Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/50/eabe3722/DC1

Supplementary Materials for

Integration of intra-sample contextual error modeling for improved detection of somatic mutations from deep sequencing

Sagi Abelson*, Andy G. X. Zeng, Ido Nofech-Mozes, Ting Ting Wang, Stanley W. K. Ng, Mark D. Minden, Trevor J. Pugh, Philip Awadalla, Liran I. Shlush, Tracy Murphy, Steven M. Chan, John E. Dick*, Scott V. Bratman*

*Corresponding author. Email: sagi.abelson@oicr.on.ca (S.A.); scott.bratman@rmp.uhn.ca (S.V.B.); john.dick@uhnresearch.ca (J.E.D.)

Published 9 December 2020, *Sci. Adv.* **6**, eabe3722 (2020)
DOI: 10.1126/sciadv.abe3722

The PDF file includes:

Supplementary Notes
Figs. S1 to S9

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/50/eabe3722/DC1)

Tables S1 to S9

SUPPLEMENTARY INFORMATION:

Supplementary Note

Effect of "Pre-filters" on error modeling

The growing catalogs of somatic mutations in cancer and the large databases of human germline variants make it possible to define genomic loci more likely to be associated with authentic mutations. Therefore, one can attempt to incorporate this knowledge by removing those alleles to refine error modeling for increased sensitivity. Thus, a set of filters was incorporated in the Espresso pipeline helping in mitigating the inclusion of real mutated alleles in contextual error models (Materials and Methods). To determine the effect of our filtering step on the accuracy of mutation detection, we analyzed the cell line dilution dataset with and without applying the filters. When the filters were not used, an expected decrease in sensitivity was observed (Fig S4A).

Alleles with a high number of supporting reads are relatively infrequent as compared with the high number of typical errors observed in hybrid capture sequencing data. Yet, they can result in poorer overall model performance if not treated separately. To shed light on the observed sensitivity decrease, we focused on the effect of germline mutations on contextual model fitting following the removal of the MAF>=0.1% filter. In a representative sample, we measured the goodness of fit for the A[C>T]G contextual model as it is expected to be populated with a high number of germline alleles (Fig S4B). When the filter was removed, the model failed to appropriately match the observed and expected number of nonreference supporting reads (Fig S4C). These results illustrated that failure to enrich for error alleles prior to model derivation would compromise the detection of real mutation with relatively low read support.

Selecting the appropriate distribution fit for nonreference supporting reads

In implementing Espresso for a particular dataset, we assess the goodness of fit of different distribution models to the data. We found that any single model cannot appropriately account for the different proportions of nonreference supporting reads in the various datasets tested. Selecting the appropriate model to match the distribution of nonreference supporting reads in the dataset (Fig. 3B, Materials and Methods) provided a more powerful and robust approach (Fig. S5).

Advantage of modeling the number of nonreference supporting reads

We evaluated whether modeling errors by their number of nonreference supporting reads (SR) will better mitigate the effect of varying coverage on mutation detection as compared with allele frequency-based modeling (Fig. S1C). Using the AML-MRD cohort, for each mutation that was reported in the diagnostic samples, we determined its VAF and SR in the follow-up samples of the patients. We then checked how many other alleles with the same genomic context had been reported with either a higher VAF or a higher number of SR. We observed that for the majority of the diagnostic alleles, higher number of other contextual nonreference alleles were reported with higher VAF as compared with higher number of SR (Fig. S6A). These results imply that probabilistic error models that are based on the VAF parameter are likely to result in a longer right-skewed distribution tail and, therefore, potentially a higher number of false negative results as compared with models that are based on the SR parameter. These results were also replicated using a random selection of alleles (Fig. S6B), illustrating that the observed advantage of using SR-based models is not specific to the reported diagnostic mutations, and it is rather a robust approach to also mitigate false positive calls.

Supplementary Figures

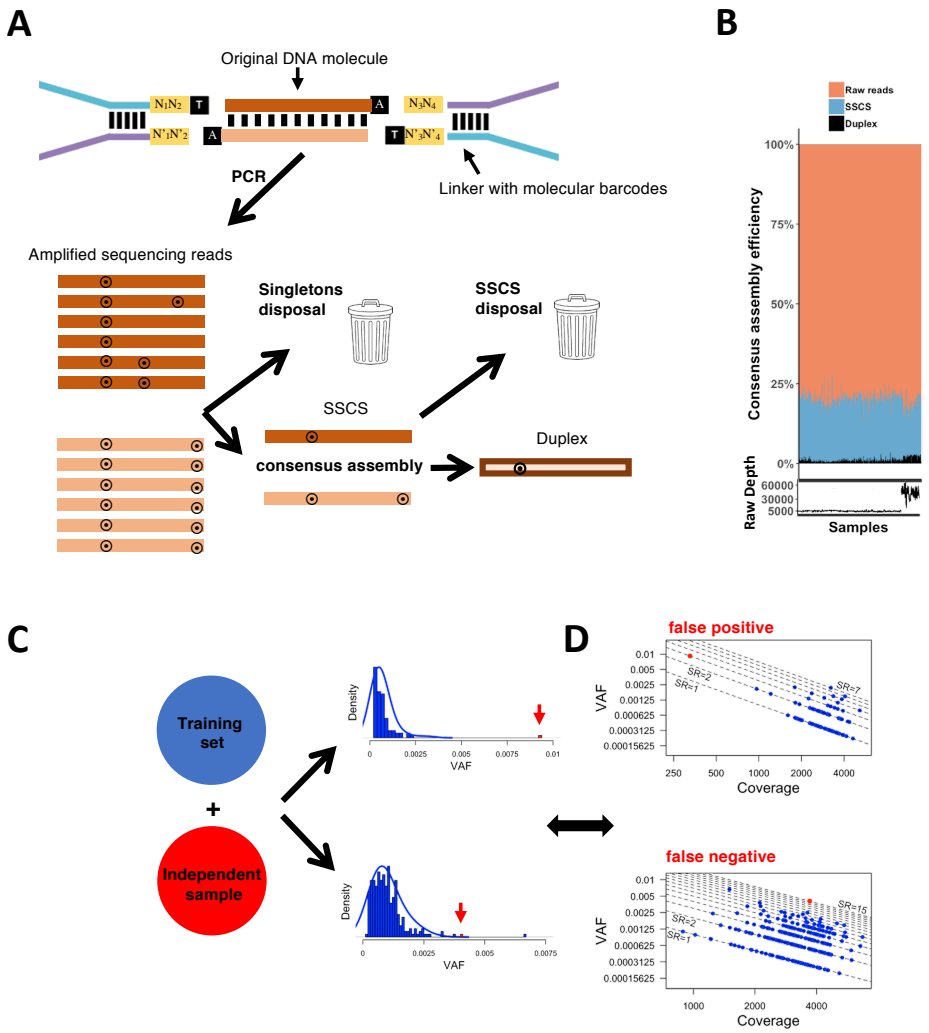

Fig. S1. Intrinsic limitations of common error suppression techniques. (**A**) Diagram illustrating the design and usage of unique molecular adapters for single-stranded consensus sequences (SSCS) and duplex implementation. During each step, lower-level reads that cannot form SSCS or duplex sequences are typically being disposed to achieve a higher level of confidence in mutation calls. Disposal of such reads results in information loss. (**B**) Low SSCS and duplex assembly efficiency (calculated as the number of SSCS or duplex reads divided by all the number of raw reads) results in lower depth. In turn, lower depth due to "lower-level" read disposal may translate to lower sensitivity and high detection limit levels. (**C**) Illustration of training set usage approach for error rate estimation. When sequencing depth is not considered, probabilistic error models may suffer from higher rates of type 1 and type 2 errors. (**D**) Modeling the number of supporting nonreference reads (SR) better captures the relationship between VAF and coverage thus helping with mitigating this issue (**Supplementary Note**). VAF (also referred to as error rates in the manuscript) and coverage (also referred to as depth in the manuscript) are plotted on a log-log scale

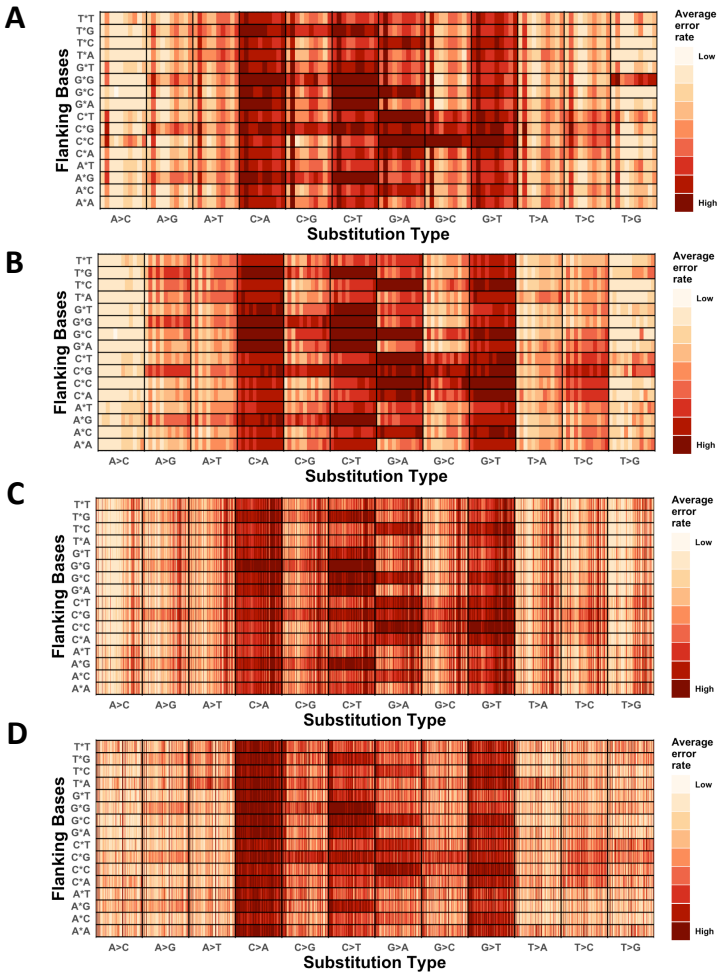

Fig. S2. Error rates differ at different trinucleotide sequence contexts. Mean error rates per sample are shown for each one of the 192 trinucleotide contextual error types across the (**A**) CB, (**B**) CL, (**C**) pre-AML1, and (**D**) pre-AML2 datasets. Vertical lines in each contextual error box represent individual samples, and their order is maintained in each box.

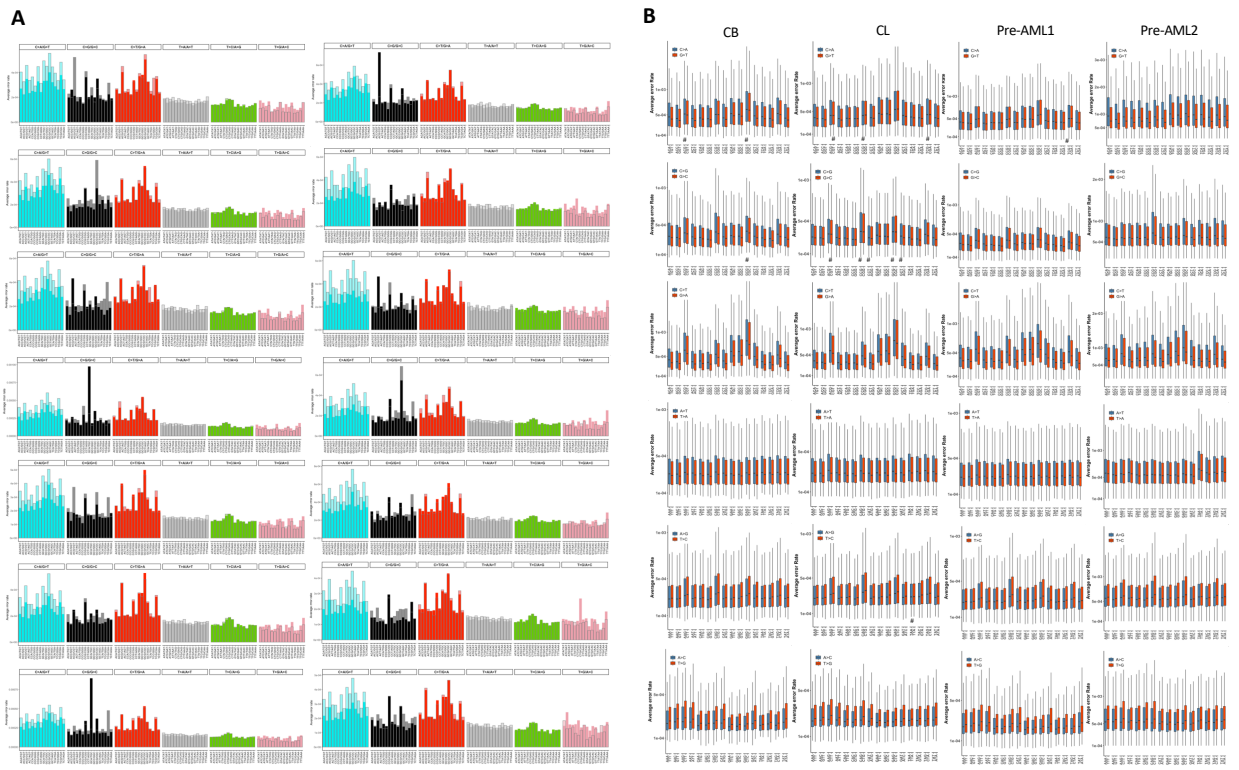

Fig. S3. Error rates asymmetries. (**A**) Contextual error rates and their reciprocals are superimposed, demonstrating intra-sample variation. Results are illustrated for 14 control samples that were sequenced together with the AML-MRD dataset. (**B**) Error rates vary significantly between error contexts and their reciprocals (Wilcoxon rank-sum test, P<0.05. # sign indicates that significance was not reached). Results from the CB, CL, pre-AML1 and pre-AML2 cohorts are shown.

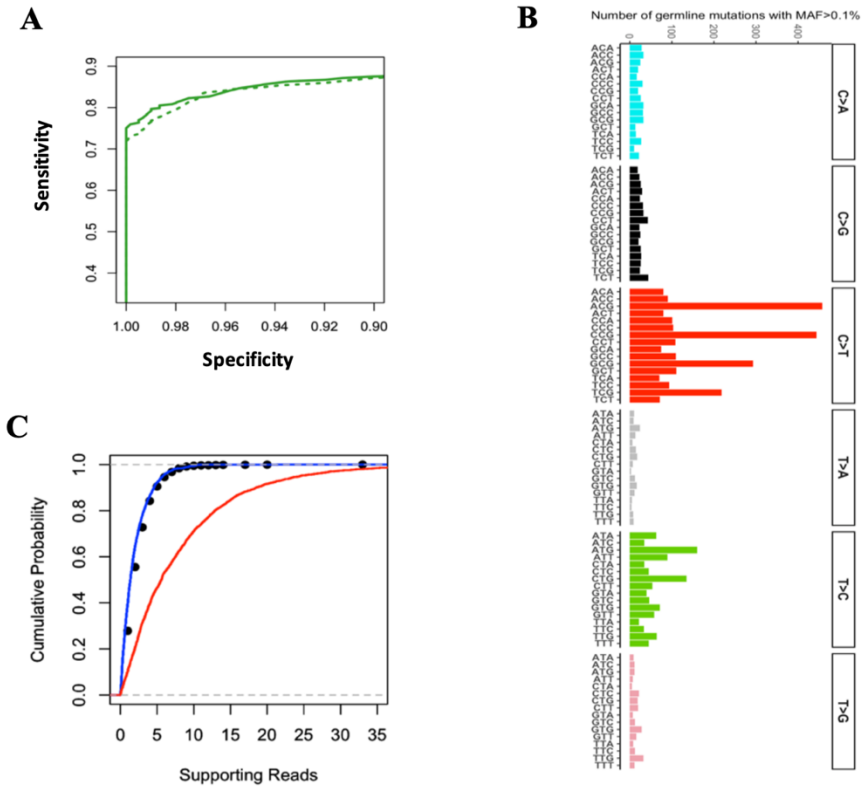

Fig. S4: Incorporating pre-filtering steps improves error modeling accuracy and performance. (A) Espresso's performance on the CL dataset when the 5 pre-filters were applied before error modeling (solid line, also reported in the main text, Fig4. A) and when they were not applied (dashed lines). **(B)** Number of contextual germline alleles defined as mutations with MAF>=0.1% either in 1000 genomes *(50)* or Kaviar *(51)* databases that are covered by the targeted panel. There are a high number of affected genomic loci corresponding to the N[C>T]G context that typically arises from methylation-mediated deamination of 5-methylcytosine. **(C)** The Cumulative Distribution Function graph displays the empirical data for the A[C>T]G genomic context in a representative sample from the CL dataset (black dots, nonreference supporting reads). The models' contextual allele counts estimates are shown when alleles with MAF>=0.1% were omitted from the sample's data (Blue line) and when these alleles were not omitted (Red line).

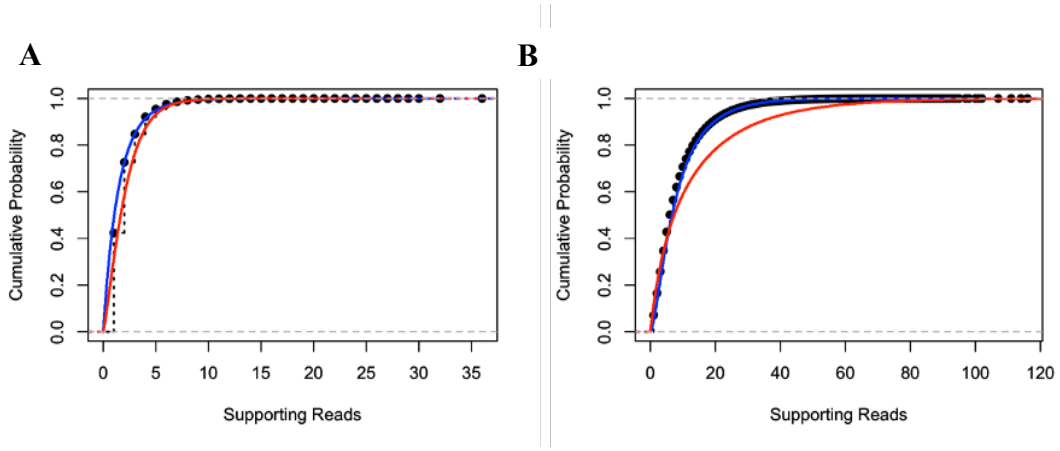

Fig. S5: Improved fit by automatic probabilistic distribution model selection. As described in Fig. 3B, the appropriate model is selected based on the modal nonreference supporting reads across the interrogated positions (exponential model for nonreference supporting reads = 1; Weibull model for nonreference supporting reads >1). The Cumulative Distribution Function graphs illustrate the goodness of fit. **(A)** The appropriate exponential model's fit (blue) to the CB data (n=10 samples) and the less suitable Weibull model's fit (red) are illustrated. **(B)** The appropriate Weibull models' fit (blue) to the AML-MRD control sample set (n=14 samples) and the less suitable exponential model's fit (red) are illustrated. Black dots represent the actual number of nonreference supporting reads observed in the samples. The merged data derived from all the contextual models is shown.

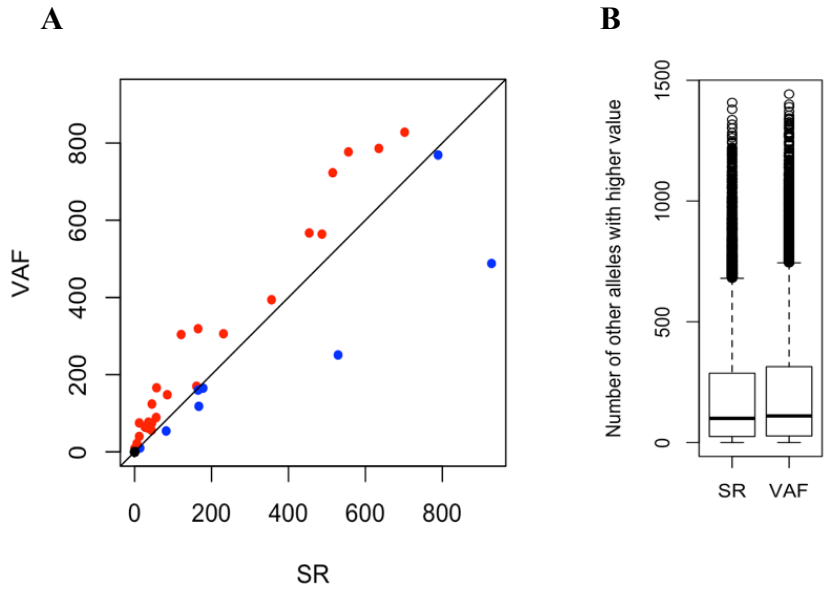

Fig. S6: Error modeling based on the number of nonreference supporting reads is likely to result in lower numbers of false positive and negatives calls as compared with VAF-based modeling. (**A**) Each dot represents a single mutation in the AML-MRD cohort that was reported in the diagnostic samples. Dot location represents the number of alleles with the same genomic

context as the diagnostic mutations that were detected with higher VAF (y-axis) or higher number of nonreference supporting reads (SR) (x-axis) in the patient's follow-up samples. Black dots (x=y), represent that an equal number of alleles with the same genomic context as the mutation were detected with higher VAF as compared with alleles with higher number of nonreference SR. For red dots (above the line of equality), a greater number of alleles were detected with higher VAF as compared with alleles with higher number of nonreference SR. For blue dots (below the line of equality), a lesser number of alleles were detected with higher VAF as compared with alleles with higher number of nonreference SR. (**B**) Single alleles were randomly selected by iterating over every trinucleotide contextual pattern in every sample in the AML-MRD dataset resulting in 18,432 interrogated alleles. In each sample, the sums of alleles with the same genomic context as the randomly queried allele with higher VAF or higher SR were calculated. A significant higher number of contextual alleles with with higher VAF as compared with alleles with higher SR is observed (Wilcoxon signed-rank test: $P = 4.9 \times 10^{-9}$).

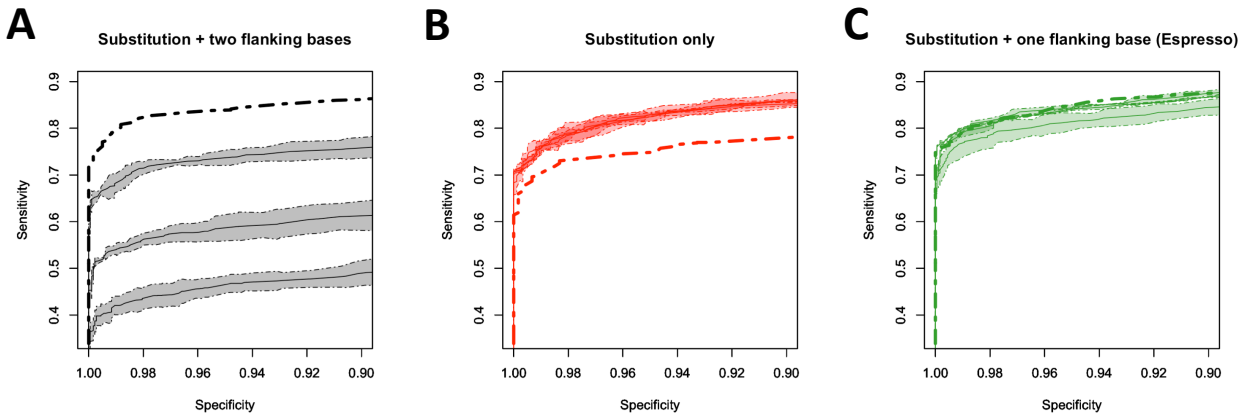

Fig. S7. Sensitivity versus specificity of various contextual error modeling approaches. **(A,B,C)** Sensitivity versus specificity trade-offs derived by Espresso, and the reduced and extended contextual error modeling approaches are illustrated at serially decreased panel sizes. The figure illustrates the same data as in Fig. 4E-H, yet in here, the data is being grouped based on the modeling approached rather than the panel size.

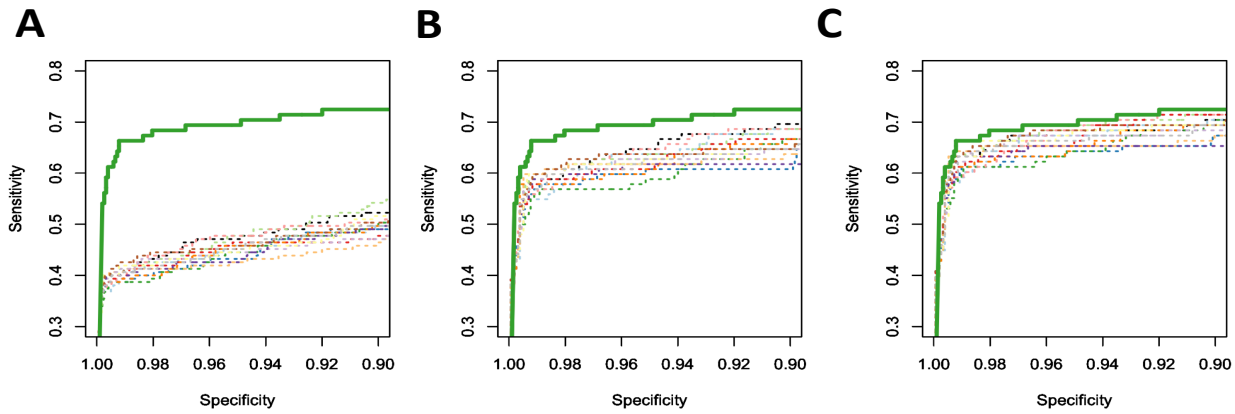

Fig. S8. Performance comparisons with deepSNV. Variable performance was calculated with deepSNV when different control samples were used (dashed color lines, n=14 iterations). A solid green line represents the performance of Espresso. deepSNV performance is being shown (**A**) at default settings. P-values were determined for each observed nonreference allele in the 78 genomic positions reported to be mutated in the AML-MRD cohort at AML diagnosis. Sensitivities and specificities are illustrated at each p-value cut-off. (**B**) deepSNV's performance is illustrated when only the most abundant nonreference allele in each one of the investigated genomic positions in the sample of interest is being considered. Other nonreference alleles that were reported following the analysis of the raw samples bams files by deepSNV were automatically considered as true-negatives. (**C**) deepSNV's performance, when only the most abundant nonreference alleles reported by our bioinformatics pipeline are being considered. The increased performance indicates the added value of our bam files preprocessing and allele prefiltering schemes described herein (see Materials and Methods**)**.

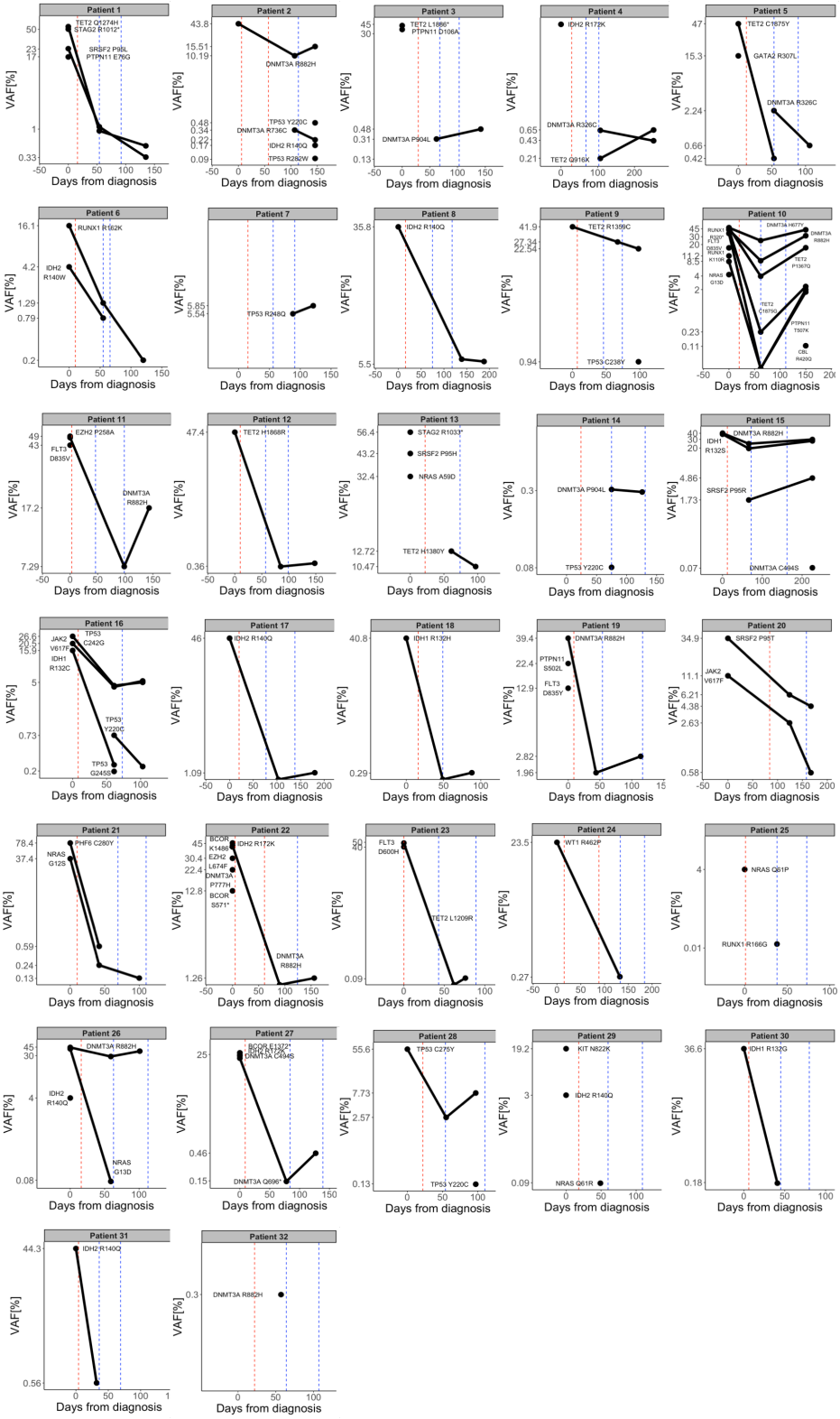

Fig. S9. Longitudinal detection of clonal and sub-clonal SNVs in AML patients. Espresso detected SNVs in 32 out of the 42 individuals recruited to this study. For each patient, the mutations' VAF is plotted at a time point relative to the time that followed diagnosis. Variants that supposedly existed or detected at multiple time points are connected with a black line (alpha≤0.05, Bonferroni adjusted. specificity=99.47% as estimated from the diagnostic mutations). Red and blue vertical lines represent induction and consolidation treatment rounds, respectively.