Highly sensitive detection of small variants in multi-sample ultra-deep tumor sequencing

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

[1]Departments of Systems Biology and Biomedical Informatics, Columbia University, New York, NY
[2]Center for Topology of Cancer Evolution and Heterogeneity, Columbia University, New York, NY
[3]The Feinstein Institute for Medical Research, Northwell Health, Manhasset, NY
[4]Institute for Cancer Genetics, Columbia University, New York, NY
[5]Rutgers Cancer Institute, Rutgers University, New Brunswick, NJ
[*]Corresponding author: h.khiabanian@rutgers.edu

Abstract

One of the main causes of cancer mortality is tumor evolution to therapy-resistant disease. Drug resistance may emerge from the rise of ancestral clones that gain fitness through therapy-induced natural selection. Previously, it was shown that the presence of drug-resistant sub-clones at diagnosis or prior to therapy could be a strong predictor of poor survival, disease transformation, and refractoriness, with direct implications for disease management. Although such prognostic mutations are most commonly identified using amplicon-based or hybrid-capture deep sequencing in a clinical setting, their sensitive detection relies on the accurate analysis of background noise, specifically sequencing errors that arise from prior polymerase chain reaction cycles. In this work, we provide a comprehensive, unbiased model that precisely describes this background noise and show that it can be approximated by aggregating negative binomial (NB) distributions, using tumor-only data. We evaluate our model and its NB approximation with simulated exponentially expanded populations, as well as ultra-deep sequencing data from cell line and patient sample dilution experiments. Our method goes beyond estimating fixed detection thresholds for all variants, having the power to assess mutation-specific sensitivities that allow identification of 1-2 mutated alleles out of 10,000 wild-type. This facilitates the design of precise treatment strategies and contributes significantly to combatting drug resistance and increasing positive outcomes.

Introduction

In the development and evolution of cancer, while genetic alterations accumulate, fitter sub-populations gain dominance and give rise to clinically diagnosable disease. Drug resistance often emerges from such tumor evolutionary patterns via Darwinian selection of sub-clones with greater fitness under therapy [1]. The detection of these low frequency disease-driving clones in the preclinical phase, at diagnosis, or during treatment, can be a strong predictor of poor survival, disease transformation and refractoriness, with direct implications for treatment strategy.

Rapid progress in genomic sequencing has enabled the molecular characterization of human neoplasms and has improved our understanding of cancer development and evolution. The underlying

biochemical mechanisms are often recurrent across different cancers: for example, aberrations leading to unregulated cell growth or inactivation of apoptosis are common to almost all neoplasms. In particular, tumors arising from different cells of origin often harbor identical low frequency genetic alterations, which often have similar prognostic consequences [2,3]. Cells bearing some of these mutations may be resistant to therapy and may exist at very low abundance (<10 in 10,000 wild-type alleles) in the preclinical phase or may persist and be positively selected during therapy-induced remission. To date, timing the rise of such resistant clones and translating this into clinical practice has been confounded by the lack of calibrated methods to accurately detect low frequency variants.

Allele-specific, real-time polymerase chain reaction (PCR) assays have been proposed to identify prognostic variants [4,5,6]. However, 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. The main hurdle in clinical utilization of ultra-deep sequencing data (depth $> 2{,}000\times$) is distinguishing real mutations from mistakes that arise during amplification. Various techniques based on unique molecular identifiers have been proposed to correct both polymerase and sequencing errors [7,8,9,10]; however, these methodologies require the generation of very large numbers of sequencing reads to assemble the genome of a single DNA molecule with high confidence. Therefore, basic amplicon-based or hybrid-capture targeted sequencing have remained the most commonly used methods to track prognostic markers in both clinical and basic research applications [11]. Precisely designing primers to generate overlapping read pairs allows their merging and facilitates correcting errors that accumulate in the sequencer after the amplification of targeted loci, while leaving the PCR errors uncorrected [12,13]. The challenge is then to determine sensitivity thresholds, i.e. the depths above which sequencing 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 deviations from this null hypothesis indicate the presence of true variants. Since different errors occur at different rates [14,15], a single threshold cannot comprehensively test the significance of all variants. Bayesian modeling of background error, in conjunction with multiple filtering criteria using sequencing data from patient-matched and normal samples has been proposed to address this issue [16,17,18,19].

Recently, we reported an algorithm called Backtrack that models the background PCR noise and establishes depth thresholds to distinguish true variants from errors in ultra-deep tumor-only sequencing [20,21]. We considered different types of error distributions: (i) the negative binomial distribution, which is known to describe the depth distribution of clones after PCR amplification through a Poisson-Gamma mixture model [22]; (ii) a single or a linear combination of Lauria-Delbrück distributions, characterizing the expected number of spontaneous mutations during growth when the PCR error rate is assumed to be constant [23]. Our empirical analysis indicated that the negative binomial distribution gives the best fit to the error depth distribution based on goodness-of-fit log-likelihood. The application of Backtrack to 309 newly diagnosed chronic lymphocytic leukemia (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 (using different primers) and allele-specific PCR [20,21]. Despite Backtrack's accuracy, we did not provide a proof for our empirical approach.

In this manuscript, we revisit this problem and introduce 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 highly sensitive, mutation-specific approaches to detect true mutations without the need for control data from unmutated normal tissue DNA.

Methods

Derivation of the error depth distribution. Let us assume an experiment in which S independent samples are subjected to ultra-deep sequencing. At each genomic locus, three possible erroneous substitutions or two types of small indels — henceforth called variants — may occur. The probability of observing n_i reads harboring a variant amongst N_i total reads that cover its position follows a binomial distribution, $\mathrm{Bino}(n_i|N_i)$. Therefore, if $M = \sum_{i\neq j}^{S} N_i$ and $m = \sum_{i\neq j}^{S} n_i$, the posterior predictive p value for having detected a true mutation in sample j, given other $S-1$ samples is

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

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

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

Variations of equation (1) have been previously derived for sequencing depths $> 100\times$ [17,18,19]. In ultra-deep data, where $N_i > 5{,}000\times$, we can assume that $n_i \ll N_i$, and hence, using Stirling's approximation, $\binom{N_i}{n_i} \approx \frac{N_i^{n_i}}{n_i!}$. Equation (1) can then be approximated by

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

which equals $\mathrm{NB}(n_j|1+m, \frac{N_j}{N_j+M})$, where NB indicates the negative binomial distribution.

Exponential expansions at varying error rates. In an exponentially expanded population that is generated through c amplification cycles, if errors accumulate at a rate of μ substitutions per site per cycle, the average error depth (i.e. the average number of reads harboring errors) is equal to $2^c\mu$. For S such populations, the error depth distribution is described by equation (1), or is approximated by a negative binomial distribution, $\mathrm{NB}(1+(S-1)2^c\mu, 1-\frac{1}{S})$, as derived above in equation (2). Since different types of PCR mis-incorporations (e.g. transitions versus transversions) occur at differential rates, assuming R independent rates, the observed number of variants $D(v)$, with error depth v is given by,

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

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

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

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 [24,25]. Thus, the approximation of $D(v)$ in equations (3) and (4) with sums of negative binomial distributions that have success probability of $1-\frac{1}{S}$, suggests the empirical observation implemented in the Backtrack algorithm [20].

Cell line and patient sample dilution experiments. In the first experiment, a series of dilutions using the SU-DHL-6 cell line (Diffuse Large B-Cell Lymphoma), which carries a heterozygous *TP53*-Y234C missense transition substitution was generated. 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 leukemia cells from a CLL patient, harboring a heterozygous *SF3B1*-K700E missense transition substitution were analyzed and the *TP53* mutation 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 pair-end reads substantially overlapped 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 [26], and all variable sites were identified using an inclusive variant caller, adapted from the SAVI algorithm [27].

Results and Discussion

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.

Cell line dilution experiments. Next, 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).

Figure 2 shows 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 3).

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.

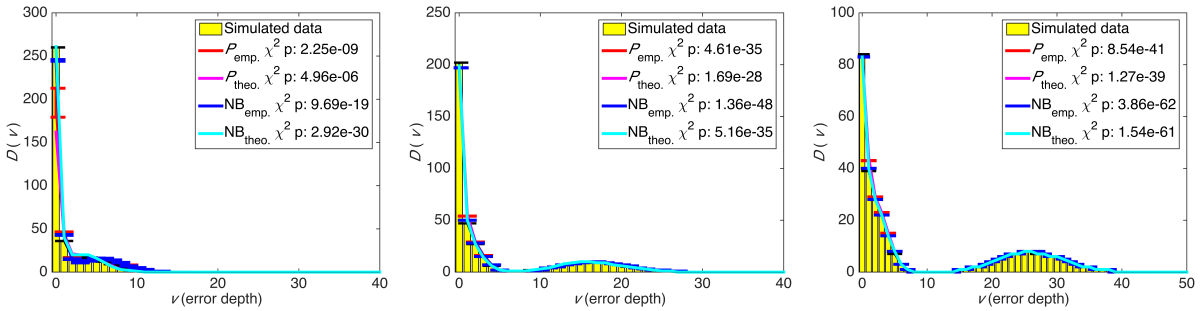
First, having established an accurate model to describe the sequencing error distribution, a threshold is determined above which sequencing errors happen with a probability below an established statistical cut-off . These thresholds can be derived from all variants or a subset of variants, for example only transitions or transversions. Figure 4 shows such thresholds for detecting the *TP53*-Y234C transition mutation in dilution experiments, where we are able to identify the mutation in 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 Figure 2, there is better sensitivity for detecting a transversion substitution.

Second, we test an individual mutation in each sample against all other sequenced samples and calculate cumulative P using equation (1). After correcting for multiple hypotheses using the Benjamini and Hochberg method, we generate a list of variants that satisfies 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).

In the absence of matched normal samples, the first approach is especially practical for identifying mutations that may exist in more than one tumor sample. The second approach, however, can accurately identify sample-specific mutations by comparing multiple samples at the same exact mutated base. In comparison, amongst various published variant calling algorithms, the only comparable unbiased method is EBCall, whose implementation is based on beta-binomial distributions and establishing priors from normal sequencing data [18]. EBCall requires normal samples; therefore, we remove the reads harboring the diluted mutations to simulate matched normal data. EBCall, with a sensitivity-adjusted configuration, successfully identifies 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 fails to detect the mutation in higher dilution levels, and also results in four false positive calls at 1M×.

Conclusion

Therapeutic resistance, one of the main causes of eventual disease relapse and mortality in cancer patients, is often associated with the natural selection of pre-existing resistant clones under treatment [1,20]. The detection of such low frequency sub-clones is hindered by a lack of precision-tested diagnostic assays. In this manuscript, we address this important problem in cancer therapy by introducing a highly sensitive method to detect prognostic markers of disease recurrence using ultra-deep targeted sequencing (depth $>$ 2,000×), a commonly utilized technology in clinical prac-



(a) 12 cycles (b) 14 cycles (c) 18 cycles

Figure 1: Number of variants with error depth of v from aggregated cycles of PCR amplification at four error rates. $P_{\text{theo.}}$ and $\text{NB}_{\text{theo.}}$ are calculated using equation (3), and $P_{\text{emp.}}$ and $\text{NB}_{\text{emp.}}$ are calculates using equation (4).

tice. Our approach is based on interrogating data from 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. Therefore, 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.

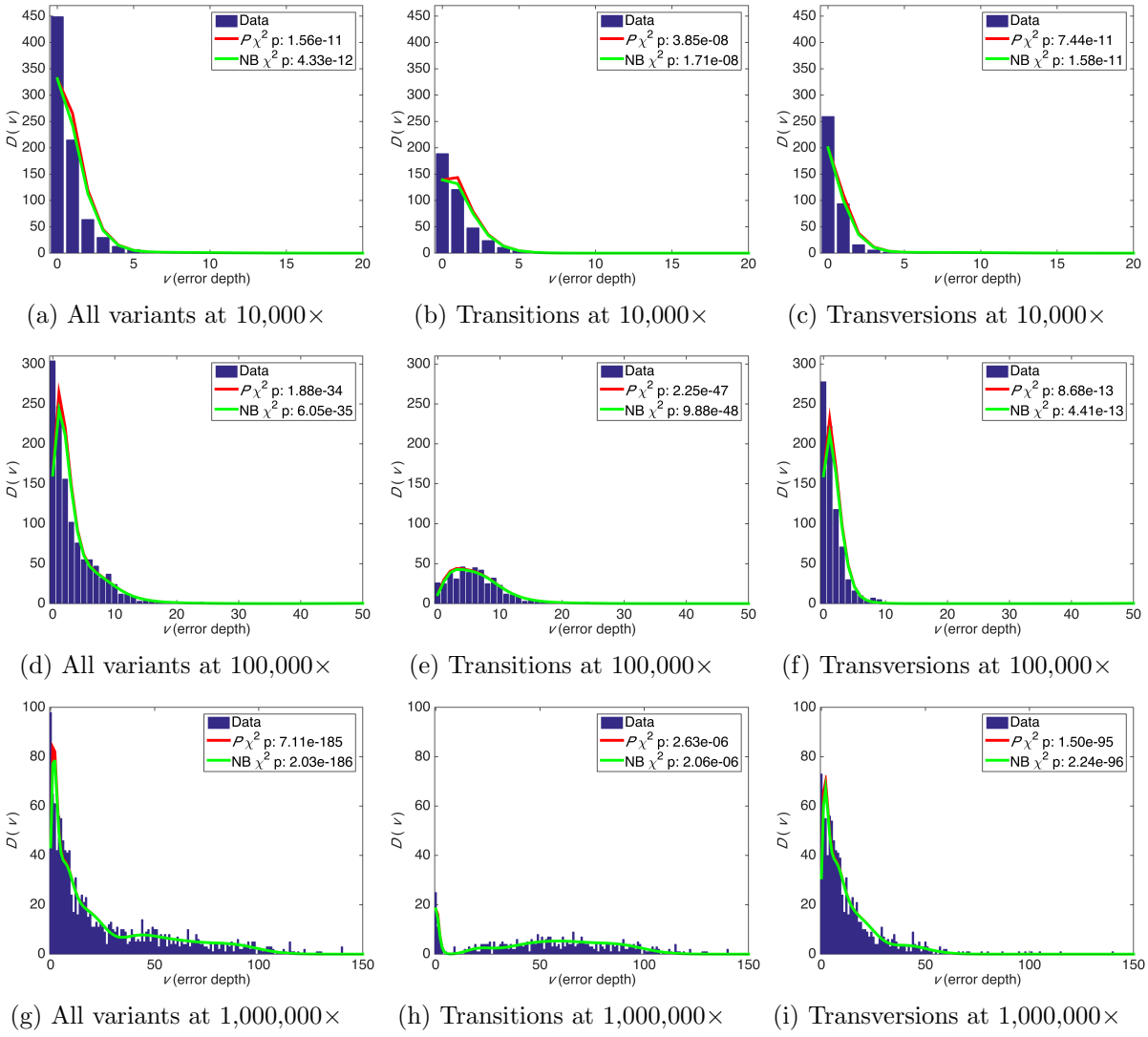
Acknowledgments

The authors gratefully acknowledge the constructive feedback of Gyan Bhanot, 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] A. N. Hata, M. J. Niederst, H. L. Archibald, M. Gomez-Caraballo, F. M. Siddiqui, H. E. Mulvey, Y. E. Maruvka, F. Ji, H. E. Bhang, V. Krishnamurthy Radhakrishna, G. Siravegna, H. Hu, S. Raoof, E. Lockerman, A. Kalsy, D. Lee, C. L. Keating, D. A. Ruddy, L. J. Damon, A. S. Crystal, C. Costa, Z. Piotrowska, A. Bardelli, A. J. Iafrate, R. I. Sadreyev, F. Stegmeier, G. Getz, L. V. Sequist, A. C. Faber, and J. A. Engelman, "Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition," *Nat Med*, vol. 22, no. 3, pp. 262–9, 2016.

[2] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander, "Emerging landscape of oncogenic signatures across human cancers," *Nat Genet*, vol. 45, no. 10, pp. 1127–33, 2013.

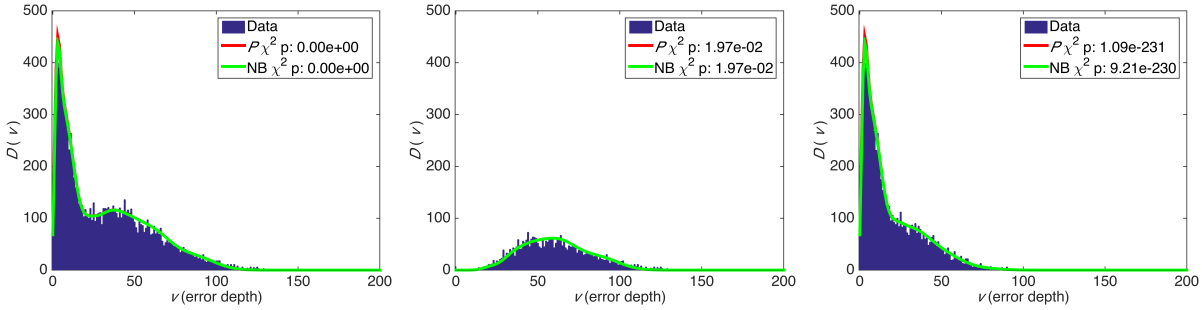


(a) All variants at 10,000× (b) Transitions at 10,000× (c) Transversions at 10,000×

(d) All variants at 100,000× (e) Transitions at 100,000× (f) Transversions at 100,000×

(g) All variants at 1,000,000× (h) Transitions at 1,000,000× (i) Transversions at 1,000,000×

Figure 2: Error depth distribution in ultra-deep sequencing of a *TP53* locus.



(a) All variants (b) Transitions (c) Transversions

Figure 3: Error depth distribution in ultra-deep sequencing of a *SF3B1* locus at mean 620,000×.

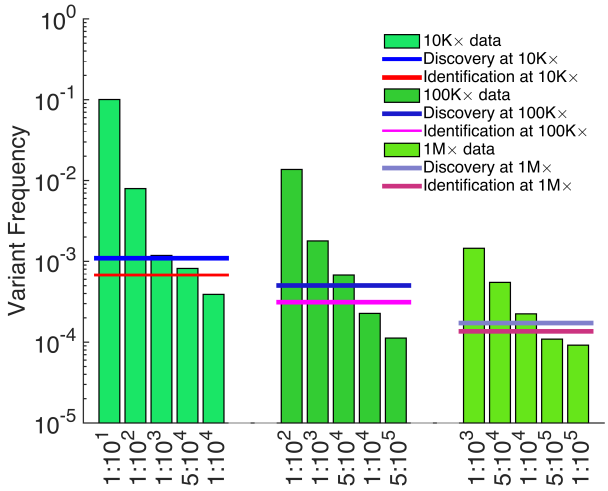


Figure 4: Sensitivity of detecting *TP53*-Y234C mutation dilutions. Assessing the presence of a variant requires correcting for multiple hypotheses based on the number of sequenced genomic positions (Bonferroni correction). Testing the presence of a discovered variant does not require such a correction; here, the p value of significance is set at 0.01.

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

[3] M. D. Leiserson, F. Vandin, H. T. Wu, J. R. Dobson, J. V. Eldridge, J. L. Thomas, A. Papoutsaki, Y. Kim, B. Niu, M. McLellan, M. S. Lawrence, A. Gonzalez-Perez, D. Tamborero, Y. Cheng, G. A. Ryslik, N. Lopez-Bigas, G. Getz, L. Ding, and B. J. Raphael, "Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes," *Nat Genet*, vol. 47, no. 2, pp. 106–14, 2015.

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

[5] Y. Jia, J. A. Sanchez, and L. J. Wangh, "Kinetic hairpin oligonucleotide blockers for selective amplification of rare mutations.," *Sci Rep*, vol. 4, p. 5921, Aug 2014.

[6] D. Y. Vargas, F. R. Kramer, S. Tyagi, and S. A. E. Marras, "Multiplex real-time pcr assays that measure the abundance of extremely rare mutations associated with cancer.," *PLoS One*, vol. 11, no. 5, p. e0156546, 2016.

[7] I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein, "Detection and quantification of rare mutations with massively parallel sequencing," *Proceedings of the National Academy of Sciences*, vol. 108, no. 23, pp. 9530–9535, 2011.

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

[9] J. Jee, A. Rasouly, I. Shamovsky, Y. Akivis, S. R. Steinman, B. Mishra, and E. Nudler, "Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing," *Nature*, vol. 534, pp. 693–696, 06 2016.

[10] A. M. Newman, A. F. Lovejoy, D. M. Klass, D. M. Kurtz, J. J. Chabon, F. Scherer, H. Stehr, C. L. Liu, S. V. Bratman, C. Say, L. Zhou, J. N. Carter, R. B. West, G. W. Sledge Jr, J. B. Shrager, B. W. Loo Jr, J. W. Neal, H. A. Wakelee, M. Diehn, and A. A. Alizadeh, "Integrated digital error suppression for improved detection of circulating tumor dna," *Nat Biotech*, vol. 34, pp. 547–555, 05 2016.

[11] V. Grossmann, A. Roller, H. U. Klein, S. Weissmann, W. Kern, C. Haferlach, M. Dugas, T. Haferlach, S. Schnittger, and A. Kohlmann, "Robustness of amplicon deep sequencing underlines its utility in clinical applications," *J Mol Diagn*, vol. 15, no. 4, pp. 473–84, 2013.

[12] H. Chen-Harris, M. K. Borucki, C. Torres, T. R. Slezak, and J. E. Allen, "Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs," *BMC Genomics*, vol. 14, no. 1, p. 96, 2013.

[13] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, "Pear: a fast and accurate illumina paired-end read merger," *Bioinformatics*, vol. 30, pp. 614–620, 03 2014.

[14] M. Costello, T. J. Pugh, T. J. Fennell, C. Stewart, L. Lichtenstein, J. C. Meldrim, J. L. Fostel, D. C. Friedrich, D. Perrin, D. Dionne, S. Kim, S. B. Gabriel, E. S. Lander, S. Fisher, and G. Getz, "Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative dna damage during sample preparation," *Nucleic Acids Res*, vol. 41, no. 6, p. e67, 2013.

[15] L. Chen, P. Liu, T. C. Evans, and L. M. Ettwiller, "Dna damage is a pervasive cause of sequencing errors, directly confounding variant identification," *Science*, vol. 355, no. 6326, pp. 752–756, 2017.

[16] M. Li and M. Stoneking, "A new approach for detecting low-level mutations in next-generation sequence data," *Genome Biology*, vol. 13, no. 5, pp. R34–R34, 2012.

[17] M. Gerstung, C. Beisel, M. Rechsteiner, P. Wild, P. Schraml, H. Moch, and N. Beerenwinkel, "Reliable detection of subclonal single-nucleotide variants in tumour cell populations," *Nature Communications*, vol. 3, pp. 811 EP –, 05 2012.

[18] Y. Shiraishi, Y. Sato, K. Chiba, Y. Okuno, Y. Nagata, K. Yoshida, N. Shiba, Y. Hayashi, H. Kume, Y. Homma, M. Sanada, S. Ogawa, and S. Miyano, "An empirical bayesian framework for somatic mutation detection from cancer genome sequencing data.," *Nucleic Acids Res*, vol. 41, p. e89, Apr 2013.

[19] M. Gerstung, E. Papaemmanuil, and P. J. Campbell, "Subclonal variant calling with multiple samples and prior knowledge," *Bioinformatics*, vol. 30, pp. 1198–1204, 05 2014.

[20] D. Rossi, H. Khiabanian, V. Spina, C. Ciardullo, A. Bruscaggin, R. Fama, S. Rasi, S. Monti, C. Deambrogi, L. De Paoli, J. Wang, V. Gattei, A. Guarini, R. Foa, R. Rabadan, and G. Gaidano, "Clinical impact of small tp53 mutated subclones in chronic lymphocytic leukemia," *Blood*, vol. 123, no. 14, pp. 2139–47, 2014.

[21] S. Rasi, H. Khiabanian, C. Ciardullo, L. Terzi-di Bergamo, S. Monti, V. Spina, A. Bruscaggin, M. Cerri, C. Deambrogi, L. Martuscelli, A. Biasi, E. Spaccarotella, L. De Paoli, V. Gattei, R. Foa, R. Rabadan, G. Gaidano, and D. Rossi, "Clinical impact of small subclones harboring notch1, sf3b1 or birc3 mutations in chronic lymphocytic leukemia," *Haematologica*, vol. 101, no. 4, pp. e135–8, 2016.

[22] W. Ndifon, H. Gal, E. Shifrut, R. Aharoni, N. Yissachar, N. Waysbort, S. Reich-Zeliger, R. Arnon, and N. Friedman, "Chromatin conformation governs t-cell receptor jbeta gene segment usage," *Proc Natl Acad Sci U S A*, vol. 109, no. 39, pp. 15865–70, 2012.

[23] D. A. Kessler and H. Levine, "Large population solution of the stochastic luria-delbruck evolution model," *Proc Natl Acad Sci U S A*, vol. 110, no. 29, pp. 11682–7, 2013.

[24] E. Furman, "On the convolution of the negative binomial random variables," *Statistics and Probability Letters*, vol. 77, no. 2, pp. 169 – 172, 2007.

[25] P. Vellaisamy and N. S. Upadhye, "On the sums of compound negative binomial and gamma random variables," *Journal of Applied Probability*, vol. 46, no. 1, pp. 272–283, 2009.

[26] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows-wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–60, 2009.

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