, \omega), \text{ with } \mu = \frac{\alpha}{\alpha+\beta} \text{ and } \omega = \alpha + \beta \tag{4}$$

We assume the BMER in a tri-nucleotide follows a beta distribution. The prior parameters of the beta distribution can be estimated based on the BMER distribution of nucleotides in the tri-nucleotide using method of moments [25]. The mean parameter μ can be estimated by the average mutation rate ($\hat{\mu}$) of nucleotides in the tri-nucleotide. For a position with x mutation count out of n

total reads, the posterior distribution of the BMER at this position will be a $\mathrm{Beta}(\alpha + x, \beta + \mathrm{n} - x)$ with a mean parameter

$$\mu' = \frac{\alpha + x}{\alpha + \beta + n} = w\mu + (1 - w)x/n \tag{5}$$

with w=(α+β)/(α+β+n). The posterior mean can be estimated with a shrinkage estimator, that is, a weighted average of the tri-nucleotide level mutation rate ($\hat{\mu}$) and the position specific rate $\hat{\pi}_{ij}$.

$$\tilde{\pi}_{ij} = w_{ij}\hat{\mu}_i + (1 - w_{ij})\hat{\pi}_{ij} \tag{6}$$

The w_{ij} is a simplified weight that balances the relative size of the tri-nucleotide mutation error rate and the position specific mutation error rate works as well

$$w_{ij} = \frac{\hat{\mu}_i}{\hat{\mu}_i + \hat{\pi}_{ij}} \tag{7}$$

This weight function provides less shrinkage when the position specific mutation error rate is high - a property that helps retain the position specific background when it is much higher than the tri-nucleotide level background. However it does not reflect the impact of sample size on the weight, therefore is a more heuristic approach than the full Bayesian method.

Once we have an estimate of the BMER π using eq. (6), the threshold for mutation detection can be defined based on the upper confidence limit of π. At α level, the upper 1-α/2 Clopper-Pearson confidence limit [11] for a binomial proportion parameter is

$$B_{ij} = \beta(1 - \frac{\alpha}{2}, x_{ij} + 1, N_j - x_{ij}) \tag{8}$$

where $\beta()$ is the quantile function of beta distribution; x_{ij} is the mutation read count at position j and N_j is the total reads for this position. In the Bayesian model, we don't have to worry about the multiple comparison correction and the prior distribution allows pooling information between positions and avoids false positive call when variation is low [26]. In our analysis, the false positive calls are very rare. Similar beta-binomial model has been used in other studies [27-29]. However, none of them used the model to estimate the BMER distribution with tri-nucleotide contexts, nor did they apply to ctDNA NGS data.

3. Results

We first evaluated the TNER model on the healthy subject data using the leave-one-out method. We built the background model using data from 13 healthy subjects and predicted the mutation in the leave-out subject. Similar to Newman et al [19], we counted the number of error free positions, defined as those positions with exclusively reference allele reads, for each of the 14 healthy subjects at all 147k nucleotide positions, and compared the different error suppression methods, including the background polishing from iDES and the TNER method (Figure 1). For TNER we used α=0.01, although the results were similar for α=0.05. We also calculated the panel-wide

error rate which is defined as the number of non-reference allele reads (frequency<5%, to exclude SNPs) divided by the total reads.

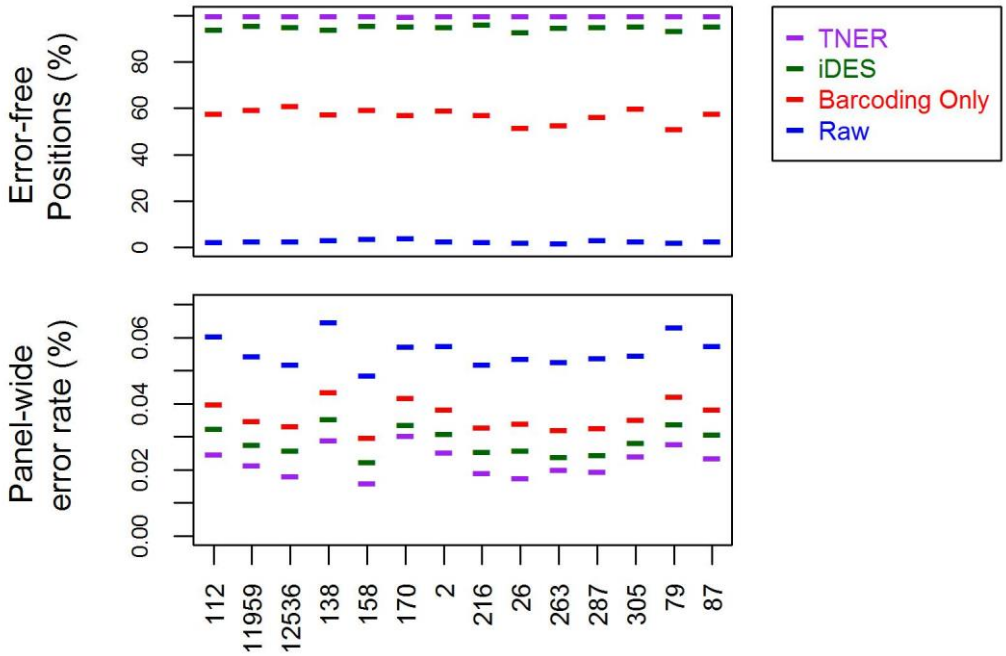

Figure 1. Error free position (%) and panel-wide error rate of the 14 healthy subjects' data (sample labels on x-axis) from the leave-one-out analysis with different methods. Raw = raw data, Barcoding Only = Barcoding error reduction only. The tri-nucleotide error reduction (TNER) method provides the lowest error rate in the healthy subjects, and demonstrates its superior specificity in reducing false positive error.

To test the sensitivity of the method, we used data from three healthy subjects that were not part of the background cohort. One subject had 10 unique private SNPs that were not shared by any of the healthy subjects. We did an in-silico experiment to dilute this subject's data with the other two health subjects in a 1:250:250 ratio and assume heterozygosity so we have an expected allele frequency of 0.1% for the 10 private SNPs. We found both iDES and TNER (α=0.01) were able to detect all the 10 SNPs in this experiment.

To compare the performance of the position specific background polishing method and the TNER method more rigorously, we evaluated their sensitivity and specificity at various detection thresholds using simulation studies. The simulation used the average position specific mutation error rate from the 14 healthy subjects as the BMER which is a matrix of 147k rows and four columns. Each column is a nucleotide the reference base can mutate to, including the reference nucleotide which is zero. We randomly selected 1000 bases (rows) out of the 147k total, and then at each of the selected base a simulated allele frequency (simulated signal) was added to the existing BMER of a selected non-reference nucleotide (column). Specifically, for each of the 1000 positions, there are three possible non-reference nucleotides it can mutate to. We chose the nucleotide with the largest BMER value as the selected nucleotide to add the simulated signal. If the BMER are all zeros at this position, we used the first non-reference letter (A-C-T-G) as the selected nucleotide to add signal. This updated BMER matrix is the same as the original matrix

except that 1000 rows have a signal added to a selected column, with the updated BMER matrix we simulated the read counts with a total coverage of 10000 per position using a binomial and a normal distribution. For a normal distribution, the standard deviation is the square root of the BMER divided by 100. The simulated counts were further split into forward and reverse strand with a random forward to reverse strand ratio centered around 1. The position specific Gaussian models from the iDES and the TNER method were then separately applied to the simulated data. As the true positives and true negatives are known, sensitivity and specificity were calculated under various detection thresholds (α values). Receiver operating characteristic (ROC) curves in Figure 2 compare the two methods under different scenarios. The TNER method performed better than the position specific Gaussian model in all cases – data simulated under different distributions and different mutation rate (MR) as shown by ROC curves.

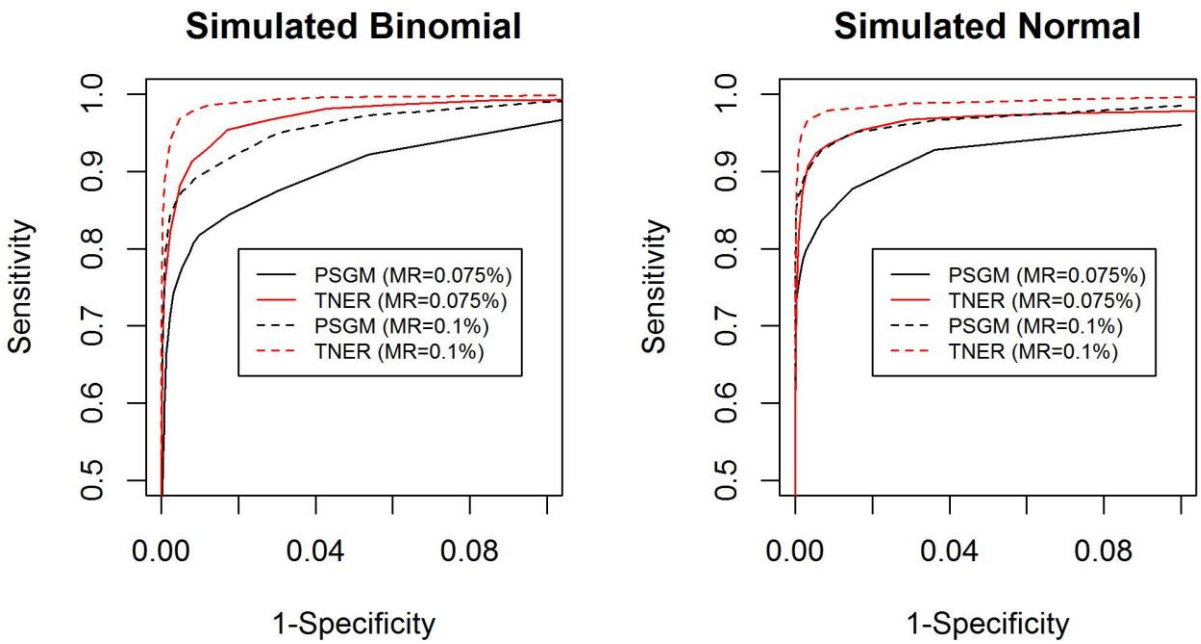

Figure 2. ROC curves for position specific Gaussian model (PSGM) (black) and TNER (red) methods in simulated cfDNA data. Two mutation rates (MR) were simulated: 0.075% (solid line) and 0.1% (dashed line) with a total coverage of 10000 at each position.

One of the advantages of the TNER method is that it uses information from other positions with the same tri-nucleotide context through a Bayesian model and thus stabilizes parameter estimates of the BMER. Therefore we would expect TNER performs better than position specific error models when the available sample size for healthy subjects becomes small. To evaluate the effect of healthy subject sample size on the performance of mutation detection methods, we used half the available healthy subjects (n=7) as our background mutation estimate and compared the results from both position specific Gaussian model and TNER in the simulation studies. As expected, we found that the smaller sample size of healthy subjects did not substantially reduce the performance of TNER, but greatly reduced the performance of the position specific Gaussian

method (Figure 3). This illustrates the robustness of the TNER method when the number of healthy subjects is small.

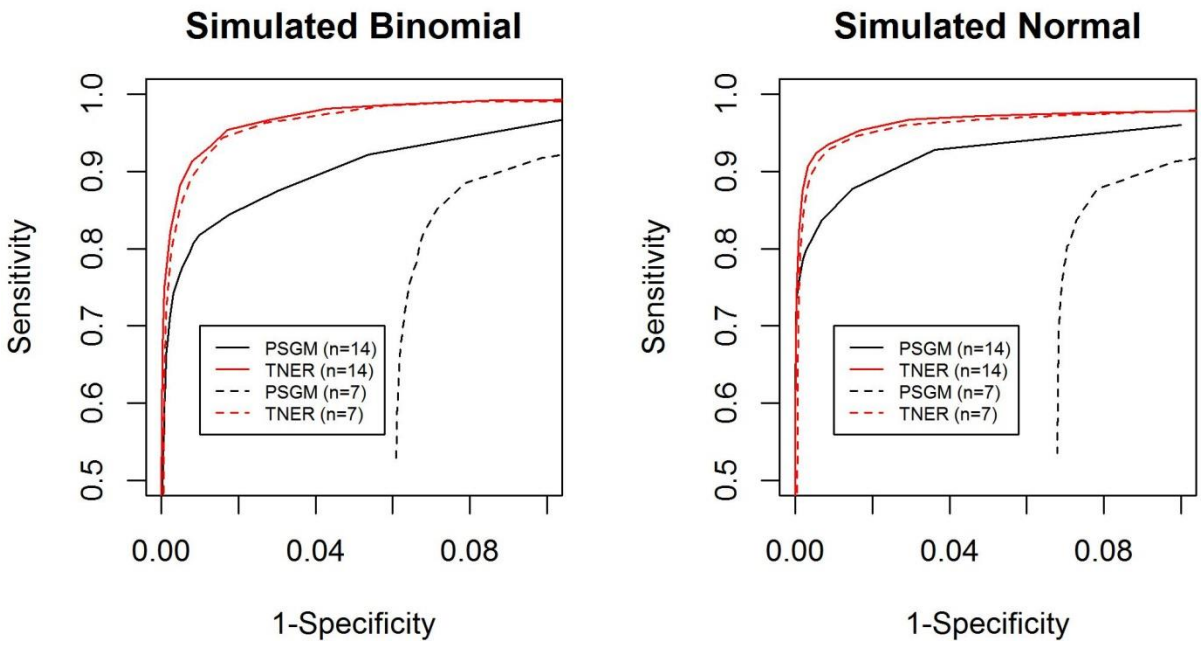

Figure 3. ROC curves of the position specific Gaussian model (PSGM) (black) and the TNER (red) methods with different input numbers of healthy subjects: n=7 (dashed line) and n=14 (solid line). Mutation rate used is 0.075%.

4. Discussion

We propose a new background polishing method for the detection of ctDNA in liquid biopsy samples based on tri-nucleotide level data. The TNER method estimates the background mutation errors from healthy subjects using a hierarchical model to incorporate both the tri-nucleotide level mutation rate and position specific mutation rate. The additional information from tri-nucleotide level data helps stabilize the estimate of background errors and proves to be more robust when the number of healthy subjects is small.

We could have used dinucleotide context or more complicated local sequence context such as pentanucleotide (2 flanking nucleotides on each side) or heptanucleotide (3 flanking nucleotides on each side). The larger local sequence context may provide better model fit to the mutation error rate [30], but the increasing model complexity with pentanucleotide (1,536 unique contexts) and heptanucleotide (24,576 unique contexts) becomes impractical for a targeted panel, like the one tested here with a total of 147k bases. The Bayesian prior parameter will not be well estimated due to small number of bases within each context. Trinucleotide context provided a better fit than dinucleotide [31] but not too complicated than the larger local sequence context [30], so is a more balanced approach for a common NGS targeted panel.

One of the assumptions for analyzing NGS data by TNER is that individual nucleotides within a tri-nucleotide share a more similar mutation error rate than those between-trinucleotides. We looked at the average mutation error rate from healthy subjects at the tri-nucleotide level and compared the intra-trinucleotide variability and the inter-trinucleotide variability. 94% of tri-nucleotides have intra-trinucleotide variability smaller than the inter-trinucleotide variability. Figure 4 displays an example of three tri-nucleotides, all with C to T mutation, showing very different distribution. The dashed lines are fit of beta distribution using the parameter estimates calculated by the method of moment. In general beta distribution fits the intra-trinucleotide mutation error rate very well.

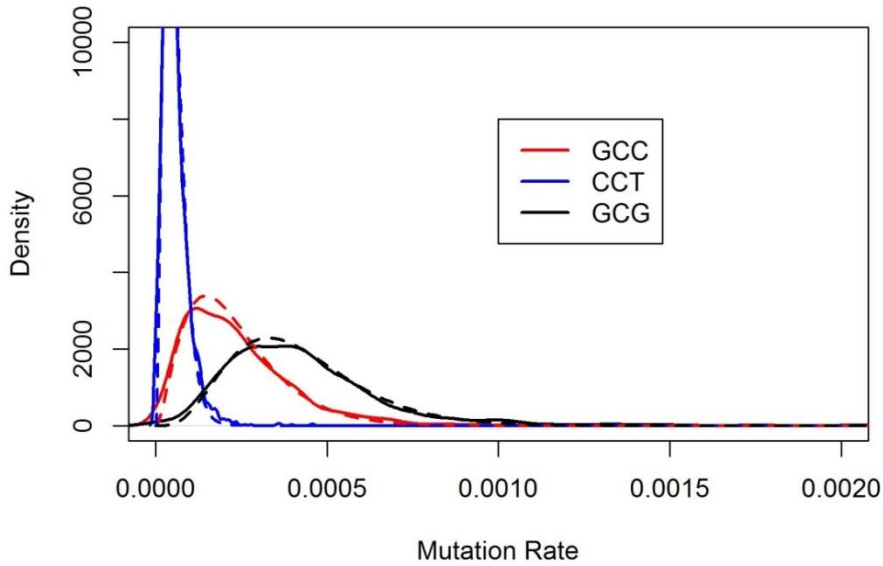

Figure 4. Examples of mutation error rate distribution of tri-nucleotides with C-T mutation. Solid lines are the probability density of average position specific error rate within a tri-nucleotide. The dashed lines are corresponding fit of a beta distribution using estimated parameters from the data.

In genomic data analysis, when the sample size is small, it is common to analyze data for individual genes using information from other genes. This is implemented in the *limma* method [32] for microarray data analysis and the DESeq method [33] for RNAseq data analysis. In our approach, we take advantage of the large number of bases shared in the same nucleotide context and use these data to model the individual base mutation error rate. We found the TNER method improves the imprecise estimate of background associated with small sample size at individual base level.

Sequence data are read counts that are best described by distribution from discrete data families, such as Poisson distribution or binomial distribution, particularly when the read count is low and mutation rate is small such as in ctDNA data. We found that the Poisson distribution fit the count data well in general. A more sophisticated distributions taking over-dispersion into consideration and the zero-inflated nature of ctDNA data may further improve the method.

Currently, ctDNA is rapidly becoming established as an important tool to supplement conventional biopsies for cancer early detection, molecular characterization and monitoring of

tumor dynamics and the TNER method provides a novel Baysian approach to improve the accuracy of mutation detection in ctDNA.

References

1. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646-674.
2. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, et al. (2008) Circulating mutant DNA to assess tumor dynamics. Nat Med 14: 985-990.
3. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108: 9530-9535.
4. Venesio T, Siravegna G, Bardelli A, Sapino A (2017) Liquid Biopsies for Monitoring Temporal Genomic Heterogeneity in Breast and Colon Cancers. Pathobiology.
5. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 366: 883-892.
6. Underhill HR, Kitzman JO, Hellwig S, Welker NC, Daza R, et al. (2016) Fragment Length of Circulating Tumor DNA. PLoS Genet 12: e1006162.
7. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J (2016) Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell 164: 57-68.
8. Jiang P, Chan CW, Chan KC, Cheng SH, Wong J, et al. (2015) Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci U S A 112: E1317-1325.
9. Volik S, Alcaide M, Morin RD, Collins C (2016) Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies. Mol Cancer Res 14: 898-908.
10. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, et al. (2014) Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 6: 224ra224.
11. Lippman M, Osborne CK (2013) Circulating tumor DNA--ready for prime time? N Engl J Med 368: 1249-1250.
12. Schwarzenbach H, Hoon DS, Pantel K (2011) Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer 11: 426-437.
13. Crowley E, Di Nicolantonio F, Loupakis F, Bardelli A (2013) Liquid biopsy: monitoring cancer-genetics in the blood. Nat Rev Clin Oncol 10: 472-484.
14. Openshaw MR, Page K, Fernandez-Garcia D, Guttery D, Shaw JA (2016) The role of ctDNA detection and the potential of the liquid biopsy for breast cancer monitoring. Expert Rev Mol Diagn 16: 751-755.
15. Chen YH, Hancock BA, Solzak JP, Brinza D, Scafe C, et al. (2017) Next-generation sequencing of circulating tumor DNA to predict recurrence in triple-negative breast cancer patients with residual disease after neoadjuvant chemotherapy. NPJ Breast Cancer 3: 24.
16. Thierry AR, Mouliere F, El Messaoudi S, Mollevi C, Lopez-Crapez E, et al. (2014) Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. Nat Med 20: 430-435.

17. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, et al. (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39: e90.
18. Kirsch S, Klein CA (2012) Sequence error storms and the landscape of mutations in cancer. Proc Natl Acad Sci U S A 109: 14289-14290.
19. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, et al. (2016) Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 34: 547-555.
20. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, et al. (2014) An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med 20: 548-554.
21. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al. (2013) Signatures of mutational processes in human cancer. Nature 500: 415-421.
22. Yang Z, Jones A, Widschwendter M, Teschendorff AE (2015) An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. Genome Biol 16: 140.
23. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C (2016) DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 17: 31.
24. Park G, Park JK, Shin SH, Jeon HJ, Kim NKD, et al. (2017) Characterization of background noise in capture-based targeted sequencing data. Genome Biol 18: 136.
25. Bowman K, Shenton L (1998) Estimator: Method of Moments. Encyclopedia of statistical sciences: 8.
26. Gelman A, Hill J, Yajima M (2012) Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness 5: 22.
27. Gerstung M, Papaemmanuil E, Campbell PJ (2014) Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics 30: 1198-1204.
28. He Y, Zhang F, Flaherty P (2015) RVD2: an ultra-sensitive variant detection model for low-depth heterogeneous next-generation sequencing data. Bioinformatics 31: 2785-2793.
29. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, et al. (2015) Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348: 880-886.
30. Aggarwala V, Voight BF (2016) An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet 48: 349-355.
31. Chen T, Wang Z, Zhou W, Chong Z, Meric-Bernstam F, et al. (2016) Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types. BMC Genomics 17 Suppl 2: 394.
32. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.
33. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.