Open access • Posted Content • DOI:10.1101/184291

MERIT: a Mutation Error Rate Identification Toolkit for Ultra-deep Sequencing Applications — Source link

Mohammad Hadigol, Hossein Khiabanian

Institutions: Rutgers University

Published on: 04 Sep 2017 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Deep sequencing

Hadigol and Khiabanian *BMC Bioinformatics* (2018) 19:219
https://doi.org/10.1186/s12859-018-2223-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

MERIT reveals the impact of genomic context on sequencing error rate in ultra-deep applications

Mohammad Hadigol[1] and Hossein Khiabanian[1,2*]

Abstract

Background: Rapid progress in high-throughput sequencing (HTS) and the development of novel library preparation methods have improved the sensitivity of detecting mutations in heterogeneous samples, specifically in high-depth (> 500×) clinical applications. However, HTS methods are bounded by their technical and theoretical limitations and sequencing errors cannot be completely eliminated. Comprehensive quantification of the background noise can highlight both the efficiency and the limitations of any HTS methodology, and help differentiate true mutations at low abundance from artifacts.

Results: We introduce MERIT (Mutation Error Rate Inference Toolkit), designed for in-depth quantification of erroneous substitutions and small insertions and deletions. MERIT incorporates an all-inclusive variant caller and considers genomic context, including the nucleotides immediately at 5′and 3′, thereby establishing error rates for 96 possible substitutions as well as four single-base and 16 double-base indels. We applied MERIT to ultra-deep sequencing data (1,300,000×) obtained from the amplification of multiple clinically relevant loci, and showed a significant relationship between error rates and genomic contexts. In addition to observing significant difference between transversion and transition rates, we identified variations of more than 100-fold within each error type at high sequencing depths. For instance, T>G transversions in trinucleotide GTCs occurred 133.5± 65.9 more often than those in ATAs. Similarly, C>T transitions in GCGs were observed at 73.8± 10.5 higher rate than those in TCTs. We also devised an *in silico* approach to determine the optimal sequencing depth, where errors occur at rates similar to those of expected true mutations. Our analyses showed that increasing sequencing depth might improve sensitivity for detecting some mutations based on their genomic context. For example, T>G rate of error in GTCs did not change when sequenced beyond 10,000×; in contrast, T>G rate in TTAs consistently improved even at above 500,000×.

Conclusions: Our results demonstrate significant variation in nucleotide misincorporation rates, and suggest that genomic context should be considered for comprehensive profiling of specimen-specific and sequencing artifacts in high-depth assays. This data provide strong evidence against assigning a single allele frequency threshold to call mutations, for it can result in substantial false positive as well as false negative variants, with important clinical consequences.

Keywords: Deep sequencing, Sequencing noise, Genomic context, Polymerase fidelity, Optimal depth

*Correspondence: h.khiabanian@rutgers.edu
[1]Center for Systems and Computational Biology, Rutgers Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ, USA
[2]Department of Pathology and Laboratory Medicine, Rutgers Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, USA

Background

The rising utilization of high-throughput sequencing (HTS) in clinical oncology has transformed our understanding of cancer evolution and has provided clinicians with an invaluable tool for precise diagnosis and prognosis.

In clinical cancer genomic testing, target-capture library preparation assays are favored over whole genome or whole exome sequencing approaches because of their lower cost in obtaining higher sequencing depth – the number of reads covering a specific locus [1]. High-depth DNA sequencing enables confident detection of small clones of somatically mutated cells in heterogenous tumor samples, where in addition to genomically diverse cancer cells, contaminating normal cells may also be present. Using polymerase chain reaction (PCR)-based amplicon or hybridization-capture enrichment techniques, clinical-grade cancer sequencing panels are capable of producing 500 to $>$ 10,000 reads mapping to each targeted locus [2, 3]. Specifically, a minimum average depth of 500$\times$ is strongly advised by regulatory bodies for reliable detection of somatic mutations with variant allele frequencies (VAFs) as low as 5% in tumor specimens [4].

The power to detect small clones in heterogenous samples may improve by increasing depth; however, confident detection and differentiation of true mutations with low VAFs, e.g., $<$ 0.1%, from the sequencing artifacts remains a challenge. HTS errors are dominated by misreading a base within the instrument or nucleotide misincorporations during library enrichment with PCR. Differential rate of substitution errors in HTS has been observed and attributed to common DNA damaging events such as spontaneous deamination, presence of oxidized bases in cells in addition to *ex vivo* oxidation during DNA extraction [5], or short-lived high temperatures during acoustic shearing [6]. Such events often lead to higher rates of transitions versus transversions [7–11] or increased number of errors in specific genomic contexts. These differences can be more pronounced at higher sequencing depths and directly impact the sensitivity for detecting true mutations with low VAFs. Here, we hypothesize that the genomic context of substitution errors, i.e., the nucleotides immediately at their 5′ and 3′, is a determinant factor in estimating their rates at high sequencing depths. To this end, we generated ultra-deep sequencing data (1,300,000$\times$) and developed MERIT (Mutation Error Rate Inference Toolkit), a comprehensive pipeline designed for in-depth quantification of erroneous HTS calls. Using MERIT, we show a significant relationship between substitution error rates and their sequence contexts. In addition to observing more than three orders of magnitude difference between transition and transversion error rates, we identify variations of more than 130-fold within each error type at high sequencing depths. We also propose an *in silico* depth reduction approach to provide insights on estimating optimal depth – where sequencing errors exist at rates similar to those of true mutations. Finally, we propose an assay for detailed assessment of nucleotide-incorporation fidelity for four high-fidelity DNA polymerase molecules.

Methods

DNA sample

We obtained HapMap NA19240 human genomic DNA (5 μg) from Coriell, purified from immortalized lymphocytes using the Qiagen Autopure LS instrument in TE buffer (10 mM Tris, pH 8.0/1 mM EDTA) with concentration of 301 ng/L. We assessed sample quality and concentration using Nanodrop and Qubit dsDNA assays before library preparation.

DNA polymerase enzymes and primer design

We used four high-fidelity DNA polymerase enzymes – NEBNext® High-Fidelity 2X PCR Master Mix (Hi-Fi 2X), NEBNext® Ultra™ II Q5® Master Mix (Ultra II), KAPA HiFi PCR kits with ReadyMix (KAPA), and Invitrogen™ Platinum™ SuperFi™ DNA polymerase (SuperFi) – for PCR amplification. We designed the primers using Primer3 [12] to target four loci in the *TP53* and *SF3B1* genes such that the paired-end reads (R1 and R2) are significantly overlapped (Additional file 1: Tables S1 and S2).

PCR amplification, indexing, and sequencing

We performed twenty PCR cycles using the Hi-Fi 2X, KAPA, and SuperFi polymerases, and 16 cycles using the Ultra II polymerase in the first round of amplification (Additional file 1: Table S3). The cycle numbers were determined after initial PCR amplification tests in order to obtain similar amount of DNA for each enzyme. The second round of PCR for multiplexing and cluster generation included seven cycles for all four polymerases (Additional file 1: Table S4). After each PCR amplification, AMPure Bead cleanup was performed. First, 0.4$\times$ ratio (20 μL AMPure bead to 50 μL PCR product) was used to remove gDNA and larger fragments (i.e., $>$ 600 bp). For the saved supernatant, additional 80 μL AMPure Bead was added to bring the total to a 2$\times$ ratio. The beads were eluted with 22 μL EB (10 mM Tris, pH 8.0). The annealing temperature of 66°C was determined based on the product specificity and yield for all polymerases after performing gradient PCR optimization at eight different temperatures (Additional file 1: Figure S3). Qubit quantification and Bioanalyzer analysis were performed for quality assessment. Custom amplicon-based sequencing and library preparation were performed at GeneWiz (South Plainfield, NJ) using Illumina HiSeq2500 Rapid Run.

MERIT: a comprehensive error rate estimator

Comparative performance analysis of the commonly used HTS variant callers [13–15] suggests a significant disagreement between their identified variants [16–18]. These differences are mainly rooted in each pipeline's specific filtering and statistical methodology. For example, a number of filters is automatically applied to reads by HaplotypeCaller implemented in the Genome Analysis Toolkit (GATK) [13] to exclude uninformative reads from the analysis. This practice is aligned with the goal of the majority of variant callers, which is distinguishing true mutations from the artifacts. However, for a precise quantification of the sequencing noise, i.e., error rate profiling, all the reads need to be included in the analysis as the ultimate goal is understanding the nature of artifacts. SAMtools [14] is the basis of a number of alignment-based variant callers [19, 20], and has high flexibility for changes in its filters. Therefore, MERIT uses SAMtools to identify all positions with alternate alleles from the aligned, indexed sequencing reads. By extracting allele frequencies of substitutions directly from the Pileup file generated by SAMtools mpileup, we make sure all reads are included in the analysis.

As MERIT is designed for ultra-deep HTS applications, the input options of its SAMtools mpileup are set to accommodate high depths while providing the users the ability to modify these parameters based on each sequencing data's characteristics. Additional file 1: Table S5 summarizes the default input parameters of SAMtools mpileup versus those used in MERIT. These parameters allow MERIT to probe SAMtools Pileup data and extract sequencing information for all substitutions, even when they are present in only a single read amongst tens of thousands. Accurate identification of indels is a challenging problem [21, 22] and beyond the scope of this work. Specifically, SAMtools's filtering criteria in introducing and extending gaps, could affect calling complex indels, especially insertions, rendering error rate estimates sequencing depth-dependent.

Next, MERIT obtains the Phred quality score of base substitutions as well as the average Phred quality of bases before and after indels. These quantities are not provided in the VCF files generated by SAMtools. Of note, we observed that the alternate allele and total depths at indel loci are only accurate in SAMtools's Pileup files and not in its VCF. Therefore, to ensure allele frequency accuracy for both indels and substitutions identified, MERIT extracts the reference and alternate alleles' depths as well as the total depths for all the variants from the Pileup file. MERIT also extracts the position-in-read for all variants. Such information, especially in hybrid-capture sequencing, helps to better quantify the source of errors in HTS platforms. An optional annotation step is also available. Finally, MERIT obtains the genomic context of the variants from the reference genome, including the nucleotides immediately at their 5′ and 3′, and estimates error rates for 96 possible single nucleotide substitutions as well as four single-base and 16 double-base insertions/deletions (indels). Details of MERIT's workflow are shown in Fig. 1.

Error rate estimation

We used a single HapMap sample to generate ultra-deep data, and although there may be small, uncharacterized variations within initial cell population, we assumed that all detected variants were errors accumulated in library preparation or during sequencing.

We considered context-specific erroneous base calls at each locus to follow a binomial distribution. More precisely, the probability of a single nucleotide X_i with the genomic context ZX_iZ', $Z, Z' \in \{A, C, T, G\}$, in a specific locus i being misread as Y_i, i.e., $P_{ZX_iZ' \to ZY_iZ'}$ followed

$$P_{ZX_iZ' \to ZY_iZ'} = P(x_i|n_i, p) = \binom{n_i}{x_i} p^{x_i}(1-p)^{n_i - x_i},$$

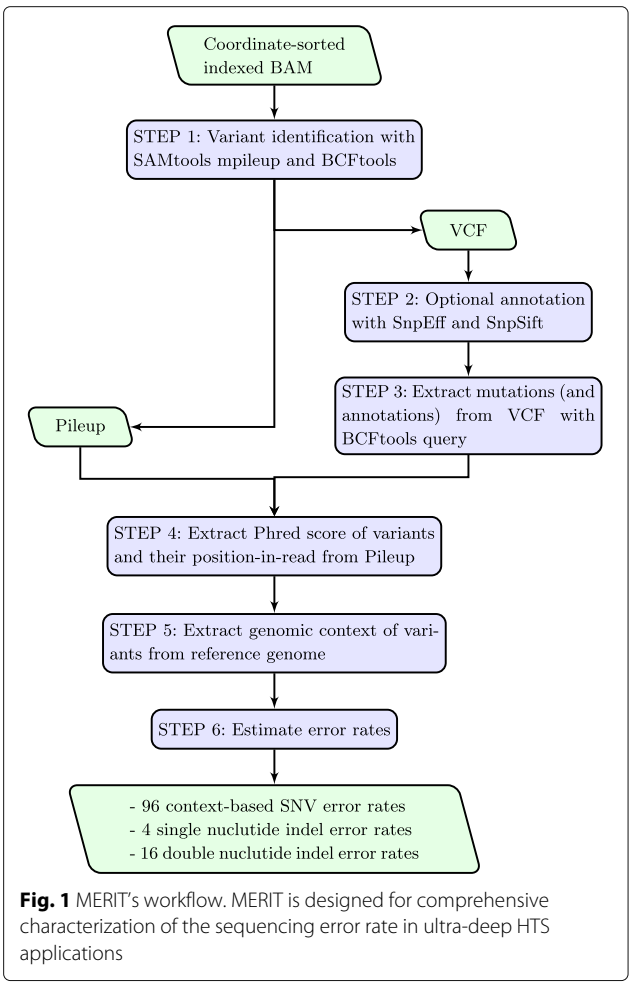

Fig. 1 MERIT's workflow. MERIT is designed for comprehensive characterization of the sequencing error rate in ultra-deep HTS applications

where p is the combined PCR and sequencing error rate and n_i and x_i are the total read depth and the number of erroneous calls at position i, respectively. Assuming a position-independent p, the probability of observing m instances of $ZXZ' \rightarrow ZYZ'$ error within each sample was then given by

$$P_{ZXZ'\rightarrow ZYZ'} = P\left(\sum_{i=1}^{m} x_i | \sum_{i=1}^{m} n_i, p\right) = \binom{\sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} x_i} p^{\sum_{i=1}^{m} x_i} (1-p)^{\sum_{i=1}^{m} (n_i - x_i)}. \quad (1)$$

(See Remark 1 in Additional file 1 on the sum of binomial random variables.) For the case of indels, a binomial model was used to describe the error rate as well, but instead of categorizing them based on their context, indels were classified based on the type of inserted/deleted base, as no differential error rates were observed for context-specific indels.

Polymerase fidelity estimation

The estimated error rate in Eq. (1) has a unit of [error/base]. It is also common to report the fidelity of polymerase enzymes as [error/base/doubling] in the literature where template doubling d is given by

$$2^d = \frac{\text{final DNA amount after PCR}}{\text{starting DNA amount for PCR}}.$$

Since precise amounts of input and output DNA were known for our experiment in its second round of PCR, we calculated template doubling and estimated polymerase replication efficiency as the ratio of template doubling d over the number of PCR cycles performed (Additional file 1: Table S4). To obtain the total amount of template doubling after performing two rounds of PCR amplification, the total number of PCR cycles were multiplied by the polymerase efficiency which resulted in 20.83, 16.19, 16.87, and 20.98 total template doubling for the Hi-Fi 2X, Ultra II, KAPA, and SuperFi polymerases, respectively.

Alignment and merging

We cleaned the paired-end (PE) reads of adapters using bcl2fastq Conversion Software (v2.17), and aligned them to the reference human genome hg19 assembly using the Burrows-Wheeler Aligner (BWA) tool [23] (`bwa sampe` for PE and `bwa samse` for merged reads along with `bwa aln`). We then merged the PE reads that properly mapped to the targeted loci. In our merging scheme, if R1 and R2 reads did not match at a base, an N was assigned for that position. We discarded read pairs with smaller than 50 base overlaps or with more than five mismatches. We calculated Phred quality score (Q) of a successfully merged locus as the sum of the qualities in R1 and R2 reads since these are independent events; Q is given by $Q = -10\log_{10} p$ where p is the probability that the base is called incorrectly. Merged reads were then mapped to the reference human genome hg19 assembly, and were filtered so that they were uniquely mapped (BWA tags `X0:1` and `X1:0`). Finally, in order to make a fair comparison between the error rate of merged and PE reads, we only considered PE reads that were merged successfully and uniquely mapped. Additional file 1: Table S2 represents the average depth of merged and PE reads in different loci. To assess the effect of alternate alignment approaches, we tested Bowtie [24] in addition to BWA to map the merged reads to the reference genome.

In silico depth reduction

The sequencing assay was designed to obtain an average depth of $> 1,000,000\times$ bp, but for some amplicons the average depth was substantially larger (Additional file 1: Table S2). Therefore, an *in silico* depth reduction procedure was performed to reduce the high depths and more importantly, generate enough independent samples to estimate low error rates confidently. It should be noted that one of the main hurdles in error rate estimation of high fidelity polymerases via HTS is the lack of signal as errors occur infrequently with increased fidelity, hence, a large number of samples is required to accurately estimate errors. As performing ultra-deep sequencing on a large number of samples is not cost-effective, alternatively, *in silico* data at lower depths can be generated from one ultra-deep sequencing run by randomly selecting reads from the original raw sequencing data.

Clinical samples

We obtained 29 hematopoietic samples collected from 9 patients with chronic lymphocytic leukemia, previously analyzed by amplicon deep-sequencing (NCBI BioProject PRJNA411889). These samples were sequenced using a custom 88-gene panel, targeting 92 amplicons on Illumina HiSeq (2x150bp) at GeneWiz (South Plainfield, NJ) (Supplementary Table 5 in [25]). The reads were cleaned, merged, and aligned to the reference genome as previously described [25]. We removed previously detected germline and somatic mutations to ensure that the remaining variants represented only the errors.

Results and discussions

Impact of merging reads on context-specific error correction

Independent analysis of R1 and R2 reads at $1{,}300{,}000\times$ indicated significant variations in estimated error rates across 96 possible sequence contexts (Fig. 2). High error rates and low Phred quality scores observed in R2 relative to R1 may be associated with sequencing errors caused by misreading a base, attributed to image analysis

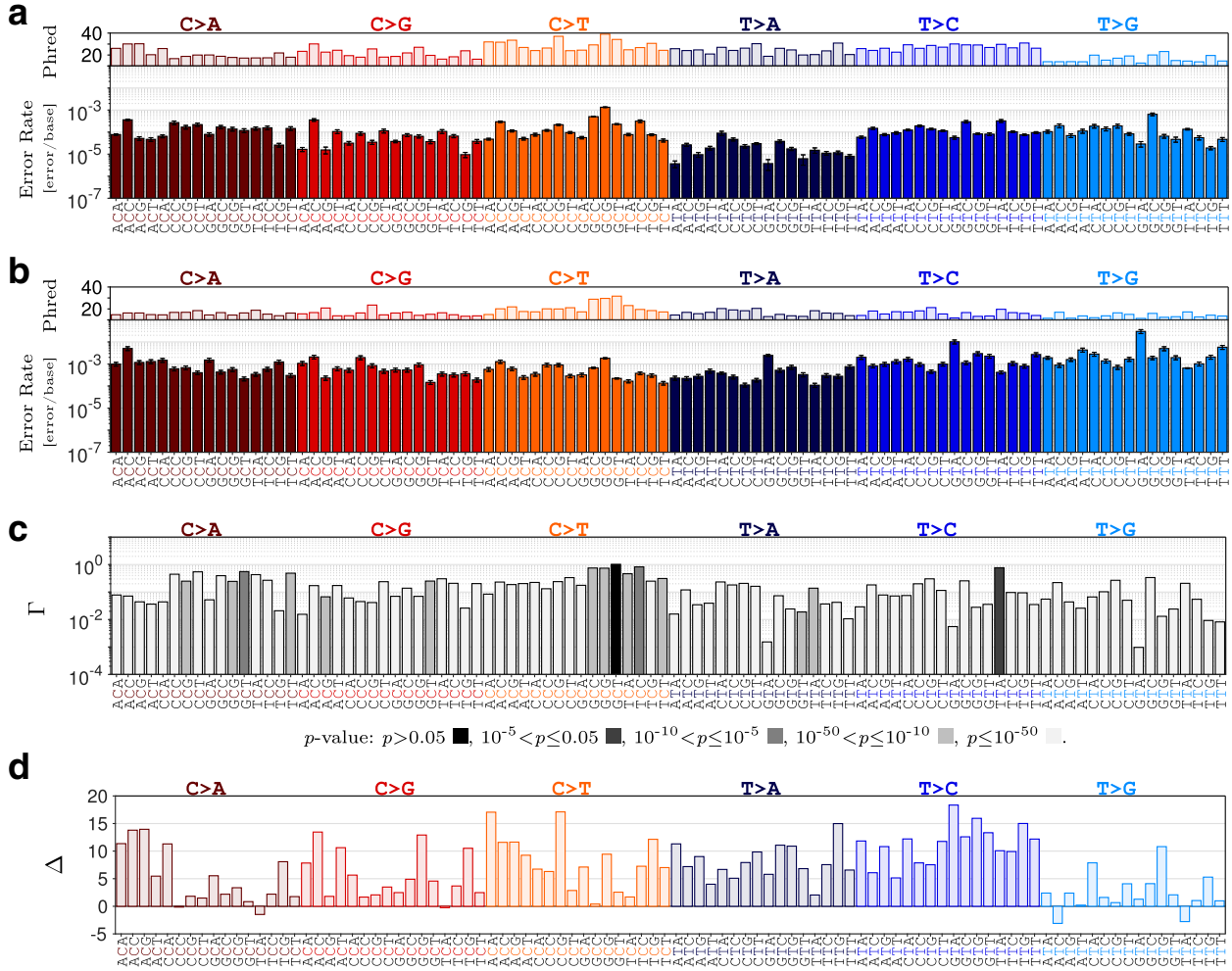

Fig. 2 Estimated context-specific substitution error rates for polymerase Hi-Fi 2X. **a**) R1 reads. **b**) R2 read. **c**) Γ, the ratio of error rate in R1 over R2. *P*-values were computed by performing a two-tailed z-test. **d**) Δ, the difference between their corresponding Phred quality scores. We reduced the depth of paired-end reads to approximately 1,300,000× through an *in silico* depth reduction procedure. Results were obtained by averaging over 100 independent samples to establish error bars which indicate one standard deviation from the average

biases [26] or phasing/pre-phasing [11]. These sequencing errors that dominated the R1 and R2 profiles can be distinguished from polymerase errors by merging the overlapped paired-end reads [27–29]. Merging, however, cannot eliminate errors randomly accumulated during the amplification processes and present in both reads. In contrast to higher rates of transversion versus transition errors in paired-end reads (Fig. 3a), the remaining PCR-related errors in the merged reads were dominated by transitions, often with high Phred quality scores (Fig. 3b). MERIT provides further insight for profiling these errors, which are the main hurdle in distinguishing real mutations from sequencing noise:

- Merging R1 and R2 reads lowered all the context-specific error rates. The highest reduction in rate was observed for GTA>GGA transversions ($5{,}025 \pm 2{,}794\times$) while GCG>GTG transition errors only improved by a factor of $1.22 \pm 0.07\times$. Moreover, these improvements were context-specific. For example, T>A transversion in GTA trinucleotides showed substantial reduction ($568 \pm 249\times$) compared to those in CTAs ($1.43 \pm 0.31\times$).
- Transition errors occurred at higher rates relative to transversions, in agreement with previous reports [7–11]. This difference was pronounced further when errors were classified based on their context, denoting a rate of $1.29 \pm 0.04 \times 10^{-3}$ [error/base] for GCG>GTG versus that of $2.17 \pm 0.92 \times 10^{-6}$ [error/base] for GTA>GAA (Fig. 3b). MERIT also revealed considerable variation within each substitution type. For example, T>G transversions in GTCs occurred $133.5 \pm 65.9\times$ more often than those in ATAs. Similarly, C>T transitions in GCGs were

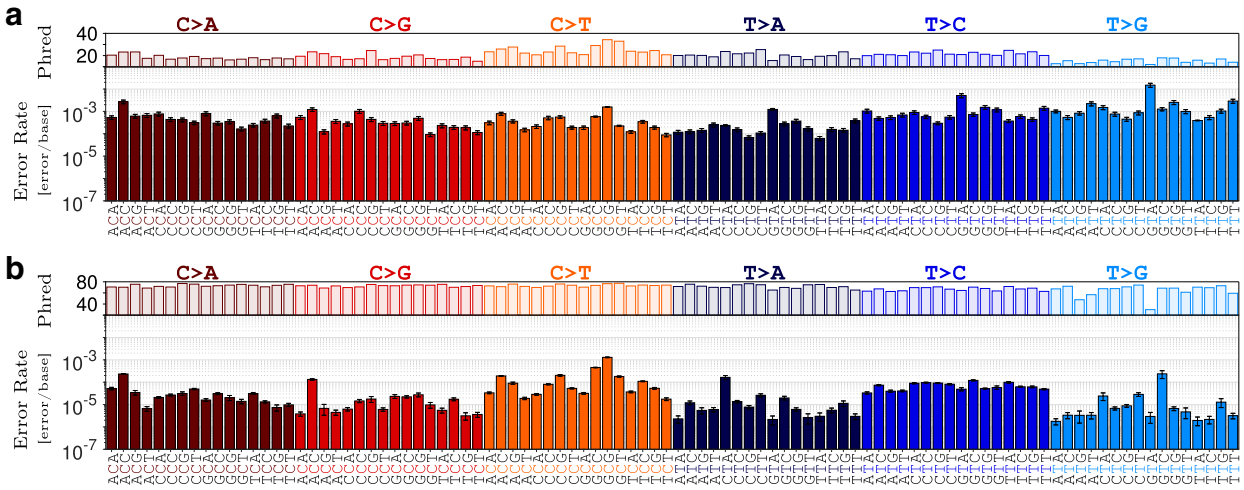

Fig. 3 Merging of overlapped PE reads reduces context-specific error rates. **a**) R1 and R2 reads. **b**) Merged reads. Depth of merged reads for polymerase Hi-Fi 2X were reduced *in silico* to approximately 650,000×. Error bars indicate one standard deviation from the average of 100 independent sub-samples

observed at 73.8± 10.5× higher rate than those in TCTs (Fig. 3b).

- The rate of C>A errors in ACCs was the highest of all such transversions. These errors are linked to the conversion of guanine to 8-oxoG resulting in mismatched pairing with adenine [30, 31]. Oxidation of guanine to 8-oxoG happens naturally in living cells and can be increased by DNA damaging factors such as acoustic shearing [32].
- Merging R1 and R2 can correct for the low quality erroneous bases associated with sequencing errors. Our analysis suggests that such sequencing errors can be identified and eliminated based on their quality, when merging the reads is not possible (e.g., in hybrid-capture-based sequencing where read pairs are not designed to necessarily overlap).

Finally, we tested whether an alternative alignment method, such as Bowtie [24], would affect error rate estimations, and found minimal changes across the 96 genomic contexts (Additional file 1: Figure S4).

Effect of mutation context on amino acid variations

In a single codon, the context-specific rate of error for each base change directly affects the sensitivity of detecting the resulting amino acid variation. Our data indicated that the most commonly mutated residues in *TP53* and *SF3B1* were often more prone to errors and hence comparatively less likely to be distinguished from sequencing errors. For example, in *TP53*, R248Q and R248W are among the most common mutations found in cancer patients [33]. The transition base changes that result in these mutations could be confounded by the HTS errors at an 8-fold higher rate than the transversion alterations that lead to R248L, and 55-fold higher than those that lead to R248G (Fig. 4a). Similarly, the K700E mutation in *SF3B1* is the most frequently mutated residue in the gene's exon 10 [34, 35]; it results from a T>C mutation in a TTC trinucleotide that showed the highest rate of error for a non-synonymous amino acid change in its codon ($4.74\pm 0.42 \times 10^{-5}$ [error/base]). In contrast, the comparatively rarer I704F mutation – a T>A in a ATG reference trinucleotide – had one of the lowest rates of error in its respective codon ($5.15\pm 1.13 \times 10^{-6}$ [error/base]; Fig. 4b). K700E's 9-fold higher rate of error than that of I704F indicated marked reduction in its relative detection sensitivity.

Optimal sequencing depth

Insufficient sequencing depth reduces the sensitivity of detecting variants and leads to loss of statistical significance for a confident variant calling [36]. Consequently, sequencing at higher depths is expected to provide robust error rate estimates and improved sensitivities in detecting true mutations. Accurate estimation of optimal sequencing depth, beyond which the inferred background error is not further reduced, not only provides a precise view of intrinsic limitations in HTS assays, but also leads to preserving time and resources by avoiding unproductive ultra-deep sequencing experiments.

To provide insight on optimal sequencing depth, we performed *in silico* experiments and estimated context-specific error rates as a function of depth. We randomly selected merged reads and constructed simulated sequencing data at depths ranging from 1,000× to 700,000×, with 500 independent replicates at each depth

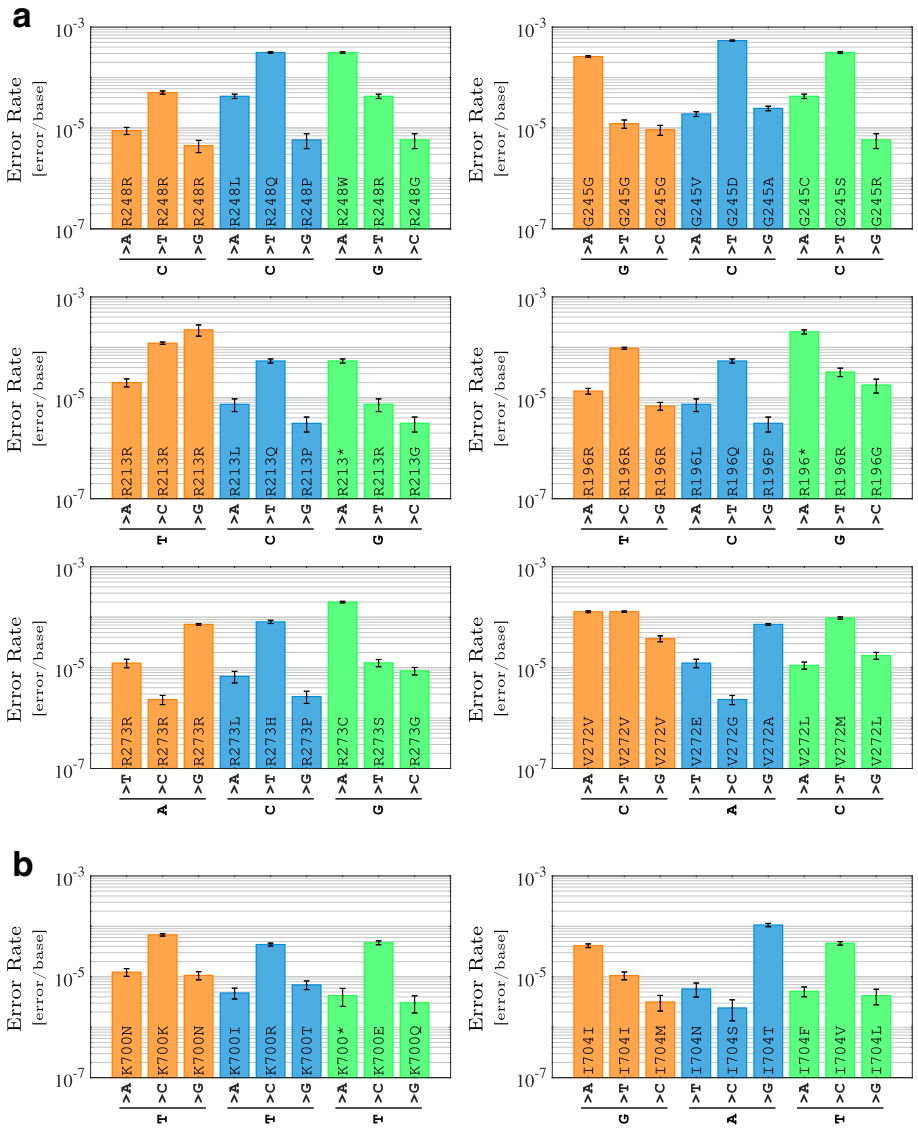

Fig. 4 Significant variation in error rates for possible amino acid changes at individual codons. **a**) Six frequently mutated residues in the *TP53* gene. **b**) Two hotspot residues in the *SF3B1* gene. The higher the rate of error for a specific base change, the lower the power to distinguish true mutations from sequencing artifacts at its position. Here, the error rates represent the amplification by the Hi-Fi 2X polymerase. Error bars represent one standard deviation from the mean of 100 independent sub-samples

to establish confidence intervals (Fig. 5). MERIT showed that the type of substitution error was an important determinant in estimating the optimal depth (Fig. 5a). The error rate estimates for all transitions as well as C>A transversions did not significantly change as sequencing depth increased beyond 200,000×; however, the inferred rates for the remaining transversions marginally improved at higher depths. More importantly, this analysis highlighted the importance of context-specific error profiling in determining detection sensitivity thresholds for true mutations. For example, at 5000×, the corresponding error rates for all T>A errors, T>A errors in CTAs, and T>A errors in GTTs were $2.19 \pm 0.37 \times 10^{-4}$ [error/base], $4.27 \pm 2.28 \times 10^{-4}$ [error/base], and $1.96 \pm 0.02 \times 10^{-4}$ [error/base], while at 700,000×, these rates were reduced to $2.02 \pm 0.73 \times 10^{-5}$ [error/base], $2.5 \pm 0.99 \times 10^{-4}$ [error/base], and $2.1 \pm 0.89 \times 10^{-6}$ [error/base], respectively. Selecting a frequency threshold for these variants at 5000× based on the general T>A rate may not yield significant number of false calls independent of their sequence contexts; however, at depths > 5000×, setting a threshold based on all T>A errors would lead to substantial false positive CTA>CAA and false negative GTT>GAT calls, as their corresponding error rates

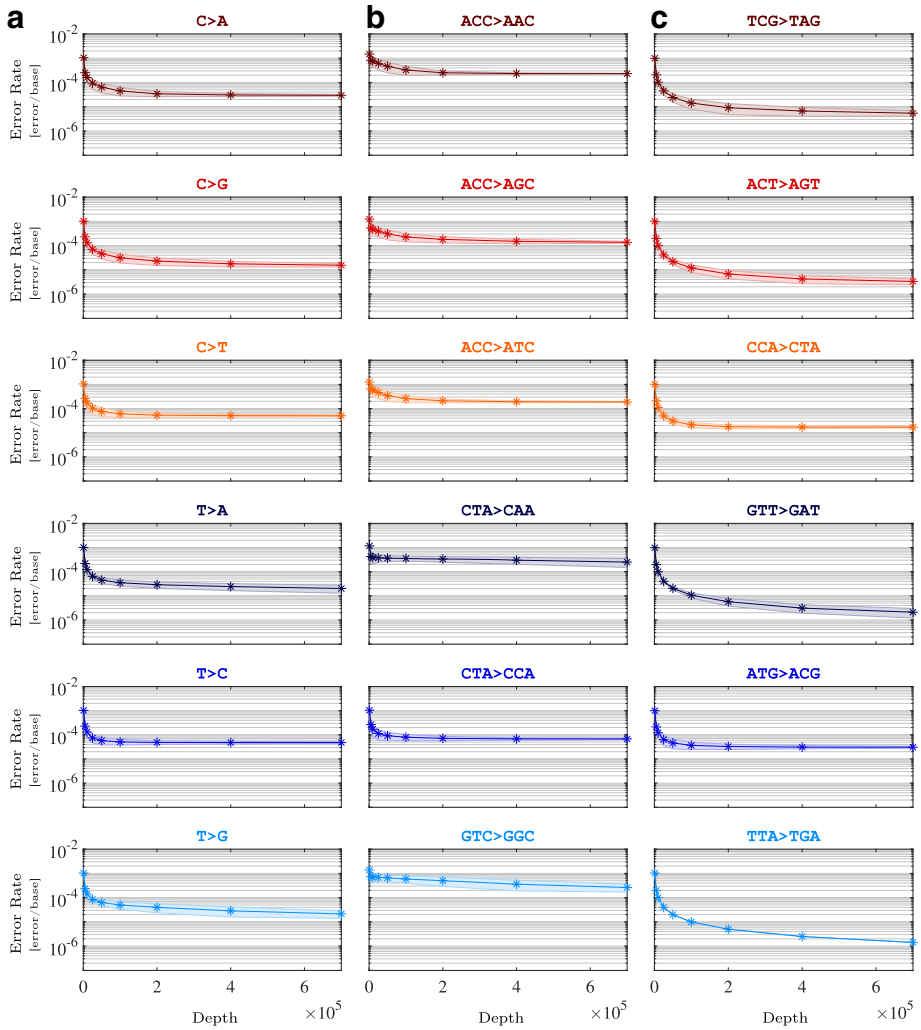

Fig. 5 Context-specific optimal sequencing depth. Substitution error rates are classified based on their type (column **a**) and context (columns **b** and **c**) at nine different depths: 1,000×, 5,000×, 10,000×, 25,0000×, 50,000×, 100,000×, 200,000×, 400,000×, and 700,000×. *In silico* depth reduction procedure was performed on merged reads, amplified by polymerase Ultra II to an average depth of 1,930,473×. The shaded areas are uncertainty bounds of one standard deviation around the average, derived from 500 independent sub-samples

diverge at high depths, reaching a difference of two orders of magnitude at 700,000×.

It should be noted, however, that SAMtools might not be able to detect all indels at all depths [21, 22]. Although comparing indel error rates might be only statistically meaningful at fixed sequencing depths, we did observe a reduction in estimated rate of error for single-nucleotide deletions relative to sequencing depth (Additional file 1: Figure S6). Calling all complex indels, especially when they are present in only a few reads, may require more sophisticated variant callers whose results can be combined with substitution calls to obtain a comprehensive error profile by MERIT.

DNA polymerase fidelity estimation

High-fidelity DNA polymerases – equipped with proofreading – result in fewer base misincorporations in PCR enrichment step, and thus, can reduce HTS error rates. The Hi-Fi 2X, Ultra II, KAPA, and SuperFi enzymes are marketed as high-fidelity polymerases, specifically designed for efficient amplification of complex templates such as those with GC-rich regions. Their providers have reported a fidelity 100× better than wild-type Taq DNA polymerase [37–40].

We applied MERIT to merged reads at equal depths of 650,000×, ensuring that the estimated fidelities were not affected by sequencing depth. When all errors were included in the analysis, global error rates suggested

that these polymerases performed fairly similarly to each other, with the highest and lowest error rates belonging to KAPA and SuperFi enzymes, respectively. Specifically, the global substitution error rates for Hi-Fi 2X, Ultra II, KAPA, and SuperFi were estimated at $2.66\pm 0.21 \times 10^{-6}$, $1.91\pm 0.19 \times 10^{-6}$, $6.95\pm 0.54 \times 10^{-6}$, and $1.76\pm 0.25 \times 10^{-6}$ [error/base/doubling], respectively (Additional file 1: Figure S1a).

Because different assays, quantification methods, and descriptive units [21, 41] are often used to estimate the polymerase fidelity, comparing the reported rates in the literature is a challenging task and beyond the scope of this work. More importantly, error rate profiles in HTS data are reported to be platform as well as batch dependent [42]. For example, using single cell sequencing technique error rates of 5.3×10^{-7} [sub/base/doubling] and 1.6×10^{-5} [sub/base/doubling] are reported in [21] for Ultra II and KAPA polymerases, respectively. In another study [43], a barcoding sequencing approach yielded a rate of 4×10^{-6} [substitutions/base] for Ultra II while 2.8×10^{-7} [substitutions/base] is reported for KAPA enzyme in [39]. Here, we use MERIT to emphasize on the importance of context-specific polymerase fidelity estimation and provide a robust comparison of these commonly used high-fidelity enzymes performed on a single sequencing platform.

Relying solely on global error rates for comparing the replication accuracy of these high-fidelity enzymes may be misleading [44]. Previous HTS-based analyses of polymerase fidelity estimation have classified substitutions into transition and transversion types and have showed preferential rates of error [21, 41, 44, 45]. Additional file 1: Figure S1b represents such classification of the substitution errors in our ultra-deep data, providing a more detailed understanding of the replication fidelity of these enzymes. For example, the global substitution fidelity of SuperFi was found $3.95\pm 0.65\times$ better than that of KAPA's; however, specific substitution fidelity differed widely. C>G errors of SuperFi were $6.88\pm 2.16\times$ less frequent than those of KAPA. In contrast, for C>A substitutions, SuperFi's advantage over KAPA was reduced to only $1.85\pm 0.30\times$.

For a more comprehensive analysis, we used MERIT to estimate 96 context specific substitutions and observed substantial variations (Fig. 6a). For example, TTA>TGA error rate of SuperFi was found $132\pm 35\times$ lower than KAPA, while for GCG>GAG errors, KAPA performed just slightly better than SuperFi. Such classification of substitution errors based on their genomic context enabled us to perform robust statistical comparisons between the replication accuracy of different DNA polymerases using Spearman's rank correlation coefficients presented in Fig. 6d, rather than just comparing them using a single global error rate. Moreover, using the data from multiple regions of the *TP53* and *SF3B1* genes, we found limited change in overall error profiles as the similarities between the genomic content of the amplified amplicons decreased (Fig. 7).

Application of MERIT to clinical samples

Sample preservation and library preparation of clinical samples can lead to specimen-specific errors. MERIT provides a tool to assess and compare such error profiles. Additional file 1: Figure S5 represents the substitution error rates estimated for hematopoietic samples collected from leukemia patients presented in [25]. The error rate estimates for these clinical samples showed a high rate of transition errors similar to previous results from a cell line. A major difference, however, was that the C>A errors proceeded by C and T bases were more frequent than those proceeded by A and G. Our data did not show a preferred 3′ base trailing the misread C base. This high rate of C>A errors has been observed in previous studies [32, 46], specifically an abnormally high rate of CCG>CAG errors in both tumor and normal samples from cancer patients [32].

Conclusions

Novel library preparation methods have succeeded in reducing the background sequencing noise, which has led to improving the sensitivity of detecting true mutations in heterogenous samples. PCR-free library preparation methods [47, 48] have forgone the bias associated with the polymerase base incorporation [49, 50], however, the large amount of input DNA required in these techniques is the main burden for their application in clinical cancer genomic testing. As the exponential PCR amplification is a crucial step in HTS, other techniques have focused on minimizing polymerase errors rather than abolishing the PCR step entirely, including Safe-Seq [6], Duplex-Seq [51], Circle-Seq [52], Cypher-Seq [53], and maximum-depth sequencing [54]. Despite all improvements, the background noise is not completely eliminated. The additional cost and complexity of these methods as well as their lower yield [54] limit their utilization in clinical cancer genomic testing. Specifically, a limited starting material, as is usually the case for tumor specimens, could results in poor sample representation due to inefficiencies in adapte r ligation and loss of genetically diverse small clones [55].

In this paper, we provided a comprehensive method for profiling sequencing artifacts and discussed their impact on accurate variant detection in amplicon-based HTS data. We proposed an approach for determining the optimal sequencing depth, where errors occur at rates similar to those of true mutations. Our data obtained from Illumina platforms confirmed previous results on the differential rates of errors in paired-end sequencing reads [11], and indicated that merging the overlapping

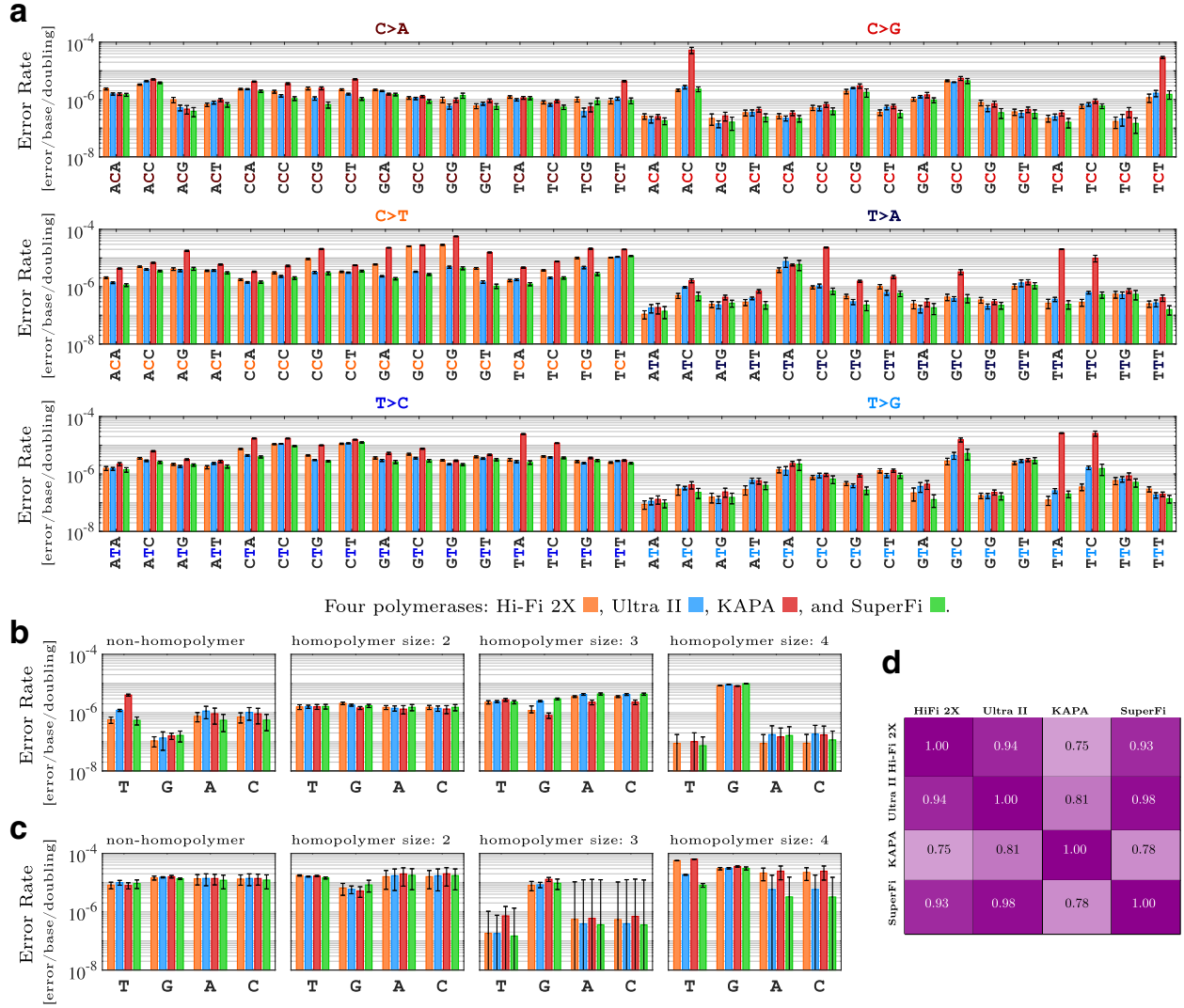

Fig. 6 In-depth comparison of the error rates for four high-fidelity polymerases. **a**) Context-specific substitutions. **b**) Single-base insertions. **c**) Single-base deletions. **d**) Spearman's rank correlation coefficient between context-specific error profiles. Results are obtained by averaging over 100 independent samples to establish error bars, which indicate one standard deviation from the average

read pairs, independent of alignment approach, can notably correct errors that accumulate in sequencing instruments [55].

We also reported the application of MERIT to ultra-deep sequencing data obtained from the amplification of multiple clinically relevant loci using four high-fidelity polymerase enzymes. Although there is limited variation in both the rates of error and dependence on the genomic content of the amplified region, our results indicated that profiling polymerase misincorporation pattern according to genomic context has important clinical consequences. Specifically, we showed that error rates obtained from deep-sequencing of clinical specimens may reflect processes that affect DNA quality during sample preparations.

Sample heterogeneity, especially when low-abundance mutations are present, can confound MERIT's sequencing error profiles. Therefore, when MERIT is applied to clinical specimens from which true mutations are not removed, the estimated rates represent the upper bound of true sequencing error rates. Our results also demonstrated that assigning a single allele frequency threshold to detect mutations may result in substantial false positive as well as false negative calls. Not only were neighboring mutational hotspots in one gene affected with markedly different error rates, there was significant variation in the sensitivity of detecting common amino acid changes within each residue. These data suggested that some of these mutations may in fact be more prevalent at sub-clonal levels in disease populations than previously

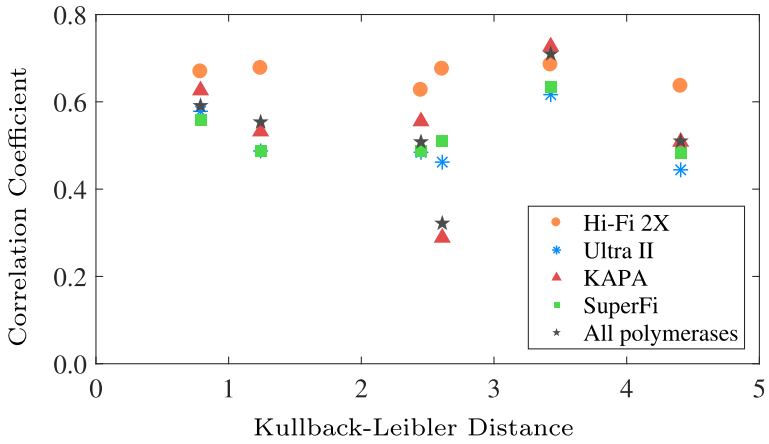

Fig. 7 Relationship between amplicon genomic content and error profiles. Spearman's rank correlation coefficient between the context-specific error profiles of the targeted genes as a function of the symmetric Kullback-Leibler distance between their content profiles presented in Additional file 1: Figure S2

reported. For instance, small mutated clones in the *TP53* gene, present in $> 0.1\%$ of alleles, are shown to be strong predictors of poor survival and possible resistance to therapy in various neoplasms [56–59]; thus, their detection at very low abundances is pertinent for patient care. Put together, our results strongly advocated mutation-specific approaches that go beyond estimating fixed detection thresholds for all variants [60–62].

As deep sequencing of patient samples becomes a routine part of precision medicine in the clinic, we believe that the application of our data-driven pipeline to tumors increases the speed with which patient data can be evaluated for presence of small prognostic mutations, hence, contributing significantly to combating drug resistance and increasing positive outcomes.

Additional file

Additional file 1: SI Materials. (PDF 4187 kb)

Abbreviations

BWA: Burrows-wheeler aligner; GATK: Genome analysis toolkit; Hi-Fi 2X: NEBNext® High-Fidelity 2X PCR master mix polymerase; HTS: High-throughput sequencing; KAPA: KAPA HiFi PCR kits with ReadyMix polymerase; MERIT: Mutation error rate inference toolkit; PCR: Polymerase chain reaction; SuperFi: Invitrogen™ Platinum™ SuperFi™ DNA polymerase; Ultra II: NEBNext® Ultra™ II Q5® master mix polymerase; VAF: Variant allele frequency

Acknowledgments

The authors gratefully acknowledge the constructive feedback by Alexandra Jacunski and assistance from GeneWiz scientific staff. This work was also supported by Rutgers Cancer Institute of New Jersey (P30CA072720) and Rutgers Office of Advanced Research Computing (NIH 1S10OD012346- 01A1).

Funding

MH is a New Jersey Commission on Cancer Research postdoctoral fellow (DFHS17PPC007). HK acknowledges support from the American Caner Society (IRG-15-168-01). The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

MERIT is open-sourced and available at www.software.khiabanian-lab.org and https://github.com/KhiabanianLab/MERIT. All the DNA sequencing data used in this study have been deposited in the NCBI's Sequence Read Archive (SRA) with accession number SRP115798 and NCBI BioProject PRJNA411889.

Authors' contributions

MH and HK conceived the study and designed the algorithm. MH implemented the algorithm and performed the statistical analyses. All authors contributed to the drafting of the manuscript and critical discussion of the results. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 11 December 2017 Accepted: 29 May 2018
Published online: 08 June 2018

References

1. Hagemann IS, Cottrell CE, Lockwood CM. Design of targeted, capture-based, next generation sequencing tests for precision cancer therapy. Cancer Genet. 2013;206(12):420–31. https://doi.org/10.1016/j.cancergen.2013.11.003. Next Generation Sequencing in Clinical Cancer Genomics.
2. Weiss GJ, Hoff BR, Whitehead RP, Sangal A, Gingrich SA, Penny RJ, Mallery DW, Morris SM, Thompson EJ, Loesch DM, Khemka V. Evaluation and comparison of two commercially available targeted next-generation sequencing platforms to assist oncology decision making. OncoTargets Ther. 2015;8:959–967. https://doi.org/10.2147/OTT.S81995.
3. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, Temple-Smolkin RL, Voelkerding KV, Nikiforova MN. Guidelines for validation of next-generation sequencing–based oncology panels. J Mol Diagn. 19(3):341–65. https://doi.org/10.1016/j.jmoldx.2017.01.011.
4. State of New York Health Department. "Next Generation" Sequencing (NGS) guidelines for somatic genetic variant detection (2016).
5. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of next generation sequencing platforms. Next Gener Sequencing Appl. 2014;1:1000106. https://doi.org/10.4172/jngsa.1000106.

6. Kinde I, Wu J, Papadopoulos N, Kinzler K. W, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci. 2011;108(23):9530–9535. https://doi.org/10.1073/pnas.1105422108.
7. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. Sequence-specific error profile of illumina sequencers. Nucleic Acids Res. 2011;39(13):90. https://doi.org/10.1093/nar/gkr344.
8. Shao W, Boltz VF, Spindler JE, Kearney MF, Maldarelli F, Mellors JW, Stewart C, Volfovsky N, Levitsky A, Stephens RM, Coffin JM. Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of low-frequency drug resistance mutations in hiv-1 dna. Retrovirology. 2013;10(1):18. https://doi.org/10.1186/1742-4690-10-18.
9. Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B, Albert J. Pcr-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. PLoS ONE. 2013;8(7):70388. https://doi.org/10.1371/journal.pone.0070388.
10. Dohm J. C, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. Nucleic Acids Research. 2008;36(16):105. https://doi.org/10.1093/nar/gkn42510.1093/nar/gkn425.
11. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. Nucleic Acids Res. 2015. https://doi.org/10.1093/nar/gku134110.1093/nar/gku1341.
12. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):115–115. https://doi.org/10.1093/nar/gks59610.1093/nar/gks596.
13. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. Genome Res. 2010;20(9):1297–303. https://doi.org/10.1101/gr.107524.110.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The sequence alignment/map format and samtools. Bioinformatics. 2009;25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp35210.1093/bioinformatics/btp352.
15. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN]. 2012.
16. Liu X, Han S, Wang Z, Gelernter J, Yang B.-Z. Variant callers for next-generation sequencing data: A comparison study. PLOS ONE. 2013;8(9). https://doi.org/10.1371/journal.pone.0075619.
17. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific Reports. 2015;5:17875.
18. Krøigård AB, Thomassen M, Lænkholm A-V, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS ONE. 2016;11(3):0151664. https://doi.org/10.1371/journal.pone.0151664.
19. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. Varscan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012;22(3):568–576. https://doi.org/10.1101/gr.129684.111.
20. Trifonov V, Pasqualucci L, Tiacci E, Falini B, Rabadan R. Savi: a statistical algorithm for variant frequency identification. BMC Syst Biol. 2013;7 Suppl 2:2. https://doi.org/10.1186/1752-0509-7-S2-S2.
21. Potapov V, Ong JL. Examining sources of error in pcr by single-molecule sequencing. PLoS ONE. 2017;12(1):1–19. https://doi.org/10.1371/journal.pone.0169774.
22. Au CH, Leung AYH, Kwong A, Chan TL, Ma ESK. Indelseek: detection of complex insertions and deletions from next-generation sequencing data. BMC Genomics. 2017;18(1):16. https://doi.org/10.1186/s12864-016-3449-9.
23. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25(14):1754–1760.
24. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol. 2009;10(3):25. https://doi.org/10.1186/gb-2009-10-3-r25.
25. Marsilio S, Khiabanian H, Fabbri G, Vergani S, Scuoppo C, Montserrat E, Shpall EJ, Hadigol M, Marin P, Rai KR, Rabadan R, Devereux S, Pasqualucci L, Chiorazzi N. Somatic cll mutations occur at multiple distinct hematopoietic maturation stages: documentation and cautionary note regarding cell fraction purity. Leukemia. 2017.
26. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. Genome Biology. 2011;12(2):18. https://doi.org/10.1186/gb-2011-12-2-r18.
27. Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. Pandaseq: paired-end assembler for illumina sequences. BMC Bioinformatics. 2012;13:31–31. https://doi.org/10.1186/1471-2105-13-31.
28. Liu B, Yuan J, Yiu S-M, Li Z, Xie Y, Chen Y, Shi Y, Zhang H, Li Y, Lam T-W, Luo R. Cope: an accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. Bioinformatics. 2012;28(22):2870. https://doi.org/10.1093/bioinformatics/bts563.
29. Zhang J, Kobert K, Flouri T, Stamatakis A. Pear: a fast and accurate illumina paired-end read merger. Bioinformatics. 2014;30(5):614–620. https://doi.org/10.1093/bioinformatics/btt593.
30. Ohno M, Sakumi K, Fukumura R, Furuichi M, Iwasaki Y, Hokama M, Ikemura T, Tsuzuki T, Gondo Y, Nakabeppu Y. 8-oxoguanine causes spontaneous de novo germline mutations in mice. Sci Rep. 2014;4:4689.
31. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA. 8-hydroxyguanine, an abundant form of oxidative dna damage, causes g—-t and a—-c substitutions. J Biol Chem. 1992;267(1):166–72.
32. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, Fostel JL, Friedrich DC, Perrin D, Dionne D, Kim S, Gabriel SB, Lander ES, 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. 2013;41(6):67–67. https://doi.org/10.1093/nar/gks1443.
33. Muller PAJ, Vousden KH. p53 mutations in cancer. Nature cell biology. 2013;15(1):2–8. Copyright - Copyright Nature Publishing Group Jan 2013; Last updated - 2014-06-15.
34. Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, Bailey SL, Bhavsar EB, Chan B, Colla S, Corson L, Feala J, Fekkes P, Ichikawa K, Keaney GF, Lee L, Kumar P, Kunii K, MacKenzie C, Matijevic M, Mizui Y, Myint K, Park ES, Puyang X, Selvaraj A, Thomas MP, Tsai J, Wang JY, Warmuth M, Yang H, Zhu P, Garcia-Manero G, Furman RR, Yu L, Smith PG, Buonamici S. Cancer-associated {SF3B1} hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. Cell Reports. 2015;13(5):1033–45. https://doi.org/10.1016/j.celrep.2015.09.053.
35. Wang L, Brooks AN, Fan J, Wan Y, Gambe R, Li S, Hergert S, Yin S, Freeman SS, Levin JZ, Fan L, Seiler M, Buonamici S, Smith PG, Chau KF, Cibulskis CL, Zhang W, Rassenti LZ, Ghia EM, Kipps TJ, Fernandes S, Bloch DB, Kotliar D, Landau DA, Shukla SA, Aster JC, Reed R, DeLuca DS, Brown JR, Neuberg D, Getz G, Livak KJ, Meyerson MM, Kharchenko PV, Wu CJ. Transcriptomic characterization of {SF3B1} mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. Cancer Cell. 2016;30(5):750–763. https://doi.org/10.1016/j.ccell.2016.10.005.
36. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121–132.
37. New England BioLabs Inc. NEBNext High-Fidelity 2X PCR Master Mix. 2017. http://www.international.neb.com. Accessed: 2017-07-06.
38. New England BioLabs Inc. NEBNext Ultra II Q5 Master Mix. 2017. http://www.international.neb.com. Accessed: 2017-07-06.
39. Kapa Biosystems. http://www.kapabiosystems.com. Accessed: 2017-07-06. 2017.
40. Thermo Fisher Scientific. Invitrogen Platinum SuperFi DNA Polymerase. 2017. http://www.thermofisher.com. Accessed: 2017-07-06.
41. Hestand MS, Houdt JV, Cristofoli F, Vermeesch JR. Polymerase specific error rates and profiles identified by single molecule sequencing. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2016;784–785:39–45. https://doi.org/10.1016/j.mrfmmm.2016.01.003.
42. Brandariz-Fontes C, Camacho-Sanchez M, Vilà C, Vega-Pla J, Rico C, Leonard JA. Effect of the enzyme and pcr conditions on the quality of high-throughput dna sequencing results. Scientific Reports. 2015;5:8056.
43. Lee DF, Lu J, Chang S, Loparo JJ, Xie XS. Mapping dna polymerase errors by single-molecule sequencing. Nucleic Acids Res. 2016;44(13):118. https://doi.org/10.1093/nar/gkw436.
44. Shagin DA, Shagina IA, Zaretsky AR, Barsova EV, Kelmanson IV, Lukyanov S, Chudakov DM, Shugay M. A high-throughput assay for

quantitative measurement of pcr errors. Scientific Reports. 2017;7(1):2718. https://doi.org/10.1038/s41598-017-02727-8.
45. McInerney P, Adams P, Hadi MZ. Error rate comparison during polymerase chain reaction by dna polymerase. Molecular Biology International. 2014;12014:1–8. https://doi.org/10.1155/2014/287430.
46. Margolin Y, Shafirovich V, Geacintov NE, DeMott MS, Dedon PC. Dna sequence context as a determinant of the quantity and chemistry of guanine oxidation produced by hydroxyl radicals and one-electron oxidants. The Journal of Biological Chemistry. 2008;283(51):35569–35578. https://doi.org/10.1074/jbc.M806809200.
47. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free illumina sequencing-library preparation facilitates improved mapping and assembly of (g+c)-biased genomes. Nat Meth. 2009;6(4):291–295.
48. Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TWB, Collins JE, Turner DJ. Frt-seq: amplification-free, strand-specific transcriptome sequencing. Nat Meth. 2010;7(2):130–132.
49. Huptas C, Scherer S, Wenning M. Optimized illumina pcr-free library preparation for bacterial whole genome sequencing and analysis of factors influencing de novo assembly. BMC Research Notes. 2016;9:269. https://doi.org/10.1186/s13104-016-2072-9.
50. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011;12(10):671–682.
51. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(36):14508–14513. https://doi.org/10.1073/pnas.1208715109.
52. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL. High-throughput dna sequencing errors are reduced by orders of magnitude using circle sequencing. In: Proceedings of the National Academy of Sciences of the United States of America; 2013. p. 19872–19877. https://doi.org/10.1073/pnas.1319590110.
53. Gregory MT, Bertout JA, Ericson NG, Taylor SD, Mukherjee R, Robins HS, Drescher CW, Bielas JH. Targeted single molecule mutation detection with massively parallel sequencing. Nucleic Acids Res. 2016;44(3):22. https://doi.org/10.1093/nar/gkv915.
54. 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. 2016;534(7609):693–696.
55. Chen-Harris H, Borucki MK, Torres C, Slezak TR, Allen JE. Ultra-deep mutant spectrum profiling: improving sequencing accuracy using overlapping read pairs. BMC Genom. 2013;14(1):96. https://doi.org/10.1186/1471-2164-14-96.
56. Cazzola M, Rossi M, Malcovati L. on behalf of the Associazione Italiana per la Ricerca sul Cancro Gruppo Italiano Malattie Mieloproliferative: Biologic and clinical significance of somatic mutations of sf3b1 in myeloid and lymphoid neoplasms. Blood. 2013;121(2):260–9. https://doi.org/10.1182/blood-2012-09-399725.
57. 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. 2014;123(14):2139–47. https://doi.org/10.1182/blood-2013-11-539726.
58. 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. 2016;101(4):135–8. https://doi.org/10.3324/haematol.2015.136051.
59. 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 MJ, Colado E, Puente XS, 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. 2016;127(17):2122–2130. https://doi.org/10.1182/blood-2015-07-659144.
60. Rabadan R, Bhanot G, Marsilio S, Chiorazzi N, Pasqualucci L, Khiabanian H. On statistical modeling of sequencing noise in high depth data to assess tumor evolution. J Stat Phys. 2017. https://doi.org/10.1007/s10955-017-1945-1.
61. 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. 2013;41(7):89–89. https://doi.org/10.1093/nar/gkt126.
62. Young AL, Challen GA, Birmann BM, Druley TE. Clonal haematopoiesis harbouring aml-associated mutations is ubiquitous in healthy adults. Nat Commun. 2016;7:12484.