bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

On statistical modeling of sequencing noise in high depth data to assess tumor evolution

Raul Rabadan[1,2], Gyan Bhanot[3], Sonia Marsilio[4], Nicholas Chiorazzi[4], Laura Pasqualucci[5], and Hossein Khiabanian[*2,6]

[1]Department of Systems Biology, Columbia University, New York, NY
[2]Center for Topology of Cancer Evolution and Heterogeneity, Columbia University, New York, NY
[3]Department of Physics and Astronomy, Rutgers University, Piscataway, NJ
[4]The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY
[4]Institute for Cancer Genetics, Columbia University, New York, NY
[5]Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ

Abstract

One cause of cancer mortality is tumor evolution to therapy-resistant disease. First line therapy often targets the dominant clone, and drug resistance can emerges from preexisting clones that gain fitness through therapy-induced natural selection. Such mutations may be identified using targeted sequencing assays by analysis of noise in high-depth data. Here, we develop a comprehensive, unbiased model for sequencing error background. We find that noise in sufficiently deep DNA sequencing data can be approximated by aggregating negative binomial distributions. Mutations with frequencies above noise may have prognostic value. We evaluate our model with simulated exponentially expanded populations as well as data from cell line and patient sample dilution experiments, demonstrating its utility in prognosticating tumor progression. Our results may have the potential to identify significant mutations that can cause recurrence. These results are relevant in the pretreatment clinical setting to determine appropriate therapy and prepare for potential recurrence pretreatment.

1 Introduction

Every extant organism is the result of over three billion years of evolution. Complex organisms consist of cells whose functions are regulated by a large number of interconnected pathways that ensure cellular, tissue, and organ homeostasis. Cancer is a result of the breakdown of this process in a single cell, which results in its unregulated growth. In most cases, the immune system is able to detect and eliminate such aberrant cells. Sometimes, however, a clone escapes this surveillance and manifests as clinically detectable disease [40]. Consequently, most clinically diagnosable tumors are clonal, i.e. they grow clonally from a single cell that finds a path to circumvent the body's defense mechanisms. The growing tumor accumulates mutations, most of which have low fitness and therefore are found at low frequencies, outcompeted by the dominant clone.

*Correspondance: h.khiabanian@rutgers.edu

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The clonal expansion process, which underlies genomic diversification within a tumor, was first studied by Salvador Luria and Max Delbrück. They designed a simple system of single-cell organisms to investigate patterns of mutation accumulation. Their rigorous quantitative methodology led them to discover that mutations arise randomly and their numbers follow a distinct probability distribution[24]. As the cell population in the tumor diversifies, it is able to explore the fitness landscape. Studying the dynamics of this genomic heterogeneity can yield insight into when the clonal expansion started, how fast the population evolved, and whether specific genomic alterations were selected in a particular host or under a treatment regimen.

The principal biochemical mechanisms in cancer are often recurrent across tumors in different tissues. For example, aberrations leading to unregulated cell growth or inactivation of the apoptotic pathway (cell suicide) are common to almost all tumors. Given the limits within which cells are regulated, the growing tumor has access to only a finite number of pathways that it can alter. As a result, tumors arising from different cells of origin often harbor identical genetic mutations, which alter the same pathways, and often have similar prognostic consequences [5].

First line therapy drugs target a tumor's dominant, fastest growing clone. Drug resistance often emerges from the rise of preexisting clones that harbor potential driver mutations that gain evolutionary fitness via therapy-induced natural selection. It has been shown that the presence of drug-resistant sub-clones in the primary tumor prior to therapy may be a strong predictor of poor survival, with direct implications for disease management [28,34,37,44]. As cancer therapy moves towards individualized treatment, it is important to identify and understand the role of such mutations, some of which may have prognostic value. Such potentially prognostic mutations are commonly identified using targeted deep sequencing of the tumor DNA in clinical settings, and their sensitive detection relies on the accurate analysis of background noise, specifically DNA sequencing errors.

Studying the evolution of chronic lymphocytic leukemia (CLL) under therapy is an illuminating example of these approaches [19,20]. CLL is the most common leukemia in adults and its clinical course ranges from asymptomatic disease that never requires therapy to rapidly progressive disease that requires intensive treatment. Genomic alterations in CLL follow a time ordered process [45]. Patients who harbor genomic defects in the *TP53* gene, which regulates many pathways including the cell suicide or apoptotic pathway, are considered at high risk of failing conventional therapies [35]. Such patients are good candidates for stem cell transplant or new gene-specific therapeutics [2,39]. The presence of such secondary mutations in genes such as *TP53* is often assessed using traditional Sanger sequencing that only provides sufficient power to detect mutations present in at least 20% of leukemia cells [32]. To assess the presence of *TP53* prognostic mutations at lower abundances in newly diagnosed CLL patients, we used deep sequencing and evaluated thousands of leukemia cells and identified small *TP53* mutations that were missed by traditional methods such as Sanger sequencing [34]. We found that *TP53* mutated sub-clones identified before treatment became the predominant population at the time of CLL relapse, as a result of therapy induced selection pressure. These results suggest that tumors harboring small *TP53* mutations have the same clinical phenotype and risk of failing therapy as those with *TP53* defects in the dominant clone [27,34], and their early detection is essential for the identification and management of high-risk CLL patients [11].

These results are also pertinent to other hematological malignancies where the presence of leukemia-associated mutations in remission is associated with significantly increased risk of relapse and poor survival[31,37]. These data lead to the conclusion that it is imperative to identify alterations

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that induce therapeutic resistance in leukemia patients in the early stages of disease in order to properly guide individualized therapy with the goal of preventing disease relapse. However, the detection of mutations at low allele frequencies (e.g., 1 mutation in 10,000 cells) is hindered by the lack of a precise model of noise in diagnostic sequencing assays.

Targeted sequencing is the most commonly used method to track prognostic markers in both clinical and basic research applications [10]. However, finding such mutations in sequencing reads is often confounded by misreading a base in the sequencing instrument or mis-incorporation of DNA bases (nucleotides) during library enrichment by polymerase chain reaction (PCR) amplification cycles. More accurate sequencing protocols, which perform overlapping reads of the same genomic DNA region, allows the merging of such reads for improved accuracy. This facilitates correcting errors accumulated in the sequencer, while leaving uncorrected PCR errors that arise during library preparation steps [4,46].

The challenge in identifying potentially functional sub-dominant mutations is to determine the sensitivity thresholds of sequencing platforms, i.e. the depths above which PCR errors happen with a probability below a statistical cut-off. Such thresholds can be estimated by hypothesizing that all variants are due to errors and using deviations from this null hypothesis to indicate the presence of true variants. This can sometimes be confounded by the fact that different sequencing errors occur at different rates [3,6]. Hence a single threshold cannot comprehensively test the significance of all variants. As a result, more sophisticated statistical modeling of the background error distribution is necessary.

To model background error one may use different types of error distributions: (i) a single or a linear combination of Luria-Delbrück distributions, characterizing the expected number of spontaneous mutations during tumor growth, where the PCR error rate is assumed to be constant [17]; (ii) the negative binomial distribution, describing the depth distribution of clones after PCR amplification through a Poisson-Gamma mixture model [29]; and iii) the beta-binomial distribution, suitable for Bayesian models, where error rates are assumed to follow the Beta distribution [21]. Although the Luria-Delbrück distribution is expected to better describe the long tail of the error depths, empirical analysis has shown that the negative binomial distribution gives the best fit to the observed error depths based on goodness-of-fit log-likelihood [34]. The beta-binomial distribution, in conjunction with multiple filtering criteria based on normal control DNA samples, has also been proposed for somatic mutation detection from cancer genomes [8,9,23,36].

In this manuscript, we revisit this problem and provide a comprehensive model that illustrates how aggregate negative binomial distributions describe PCR error depths in ultra-deep targeted sequencing. We test our model with *in silico* as well as cell line and patient dilution experiments, and propose a highly sensitive, mutation-specific approach to detect true mutations, without the need for control data from un-mutated (wild type) normal tissue DNA.

2 Methods

Derivation of the error depth distribution. Here we will only be discussing the distribution of low frequency errors in deep DNA sequencing analysis of tumor samples. Let us assume an experiment in which S independent tissue samples are subjected to ultra-deep sequencing. DNA sequencing of tumor samples produces strings of nucleotides (A, C, G, and T) of 100-200 base-pair length that correspond to the DNA sequences of different sections of the genome in the tumor sample. These sequences of DNA reads are mapped to a "reference" genome and deviations/mismatches

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

are identified as potential mutations. Ideally, the reference sequence is the sequence from the patient's "germ-line", usually obtained from blood or some other tissue with normal cells. The sequencing read depth is the average number of reads that map to the same locus (section of the genome). At a nucleotide, three potential single base substitutions can occur: A (adenine) → C, G, T, or C (cytosine) → A, G, T, or G (guanine) → A, C, T, or T (thymine) → A, C, G. Alternately, there might be an insertion (addition of one or more A, C, G, T nucleotides) or a deletion (loss of A, C, G, T nucleotides). All of these will henceforth be referred to as variants. We want to derive the posterior probability distribution for these variants, assuming they are stochastic, i.e. they represent noise (statistical random errors).

Suppose that, at a genomic DNA locus, we see n_i such variant reads amongst N_i total reads. The distribution of n_i follows a binomial distribution, $\text{Bino}(n_i|N_i,\theta)$, where θ is the *a priori* probability of a variant's occurrence. Let $M = \sum_{i\neq j}^{S} N_i$ be the total number of reads across samples at that locus and $m = \sum_{i\neq j}^{S} n_i$ be the total number of variant (erroneous) reads across samples at that DNA locus. Then, the posterior predictive p value for having detected a true mutation in sample j, given $S-1$ other samples, can be obtained from the posterior probability distribution:

$$\begin{aligned} P(n_j|N_j,\{n_i,N_i\}) &= \int_0^1 \frac{\text{Bino}(n_j|N_j,\theta)\prod_{i\neq j}\text{Bino}(n_i|N_i,\theta)}{\int_0^1 \prod_{i\neq j}\text{Bino}(n_i|N_i,\theta')\mathrm{d}\theta'}\mathrm{d}\theta' \\ &= \binom{N_j}{n_j} \times \int_0^1 \frac{\theta^{n_j+m}(1-\theta)^{N_j-n_j+M-m}}{\int_0^1 \theta^m(1-\theta)^{M-m}\mathrm{d}\theta}\mathrm{d}\theta \\ &= \binom{N_j}{n_j} \times \frac{\text{Beta}(1+n_j+m, 1+N_j-n_j+M-m)}{\text{Beta}(1+m, 1+M-m)}, \end{aligned}$$

where Beta indicates the Beta function. Simplifying the algebra yields the beta-binomial distribution,

$$P(n_j|N_j,m,M) = \frac{1+M}{1+N_j+M}\frac{\binom{N_j}{n_j}\binom{M}{m}}{\binom{N_j+M}{n_j+m}}. \tag{1}$$

Variations of equation (1) have been previously derived for sequencing depths > 100× [8,9,36]. Today, it is possible to do ultra-deep sequencing, where $N_i > 5{,}000\times$. In such cases, for low frequency variants, we can assume that $n_i \ll N_i$. Therefore, we can use Stirling's approximation, and estimate $\binom{N_i}{n_i} \approx \frac{N_i^{n_i}}{n_i!}$. Equation (1) can then be approximated by

$$P(n_j|N_j,m,M) = \binom{n_j+m}{n_j}\left(\frac{N_j}{N_j+M}\right)^{n_j}\left(\frac{M}{N_j+M}\right)^{m+1}, \tag{2}$$

which equals $\text{NB}(n_j|1+m, \frac{N_j}{N_j+M})$, with NB indicating the negative binomial distribution, and where $1+m$ and $\frac{N_j}{N_j+M}$ are its two parameters, which we can interpret as the number of detected errors and the *a priori* probability of success in detecting an error, respectively.

Exponential expansions at varying error rates. An exponentially expanded population is generated through c PCR amplification cycles, where each cycle doubles the DNA population. If errors accumulate independently at a rate of μ substitutions per site per cycle, the average error

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

depth (i.e. the average number of reads harboring errors) is $2^c\mu$. For S such populations, the error depth distribution is described by equation (1), or is approximated by a negative binomial distribution, $\mathrm{NB}(1+(S-1)2^c\mu, \frac{1}{S})$, as derived above in equation (2).

It is well known that different types of PCR errors occur at different rates. For example, transitions, that exchange two-ring purines (A and G) or one-ring pyrimidines (C and T) are more common than transversions, which replace an A or G with one of C or T. Assuming R independent rates, the observed number of variants $D(v)$, with error depth v is then given by,

$$\begin{aligned} D(v) &= \sum_{r=1}^{R} X_r P(v|2^c, (S-1)2^c\mu_r, (S-1)2^c) \\ &\approx \sum_{r=1}^{R} X_r \mathrm{NB}(v|1+(S-1)2^c\mu_r, \frac{1}{S}), \end{aligned} \tag{3}$$

where X_r represents the number of variants that occur with rate μ_r. Since error rates are often unknown and sequence context dependent, we can alternatively bin the variants based on their average error depth across samples and write $D(v)$ as

$$\begin{aligned} D(v) &= \sum_{b=1}^{B} X_b P(v|\langle N\rangle, (S-1)\langle v\rangle_b, (S-1)\langle N\rangle) \\ &\approx \sum_{b=1}^{B} X_b \mathrm{NB}(v|1+(S-1)\langle v\rangle_b, \frac{1}{S}), \end{aligned} \tag{4}$$

where B is the number of bins, X_b is the number of variants in each bin, and $\langle N\rangle$ is the average sequencing depth across S samples. It has been shown that the sum of negative binomial distributions with equal success probabilities is also a negative binomial distribution, though with a random parameter [7,43]. Thus, the approximation of $D(v)$ in equations (3) and (4) with sums of negative binomial distributions that have success probability of $\frac{1}{S}$, suggests empirical observations [34].

3 Data

In the first experiment, a series of dilutions was generated using the SU-DHL-6 cell line (Diffuse Large B-Cell Lymphoma), which carries a heterozygous (one allele altered) *TP53*-Y234C missense mutation (one that changes an amino acid in a protein sequence) [26]. The cells were serially diluted at (1:10, 1:10^2, 1:10^3, 5:10^4, 1:10^4, 5:10^5, and 1:10^5) by mixing the cell line DNA with *TP53* wild-type genomic DNA from a healthy donor. The *TP53* mutation locus was sequenced at depths of 10,000× (10K×), 100,000× (100K×), and 1,000,000× (1M×).

In the second experiment, genomic samples from 18 healthy individuals as well as samples from undiluted and 1:10^3 diluted cancer cells from a CLL patient, harboring a heterozygous *SF3B1*-K700E missense transition substitution were analyzed and the *SF3B1* mutated locus was sequenced at a mean depth of 620,000×.

For both experiments, each cell line dilution and patient sample was barcoded and targeted with amplicon multiplexed sequencing using the Illumina MiSeq (2 × 150 bp) (Genewiz, South Plainfield, NJ). The primers were designed so that the paired-end reads substantially overlapped

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

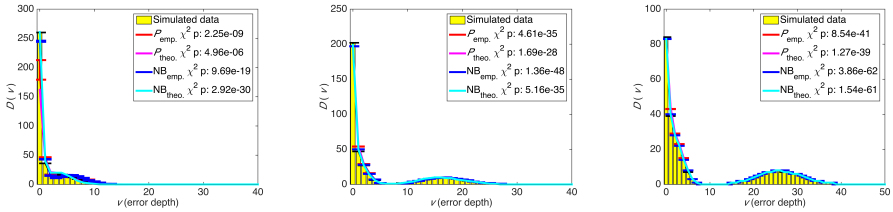

Figure 1: Number of variants with error depth of v from aggregated simulated cycles of PCR amplification at four error rates: 12 cycles (left), 14 cycles (middle), and 18 cycles (right). $P_{\text{theo.}}$ and $\text{NB}_{\text{theo.}}$ are calculated using equation (3), and $P_{\text{emp.}}$ and $\text{NB}_{\text{emp.}}$ are calculates using equation (4). The χ^2 test was used to compare the distributions.

with each other and each read pair was merged to correct sequencing errors. The merged reads were mapped to the human reference genome (hg19) using the Burrows-Wheeler Aligner (BWA) alignment tool[22], and all variable sites were identified using an inclusive variant caller, adapted from the SAVI algorithm[41].

4 Results

Simulated data. We generated a set of *in silico* experiments with exponentially expanded populations starting from a single, homogenous, 100 base-long sequence of binary bases. Each population was aggregated from four expansions that followed error rates of 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6} substitutions per site per cycle. The number 12, 14, and 18 of cycles were chosen to produce populations with 16,384, 65,536 and 1,048,576 total reads respectively. Each experiment contained 50 independent populations ($S = 50$) and for each experiment, $D(v)$, the expected number of variants with depth v was calculated using equations (3). This experiment was repeated 100 times. Figure 1 shows the results, as well as statistically significant χ^2 p values indicating high accuracy of the estimates from both the beta-binomial model and its NB approximation.

Dilution experiments. We removed the real diluted *TP53* mutation from cell line sequencing data, and arranged the erroneous variants based on their depth in 5×-sized bins. We then counted the number of variants X_b in each bin, and calculated $D(v)$ using equation (4). Figures 2, 3, and 4 show the results for sequencing depths of 10K×, 100K×, and 1M×, indicating statistically significant χ^2 p values that show a strong concordance between estimates from the beta-binomial model, its NB approximation, and ultra-deep sequencing data. Distinguishing transitions and transversions further clarified the importance of classifying variants using sequencing depth as a proxy for the error rates. We obtain similar results from modeling the ultra-deep sequencing data from the *SF3B1* locus (Figure 5).

Detecting true mutations. We propose two comprehensive approaches to assess the presence of true mutations at very low abundance relative to background. Our methodology does not require matched normal samples or extensive filtering based on variant annotation resources.

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

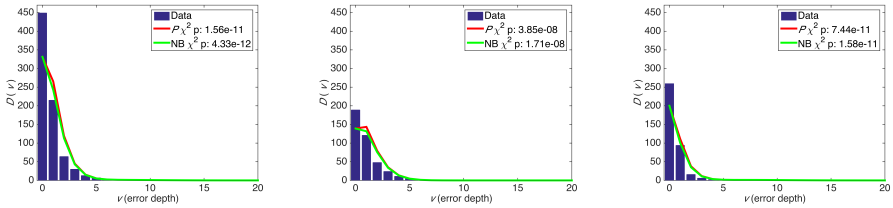

Figure 2: Error depth distribution in ultra-deep sequencing of a *TP53* locus at 10,000× for all variants (left), transitions (middle), and transversions (right).

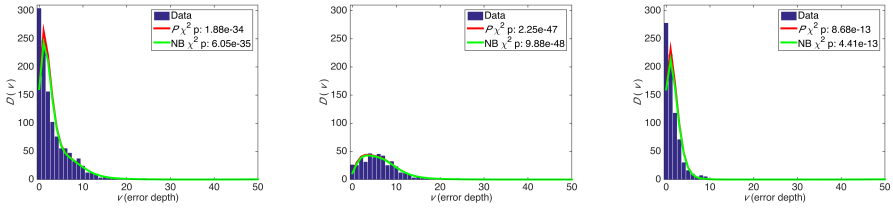

Figure 3: Error depth distribution in ultra-deep sequencing of a *TP53* locus at 100,000× for all variants (left), transitions (middle), and transversions (right).

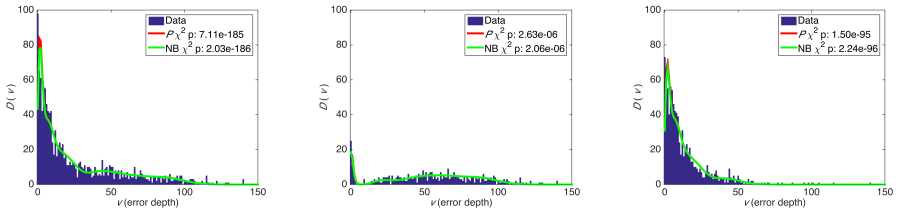

Figure 4: Error depth distribution in ultra-deep sequencing of a *TP53* locus at 1,000,000× for all variants (left), transitions (middle), and transversions (right).

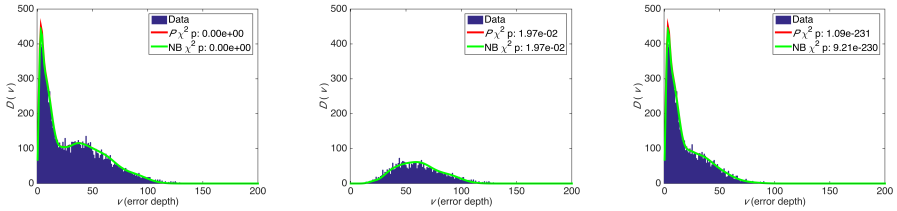

Figure 5: Error depth distribution in ultra-deep sequencing of a *SF3B1* locus at mean 620,000× for all variants (left), transitions (middle), and transversions (right).

First, having established an accurate model to describe the sequencing error distribution, a threshold is determined above which sequencing errors happen with a probability below an established statistical cut-off. These thresholds can be derived from all variants or a subset of variants, for example, only transitions or transversions. Figure 6 shows such thresholds for detecting the *TP53*-Y234C transition mutation in dilution experiments, where we are able to identify the mutation in

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

abundances as low as $5{:}10^4$ at 10K× and 100K×, and $1{:}10^4$ at 1M×, without any false positive calls. As shown in Figures 2, 3, and 4, there is better sensitivity for detecting a transversion substitution.

In the absence of matched normal samples, this approach is especially practical for identifying mutations that may exist in more than one tumor sample. Its application to 309 newly diagnosed CLL patients identified small sub-clonal prognostic mutations in four frequently mutated drivers of this neoplasm, present in 2 out of 1,000 wild-type alleles. These mutations were missed by traditional Sanger sequencing, but were validated by independent deep sequencing and allele-specific PCR [33,34].

Second, we tested an individual mutation in each sample against all other sequenced samples and calculated the cumulative P using equation (1). After correcting for multiple hypotheses using the Benjamini and Hochberg method [1], we generated a list of variants that satisfied a pre-determined false discovery rate. This approach is particularly powerful in identifying patient-specific mutations. We assess the presence of the *SF3B1*-K700E mutation in patient samples, and find the probability of observing the mutation in $1{:}10^3$ CLL dilution to be extremely significant compared to controls (Table 1). This approach can accurately identify sample-specific mutations by comparing multiple samples at the same exact mutated base.

In comparison of our method to other published variant calling algorithms, one comparable unbiased method is EBCall, whose implementation is based on beta-binomial distributions and establishing priors from normal sequencing data [36]. EBCall requires normal samples; therefore, we removed the reads harboring the diluted mutations in the EBCall analysis to simulate matched normal data. EBCall, with a sensitivity-adjusted configuration, successfully identified the *SF3B1*-K700E mutation in $1{:}10^3$ CLL dilution sample, as well as the *TP53*-Y234C mutation in the least diluted samples at all sequencing depths (i.e. 1:10 in 10K×, $1{:}10^2$ in 100K×, and $1{:}10^3$ in 1M×). However, it failed to detect the mutation at higher dilution levels, and also resulted in four false positive calls at 1M×.

5 Conclusion

Therapeutic resistance, one of the main causes of eventual disease relapse and mortality in cancer patients, is often associated with natural selection of preexisting resistant clones under treatment [12,34]. The detection of such low frequency sub-clones is hindered by a lack of precision-tested diagnostic assays.

Allele-specific, real-time PCR assays have been proposed to identify prognostic variants [15,25,42]. These approaches only target known mutations, and their adaptation to situations with large numbers of variants requires extensive primer calibration. In contrast, high-throughput sequencing provides an unbiased view of tumor heterogeneity and its genomic profile. Various techniques based on unique molecular identifiers have been proposed to correct both polymerase and sequencing errors [14,16,18,30] that facilitate distinguishing real mutations from mistakes that arise during amplification. However, the main hurdle in clinical utilization of these approached is the requirement for generating very large numbers of sequencing reads to assemble the genome of a single DNA molecule with high confidence at depth $> 2{,}000\times$.

Here, we addressed this important problem in cancer therapy by introducing a highly sensitive method to model sequencing noise, which allows the detection of prognostic markers of disease recurrence using ultra-deep targeted sequencing. Our approach is based on interrogating data from

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

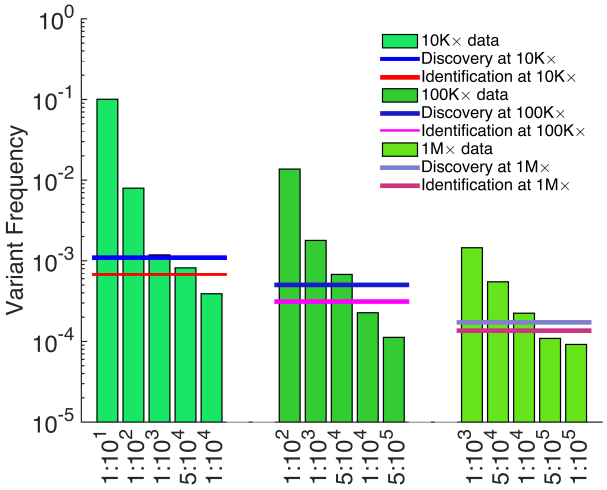

Figure 6: Sensitivity of detecting *TP53*-Y234C mutation dilutions. Assessing the presence of a variant requires correcting for multiple hypotheses based on the number of sequenced genomic positions (Bonferroni correction). Testing the presence of a discovered variant does not require such a correction; here, the p value of significance is set at 0.01.

multiple tumor samples at identical genomic regions and provides an accurate assessment of the error rate at a given position without relying on normal samples. Instead of establishing a fixed detection threshold for all variants, we directly calculate mutation-specific sensitivities. Overall, since ultra-deep sequencing methods are now routinely implemented in the clinic, we believe that the application of our comprehensive model to tumor samples will increase the speed with which patients can be evaluated during disease surveillance. Our method opens up the possibility of exploring the dynamics of cancer clones after treatment, timing the rise of resistance to therapy, and determining the clinical importance of minimal residual disease assessed from liquid biopsy samples for precise disease management [13,38].

Acknowledgments

The authors gratefully acknowledge the constructive feedback of Mohammad Hadigol and Alexandra Jacunski. R.R. acknowledges funding from the NIH (U54CA193313, R01CA185486, and R01CA179044). H.K. acknowledges support from the ACS (IRG-15-168-01), Rutgers Cancer Institute (P30CA072720), and Rutgers Office of Advanced Research Computing (NIH 1S10OD012346-01A1).

References

[1] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological **57**(1), 289–300 (1995)

[2] Burger, J.A., Tedeschi, A., Barr, P.M., Robak, T., Owen, C., Ghia, P., Bairey, O., Hillmen, P., Bartlett, N.L., Li, J., Simpson, D., Grosicki, S., Devereux, S., McCarthy, H., Coutre, S.,

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Presence of the *SF3B1*-K700E mutation in undiluted and diluted patient samples are tested against 18 samples that harbor wild-type allele.

Sample	Variant Depth (v)	Total Depth	Variant Frequency	Cumulative P	FDR	Cumulative NB
Control	64	711703	0.00009	9.48E-01	9.48E-01	8.66E-01
Control	62	642586	0.00010	8.45E-01	9.48E-01	6.85E-01
Control	74	717154	0.00010	7.02E-01	9.37E-01	5.09E-01
Control	94	630510	0.00015	2.95E-03	9.43E-03	5.00E-04
Control	56	505857	0.00011	4.68E-01	7.89E-01	2.60E-01
Control	61	509147	0.00012	2.49E-01	5.69E-01	1.07E-01
Control	88	699082	0.00013	1.12E-01	2.98E-01	4.12E-02
Control	75	749932	0.00010	7.91E-01	9.48E-01	6.22E-01
Control	62	657036	0.00009	8.84E-01	9.48E-01	7.47E-01
Control	56	581178	0.00010	8.34E-01	9.48E-01	6.63E-01
Control	81	731934	0.00011	4.75E-01	7.89E-01	2.85E-01
Control	70	636485	0.00011	4.93E-01	7.89E-01	2.89E-01
Control	40	452271	0.00009	9.15E-01	9.48E-01	7.92E-01
Control	59	511932	0.00012	3.51E-01	7.03E-01	1.72E-01
Control	46	518211	0.00009	9.27E-01	9.48E-01	8.15E-01
Control	80	714670	0.00011	4.35E-01	7.89E-01	2.50E-01
Control	85	736865	0.00012	3.33E-01	7.03E-01	1.75E-01
Control	74	691495	0.00011	5.87E-01	8.53E-01	3.82E-01
CLL	281058	630750	0.44559	0.00E+00	0.00E+00	0.00E+00
CLL 1:1000 dilution	2678	440301	0.00608	0.00E+00	0.00E+00	0.00E+00

Quach, H., Gaidano, G., Maslyak, Z., Stevens, D.A., Janssens, A., Offner, F., Mayer, J., ODwyer, M., Hellmann, A., Schuh, A., Siddiqi, T., Polliack, A., Tam, C.S., Suri, D., Cheng, M., Clow, F., Styles, L., James, D.F., Kipps, T.J.: Ibrutinib as initial therapy for patients with chronic lymphocytic leukemia. New England Journal of Medicine **373**(25), 2425–2437 (2015)

[3] Chen, L., Liu, P., Evans, T.C., Ettwiller, L.M.: Dna damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science **355**(6326), 752–756 (2017)

[4] Chen-Harris, H., Borucki, M.K., Torres, C., Slezak, T.R., Allen, J.E.: Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. BMC Genomics **14**(1), 96 (2013)

[5] Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., Sander, C.: Emerging landscape of oncogenic signatures across human cancers. Nat Genet **45**(10), 1127–33 (2013)

[6] Costello, M., Pugh, T.J., Fennell, T.J., Stewart, C., Lichtenstein, L., Meldrim, J.C., Fostel, J.L., Friedrich, D.C., Perrin, D., Dionne, D., Kim, S., Gabriel, S.B., Lander, E.S., Fisher, S., Getz, G.: Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative dna damage during sample preparation. Nucleic Acids Res **41**(6), e67 (2013)

[7] Furman, E.: On the convolution of the negative binomial random variables. Statistics and Probability Letters **77**(2), 169 – 172 (2007)

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[8] Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., Beerenwinkel, N.: Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nature Communications **3**, 811 EP – (2012)

[9] Gerstung, M., Papaemmanuil, E., Campbell, P.J.: Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics **30**(9), 1198–1204 (2014)

[10] Grossmann, V., Roller, A., Klein, H.U., Weissmann, S., Kern, W., Haferlach, C., Dugas, M., Haferlach, T., Schnittger, S., Kohlmann, A.: Robustness of amplicon deep sequencing underlines its utility in clinical applications. J Mol Diagn **15**(4), 473–84 (2013)

[11] Hallek, M.: Chronic lymphocytic leukemia: 2015 update on diagnosis, risk stratification, and treatment. American Journal of Hematology **90**(5), 446–460 (2015)

[12] Hata, A.N., Niederst, M.J., Archibald, H.L., Gomez-Caraballo, M., Siddiqui, F.M., Mulvey, H.E., Maruvka, Y.E., Ji, F., Bhang, H.E., Krishnamurthy Radhakrishna, V., Siravegna, G., Hu, H., Raoof, S., Lockerman, E., Kalsy, A., Lee, D., Keating, C.L., Ruddy, D.A., Damon, L.J., Crystal, A.S., Costa, C., Piotrowska, Z., Bardelli, A., Iafrate, A.J., Sadreyev, R.I., Stegmeier, F., Getz, G., Sequist, L.V., Faber, A.C., Engelman, J.A.: Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. Nat Med **22**(3), 262–9 (2016)

[13] Ivey, A., Hills, R.K., Simpson, M.A., Jovanovic, J.V., Gilkes, A., Grech, A., Patel, Y., Bhudia, N., Farah, H., Mason, J., Wall, K., Akiki, S., Griffiths, M., Solomon, E., McCaughan, F., Linch, D.C., Gale, R.E., Vyas, P., Freeman, S.D., Russell, N., Burnett, A.K., Grimwade, D., Group, U.K.N.C.R.I.A.W.: Assessment of minimal residual disease in standard-risk aml. N Engl J Med **374**(5), 422–33 (2016)

[14] Jee, J., Rasouly, A., Shamovsky, I., Akivis, Y., R. Steinman, S., Mishra, B., Nudler, E.: Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. Nature **534**(7609), 693–696 (2016)

[15] Jia, Y., Sanchez, J.A., Wangh, L.J.: Kinetic hairpin oligonucleotide blockers for selective amplification of rare mutations. Sci Rep **4**, 5921 (2014)

[16] Kennedy, S.R., Schmitt, M.W., Fox, E.J., Kohrn, B.F., Salk, J.J., Ahn, E.H., Prindle, M.J., Kuong, K.J., Shen, J.C., Risques, R.A., Loeb, L.A.: Detecting ultralow-frequency mutations by duplex sequencing. Nat. Protocols **9**(11), 2586–2606 (2014)

[17] Kessler, D.A., Levine, H.: Large population solution of the stochastic luria-delbruck evolution model. Proc Natl Acad Sci U S A **110**(29), 11,682–7 (2013)

[18] Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B.: Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences **108**(23), 9530–9535 (2011)

[19] Kipps, T.J., Stevenson, F.K., Wu, C.J., Croce, C.M., Packham, G., Wierda, W.G., O'Brien, S., Gribben, J., Rai, K.: Chronic lymphocytic leukaemia. Nature Reviews Disease Primers **3**, 16,096 EP – (2017)

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[20] Lazarian, G., Guize, R., Wu, C.J.: Clinical implications of novel genomic discoveries in chronic lymphocytic leukemia. Journal of Clinical Oncology **35**(9), 984–993 (2017)

[21] Lee, J.C., Sabavala, D.J.: Bayesian estimation and prediction for the beta-binomial model. Journal of Business and Economic Statistics **5**(3), 357–367 (1987)

[22] Li, H., Durbin, R.: Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics **25**(14), 1754–60 (2009)

[23] Li, M., Stoneking, M.: A new approach for detecting low-level mutations in next-generation sequence data. Genome Biology **13**(5), R34–R34 (2012)

[24] Luria, S.E., Delbrück, M.: Mutations of bacteria from virus sensitivity to virus resistance. Genetics **28**(6), 491–511 (1943)

[25] Milbury, C.A., Li, J., Makrigiorgos, G.M.: Pcr-based methods for the enrichment of minority alleles and mutations. Clin Chem **55**(4), 632–640 (2009)

[26] Morin, R.D., Mungall, K., Pleasance, E., Mungall, A.J., Goya, R., Huff, R.D., Scott, D.W., Ding, J., Roth, A., Chiu, R., Corbett, R.D., Chan, F.C., Mendez-Lago, M., Trinh, D.L., Bolger-Munro, M., Taylor, G., Hadj Khodabakhshi, A., Ben-Neriah, S., Pon, J., Meissner, B., Woolcock, B., Farnoud, N., Rogic, S., Lim, E.L., Johnson, N.A., Shah, S., Jones, S., Steidl, C., Holt, R., Birol, I., Moore, R., Connors, J.M., Gascoyne, R.D., Marra, M.A.: Mutational and structural analysis of diffuse large b-cell lymphoma using whole-genome sequencing. Blood **122**(7), 1256–1265 (2013)

[27] Nadeu, F., Delgado, J., Royo, C., Baumann, T., Stankovic, T., Pinyol, M., Jares, P., Navarro, A., Martín-García, D., Beà, S., Salaverria, I., Oldreive, C., Aymerich, M., Suárez-Cisneros, H., Rozman, M., Villamor, N., Colomer, D., López-Guillermo, A., González, M., Alcoceba, M., Terol, M.J., Colado, E., Puente, X.S., López-Otín, C., Enjuanes, A., Campo, E.: Clinical impact of clonal and subclonal tp53, sf3b1, birc3, notch1, and atm mutations in chronic lymphocytic leukemia. Blood **127**(17), 2122–2130 (2016)

[28] Naxerova, K., Reiter, J.G., Brachtel, E., Lennerz, J.K., van de Wetering, M., Rowan, A., Cai, T., Clevers, H., Swanton, C., Nowak, M.A., Elledge, S.J., Jain, R.K.: Origins of lymphatic and distant metastases in human colorectal cancer. Science **357**(6346), 55–60 (2017)

[29] Ndifon, W., Gal, H., Shifrut, E., Aharoni, R., Yissachar, N., Waysbort, N., Reich-Zeliger, S., Arnon, R., Friedman, N.: Chromatin conformation governs t-cell receptor jbeta gene segment usage. Proc Natl Acad Sci U S A **109**(39), 15,865–70 (2012)

[30] Newman, A.M., Lovejoy, A.F., Klass, D.M., Kurtz, D.M., Chabon, J.J., Scherer, F., Stehr, H., Liu, C.L., Bratman, S.V., Say, C., Zhou, L., Carter, J.N., West, R.B., Sledge Jr, G.W., Shrager, J.B., Loo Jr, B.W., Neal, J.W., Wakelee, H.A., Diehn, M., Alizadeh, A.A.: Integrated digital error suppression for improved detection of circulating tumor dna. Nat Biotech **34**(5), 547–555 (2016)

[31] Oshima, K., Khiabanian, H., da Silva-Almeida, A.C., Tzoneva, G., Abate, F., Ambesi-Impiombato, A., Sanchez-Martin, M., Carpenter, Z., Penson, A., Perez-Garcia, A., Eckert,

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

C., Nicolas, C., Balbin, M., Sulis, M.L., Kato, M., Koh, K., Paganin, M., Basso, G., Gastier-Foster, J.M., Devidas, M., Loh, M.L., Kirschner-Schwabe, R., Palomero, T., Rabadan, R., Ferrando, A.A.: Mutational landscape, clonal evolution patterns, and role of ras mutations in relapsed acute lymphoblastic leukemia. Proc Natl Acad Sci U S A (2016)

[32] Pospisilova, S., Gonzalez, D., Malcikova, J., Trbusek, M., Rossi, D., Kater, A.P., Cymbalista, F., Eichhorst, B., Hallek, M., Dohner, H., Hillmen, P., van Oers, M., Gribben, J., Ghia, P., Montserrat, E., Stilgenbauer, S., Zenz, T.: Eric recommendations on tp53 mutation analysis in chronic lymphocytic leukemia. Leukemia **26**(7), 1458–1461 (2012)

[33] Rasi, S., Khiabanian, H., Ciardullo, C., Terzi-di Bergamo, L., Monti, S., Spina, V., Bruscaggin, A., Cerri, M., Deambrogi, C., Martuscelli, L., Biasi, A., Spaccarotella, E., De Paoli, L., Gattei, V., Foa, R., Rabadan, R., Gaidano, G., Rossi, D.: Clinical impact of small subclones harboring notch1, sf3b1 or birc3 mutations in chronic lymphocytic leukemia. Haematologica **101**(4), e135–8 (2016)

[34] Rossi, D., Khiabanian, H., Spina, V., Ciardullo, C., Bruscaggin, A., Fama, R., Rasi, S., Monti, S., Deambrogi, C., De Paoli, L., Wang, J., Gattei, V., Guarini, A., Foa, R., Rabadan, R., Gaidano, G.: Clinical impact of small tp53 mutated subclones in chronic lymphocytic leukemia. Blood **123**(14), 2139–47 (2014)

[35] Rossi, D., Rasi, S., Spina, V., Bruscaggin, A., Monti, S., Ciardullo, C., Deambrogi, C., Khiabanian, H., Serra, R., Bertoni, F., Forconi, F., Laurenti, L., Marasca, R., Dal-Bo, M., Rossi, F.M., Bulian, P., Nomdedeu, J., Del Poeta, G., Gattei, V., Pasqualucci, L., Rabadan, R., Foà, R., Dalla-Favera, R., Gaidano, G.: Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. Blood **121**(8), 1403–1412 (2013)

[36] Shiraishi, Y., Sato, Y., Chiba, K., Okuno, Y., Nagata, Y., Yoshida, K., Shiba, N., Hayashi, Y., Kume, H., Homma, Y., Sanada, M., Ogawa, S., Miyano, S.: An empirical bayesian framework for somatic mutation detection from cancer genome sequencing data. Nucleic Acids Res **41**(7), e89 (2013)

[37] Shlush, L.I., Mitchell, A., Heisler, L., Abelson, S., Ng, S.W.K., Trotman-Grant, A., Medeiros, J.J.F., Rao-Bhatia, A., Jaciw-Zurakowsky, I., Marke, R., McLeod, J.L., Doedens, M., Bader, G., Voisin, V., Xu, C., McPherson, J.D., Hudson, T.J., Wang, J.C.Y., Minden, M.D., Dick, J.E.: Tracing the origins of relapse in acute myeloid leukaemia to stem cells. Nature **547**(7661), 104–108 (2017)

[38] Siravegna, G., Marsoni, S., Siena, S., Bardelli, A.: Integrating liquid biopsies into the management of cancer. Nat Rev Clin Oncol **advance online publication**, – (2017)

[39] Souers, A.J., Leverson, J.D., Boghaert, E.R., Ackler, S.L., Catron, N.D., Chen, J., Dayton, B.D., Ding, H., Enschede, S.H., Fairbrother, W.J., Huang, D.C.S., Hymowitz, S.G., Jin, S., Khaw, S.L., Kovar, P.J., Lam, L.T., Lee, J., Maecker, H.L., Marsh, K.C., Mason, K.D., Mitten, M.J., Nimmer, P.M., Oleksijew, A., Park, C.H., Park, C.M., Phillips, D.C., Roberts, A.W., Sampath, D., Seymour, J.F., Smith, M.L., Sullivan, G.M., Tahir, S.K., Tse, C., Wendt, M.D., Xiao, Y., Xue, J.C., Zhang, H., Humerickhouse, R.A., Rosenberg, S.H., Elmore, S.W.: Abt-199, a potent and selective bcl-2 inhibitor, achieves antitumor activity while sparing platelets. Nat Med **19**(2), 202–208 (2013)

bioRxiv preprint doi: https://doi.org/10.1101/128587; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[40] Stewart, T.J., Abrams, S.I.: How tumours escape mass destruction. Oncogene **27**(45), 5894–5903 (2008)

[41] Trifonov, V., Pasqualucci, L., Tiacci, E., Falini, B., Rabadan, R.: Savi: a statistical algorithm for variant frequency identification. BMC Syst Biol **7 Suppl 2**, S2 (2013)

[42] Vargas, D.Y., Kramer, F.R., Tyagi, S., Marras, S.A.E.: Multiplex real-time pcr assays that measure the abundance of extremely rare mutations associated with cancer. PLoS One **11**(5), e0156,546 (2016)

[43] Vellaisamy, P., Upadhye, N.S.: On the sums of compound negative binomial and gamma random vaariables. Journal of Applied Probability **46**(1), 272–283 (2009)

[44] Wang, J., Cazzato, E., Ladewig, E., Frattini, V., Rosenbloom, D.I.S., Zairis, S., Abate, F., Liu, Z., Elliott, O., Shin, Y.J., Lee, J.K., Lee, I.H., Park, W.Y., Eoli, M., Blumberg, A.J., Lasorella, A., Nam, D.H., Finocchiaro, G., Iavarone, A., Rabadan, R.: Clonal evolution of glioblastoma under therapy. Nat Genet **48**(7), 768–776 (2016)

[45] Wang, J., Khiabanian, H., Rossi, D., Fabbri, G., Gattei, V., Forconi, F., Laurenti, L., Marasca, R., Del Poeta, G., Fo, R., Pasqualucci, L., Gaidano, G., Rabadan, R.: Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia. eLife **3**, e02,869 (2014)

[46] Zhang, J., Kobert, K., Flouri, T., Stamatakis, A.: Pear: a fast and accurate illumina paired-end read merger. Bioinformatics **30**(5), 614–620 (2014)