GENETICS

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

Sagi Abelson[1,2]*, Andy G. X. Zeng[2,3], Ido Nofech-Mozes[1,2], Ting Ting Wang[3,4], Stanley W. K. Ng[5], Mark D. Minden[3,4], Trevor J. Pugh[1,3,4], Philip Awadalla[1,2], Liran I. Shlush[6,7], Tracy Murphy[3], Steven M. Chan[3,4], John E. Dick[2,3]*†, Scott V. Bratman[3,4,8]*†

Sensitive mutation detection by next-generation sequencing is critical for early cancer detection, monitoring minimal/measurable residual disease (MRD), and guiding precision oncology. Nevertheless, because of artifacts introduced during library preparation and sequencing, the detection of low-frequency variants at high specificity is problematic. Here, we present Espresso, an error suppression method that considers local sequence features to accurately detect single-nucleotide variants (SNVs). Compared to other advanced error suppression techniques, Espresso consistently demonstrated lower numbers of false-positive mutation calls and greater sensitivity. We demonstrated Espresso's superior performance in detecting MRD in the peripheral blood of patients with acute myeloid leukemia (AML) throughout their treatment course. Furthermore, we showed that accurate mutation calling in a small number of informative genomic loci might provide a cost-efficient strategy for pragmatic risk prediction of AML development in healthy individuals. More broadly, we aim for Espresso to aid with accurate mutation detection in many other research and clinical settings.

INTRODUCTION

The process of single-nucleotide variant (SNV) accumulation is an important universal element of cancer initiation and progression. While the genetic landscape of the most common malignancies has been broadly described (*1*–*3*), accurate identification of driver mutations in specimens with low cancer DNA purity continues to be of great importance yet presents substantial challenges. Hybrid-capture–based next-generation sequencing (NGS) is one of the most common techniques being used for circulating tumor DNA profiling (*4*, *5*), detection of therapy-resistant clones (*6*, *7*) and pre-leukemia (*8*, *9*), and monitoring disease burden during therapy (*10*). Nevertheless, in all of these settings, the relevant genomic alterations typically exist at low relative abundance.

Several different methods have been developed in recent years to address the barrier of identifying the minute fraction of DNA molecules harboring an alteration against the high background of NGS-associated errors. Among the various methods, state-of-the-art techniques for error suppression typically can be categorized into two groups: (i) those that incorporate unique molecular identifiers (UMIs) to suppress library amplification errors by the assembly of consensus sequences (*11*–*13*) and (ii) those that use probabilistic models to estimate background sequencing noise. The latter group can be further segregated into those that generate models that estimate error rates by the analysis of data from a single sample (i.e., single sample/tumor-only mode) (*14*–*16*), data from a single control sample (*16*–*18*), or data from multiple control samples (e.g., cohort of healthy controls) (*19*–*21*). In the case of paired patient's tumor and matched normal sample, Bayesian statistics models are commonly used to identify tumor-specific somatic variants that are distinguishable from the background and the germline variants detected within the matched normal sample (*22*, *23*). Some techniques rely on a ploidy assumption to calculate genotype probabilities (*24*), while others have adapted statistical models to analyze allele frequencies directly (*16*), thus allowing the identification of rare subclones in existing, complex cancer genomes. Since a single control sample cannot fully account for the stochastic nature of NGS errors, other algorithms have been developed to generate site-specific error estimations using a larger cohort of controls (*19*–*21*). This approach could be problematic as proper control samples are not always available. When control samples are completely lacking, stringent preprocessing steps can be applied to prioritize high-confidence mutations, for instance, thresholds on base quality scores, supporting read counts, and variant allele frequencies.

Despite advances enabled by the diverse approaches mentioned above, each is associated with inherent disadvantages that can lead to increased assay complexity, elevated sequencing costs, and/or suboptimal exchange between sensitivity and specificity (fig. S1, table S1, and Supplementary Note). To overcome these limitations, we characterized the contextual patterns of high-frequency errors observed during targeted hybrid-capture NGS in >1000 samples, divided across multiple technically diverse and clinically relevant human cohorts. On the basis of these patterns, we developed Espresso, a novel UMI-independent method that optimizes the suppression of artifacts from deep NGS for accurate SNV mutation calling.

RESULTS

Evaluation of error abundance and rates in multiple NGS datasets

To demonstrate the challenges associated with low–variant allele fraction (VAF) mutation calling from hybrid-capture–targeted NGS,

[1]Ontario Institute for Cancer Research, Toronto, ON, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. [3]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada. [4]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada. [5]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. [6]Division of Hematology, Rambam Healthcare Campus, Haifa, Israel. [7]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [8]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada.
*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.)
†These authors contributed equally to this work as co–senior authors.

we interrogated multiple benchmarking datasets that differ by their library preparation techniques, captured genomic loci, number of samples, and sequencing depths (Fig. 1A, table S2, and Materials and Methods). Briefly, these datasets include the following: (i) CB: a human cord blood dataset; (ii) CL: a cell line dilution series using genomic DNA from the acute myeloid leukemia (AML) cell line MOLM13 and the colon cancer cell line SW48; (iii and iv) pre-AML1 and pre-AML2: peripheral blood DNA from two separate cohorts, each composed of pre-AML cases (that is, blood was drawn before clinical diagnosis of AML) and age- and sex-matched controls (*9*); and (v) AML-MRD: a cohort composed of peripheral blood DNA samples obtained from patients with AML during the course of treatment.

Three different target panels were used to sequence these cohorts, resulting in 83,000 to 1.2 million interrogated bases (table S2). Investigating these genomic loci revealed that the percentage of positions with nonreference alleles per sample varied widely among the different datasets and, in some cases, among samples within a particular dataset (Fig. 1B). Samples with a lower percentage of positions with nonreference alleles displayed higher average error rates (Fig. 1C). Furthermore, almost all genomic positions sequenced harbored a nonreference allele in at least one sample in each dataset (Fig. 1D). Overall, these observations reveal the magnitude of the challenge presented by potential false-positive variants produced by hybrid-capture NGS. Since such a large number of technical artifacts may mask clinically relevant variants, we conducted an unbiased exploration of multiple strategies aiming to specifically suppress NGS errors while maintaining high sensitivity in identifying real mutations.

Error rates and sequencing depth vary according to different sequence contexts

To evaluate the contextual dependencies of errors in the datasets described above, we investigated how error rates differ with respect

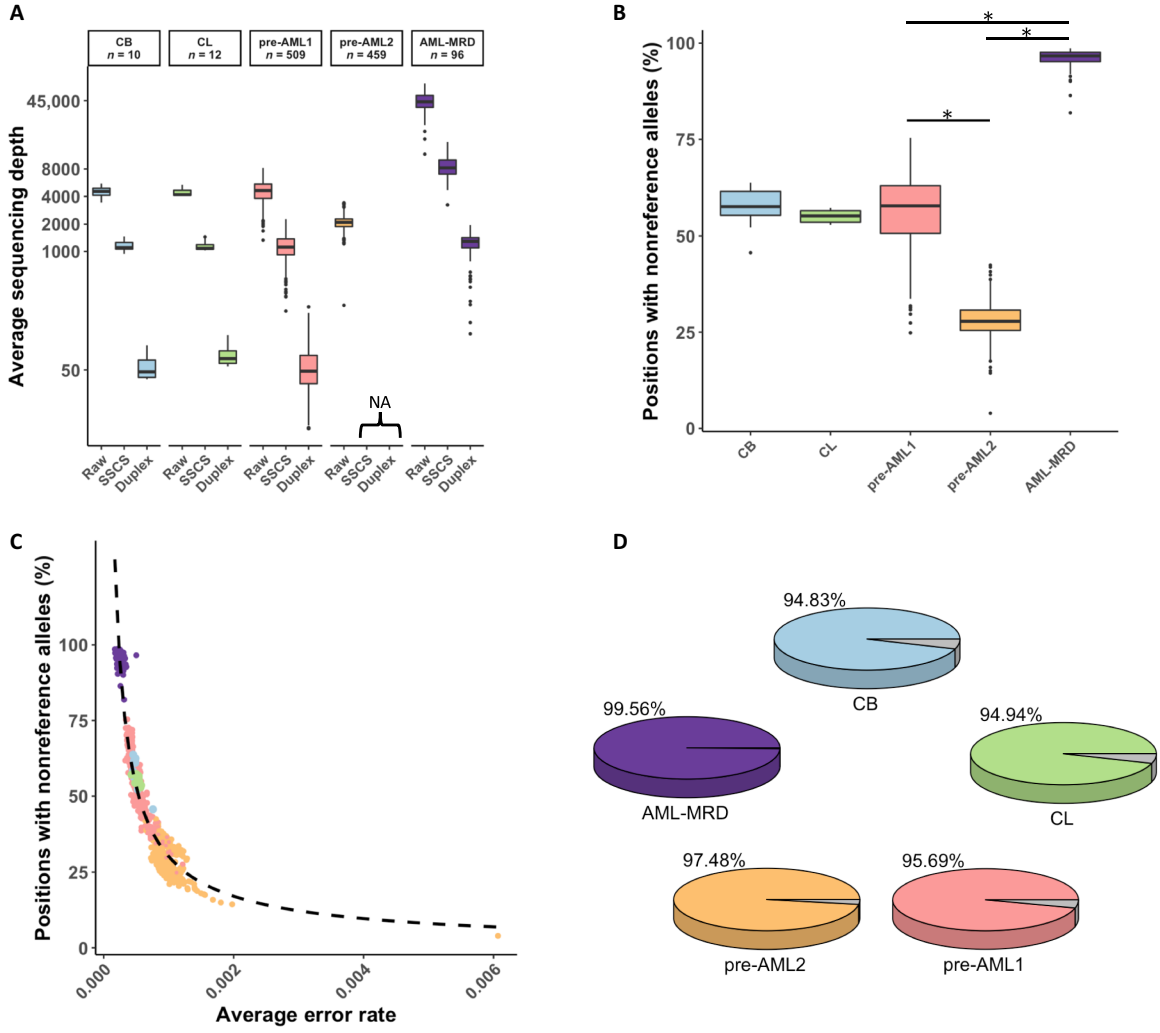

Fig. 1. Sequencing depths and error abundance in the investigated datasets. (**A**) Raw, SSCS, and duplex average sequencing depths across all the samples included in this study. Different colors represent different datasets, and these are consistent across all of the figure panels. (**B**) Sample-wide error abundance in the diverse NGS cohorts. The fraction of genomic positions being observed with at least one nonreference allele supporting read in each sample is indicated. Error burden is significantly different among the investigated datasets (Mann-Whitney test: $P < 1.2 \times 10^{-53}$ for the indicated comparisons). (**C**) Inverse correlation between the abundance of genomic positions with nonreference allele and their corresponding allele frequencies is demonstrated (Spearman's rank order correlation: $r = -0.95$; $^{*}P < 2.2 \times 10^{-16}$). Each dot represents a single sample. (**D**) Panel-wide error abundance in the diverse NGS cohorts as determined by the inclusion of positions with a minimum of one nonreference supporting read in at least one sample. NA, not applicable.

to the substitution type, and its 5′ and 3′ one-base flanking genomic sequence. We found that error rates are highly heterogeneous across the 192 distinct trinucleotide sequence contexts (Fig. 2A, top, and fig. S2) and are highly variable between samples within the same experimental cohort (Fig. 2A, bottom). High error rates were frequently observed at C>A and C>T substitutions (Fig. 2, A and B, and fig. S2). C>T error rates were particularly high when they occurred at a CpG context (Fig. 2, A and C, and fig. S2). Initiated by spontaneous deamination of 5-methylcytosine, real mutations in this context accumulate during aging (*25*), are frequent in germline cells (Supplementary Note), and are also highly prevalent in cancer genomes (*26*), emphasizing the importance of evaluating error rates in relation to their associated genomic contexts.

While contextual error patterns were generally similar between their complementary counterparts, they did not always mirror each other perfectly within any particular sample (fig. S3A). Small yet statistically significant asymmetric error rates were consistently observed among the majority of error contexts in each of the cohorts (Fig. 2D and fig. S3B). For instance, we measured asymmetric error rates involving G>T/C>A, in line with prior observations (*27*). Error rate asymmetries were markedly directional and consistently elevated in specific contexts as compared with their matched reciprocals in all of the investigated datasets. As an example, each of the 16 trinucleotide contexts containing A>T substitutions demonstrated elevated error rates as compared with their corresponding reciprocal contexts containing T>A substitutions. Together, these results indicate that 192, rather than 96, contextual error types would need to be considered to accurately model error rates.

Next, we investigated how sequencing depth may influence error frequencies. As with error rates, sequencing depth differed between distinct contextual error types (Fig. 2E). We noticed a marked inverse correlation between sequencing depth and guanine or cytosine content within specific trinucleotide contexts, a possible reflection of the systemic under-coverage in GC-rich regions reported in NGS (Fig. 2F) (*28*, *29*). Sequencing depth was also lower within trinucleotide contexts that included C>G and G>C substitutions as compared with those that included nucleotide substitutions that reduce GC content (Fig. 2G). These data illustrate how sequencing depth can be influenced by both the trinucleotide context and the nonreference allele.

Overall, a modest, statistically significant inverse correlation was observed between sequencing depth and error rates (Fig. 2H). Correlation strengths were not equal among distinct contextual error types. Further supporting this trend, individual samples with lower average sequencing depth displayed high error rates in multiple contextual error types (see arrows in Fig. 2, A and E). In contrast to the error rates, the absolute number of nonreference supporting reads at the distinct contextual error types showed reduced inter-sample differences in those samples; however, the differences between distinct contextual errors were preserved (Fig. 2I). Collectively, the results obtained here suggest that integration of intra-sample contextual error modeling of nonreference supporting reads at each of the 192 contexts may be a promising strategy for accurate suppression of errors produced by hybrid-capture NGS.

Using intra-sample contextual error modeling for reduction of false-positive calls

As described above, errors varied across samples yet were highly stereotypical according to sequence context and sequencing depth. We reasoned that intra-sample contextual error patterns could be leveraged for in silico error suppression. Such an approach could have several inherent advantages over existing error suppression methods that rely on UMIs, apply thresholds based on intra-sample–wide error rates, or use control samples to train error rate models. Therefore, we devised a computational approach, called Espresso, to model within a sample of interest the nonreference allele counts at each of the 192 distinct contextual error types. Espresso incorporates three distinct features that make it robust to different sequencing datasets (Supplementary Note): (i) pragmatic pre-filters that prepare the dataset for error modeling (fig. S4), (ii) automatic selection of the most appropriate probabilistic distribution for error modeling at a particular contextual error type (fig. S5), and (iii) utilization of nonreference supporting reads as opposed to VAF for error modeling (fig. S6). Unlike applying fixed and arbitrary cutoffs (e.g., minimum VAF, coverage, and number of supporting reads), nonreference alleles would not be indiscriminately eliminated by such an approach; rather, mutations would only be called if they reached statistical significance when compared to their corresponding error distributions (Fig. 3, A to E, and Materials and Methods).

To evaluate the performance of Espresso, we first applied it to the CB dataset. We reasoned that CB would have a minimal burden of somatic mutations, allowing for a more precise estimation of true error rates. We also tested in parallel other common error suppression techniques for unbiased comparative performance assessment (Materials and Methods). The techniques selected for comparison were representative of the spectrum of previously published tools. Specifically, we used two UMI-based methods, namely, single-strand consensus sequences (SSCSs) and duplex sequences (*12*), and two statistical methods for error correction that model background error distributions differently. Among the two statistical methods used, one relies on a training cohort to estimate error rates at the allele level (termed AL here) (*20*), and the other estimates error rates at the sample level (termed SL here) (*14*) without consideration for distinct sequence contexts.

Panel-wide error rates were highly similar among the 10 CB samples but varied significantly among the different error suppression methods (Fig. 3F). As compared with the various statistical approaches (i.e., SL, AL, and Espresso), the UMI-based methods demonstrated inferior error suppression capabilities. A minimum of nine nonreference supporting SSCS reads or three nonreference supporting duplex reads were required to achieve panel-wide error rates comparable to that of SL and Espresso in the CB dataset. We observed similar relative performance among the methods to maximize the number of error-free positions across the entire target panel (Fig. 3G). Considering the highest panel-wide error rate obtained by Espresso (2.74×10^{-6}) and the lowest of the panel-wide error rate observed without error suppression (0.025) across the CB samples, Espresso achieved an error rate reduction of more than 9000-fold.

Analytical assessment of mutation detection accuracy

To evaluate the sensitivity and specificity exchange delivered by Espresso, we analyzed the sequencing data from the CL dataset, which consisted of a dilution series using two cancer cell lines, MOLM13 and SW48. For sensitivity measurements, we assessed the ability of the different methods to detect 119 MOLM13-specific germline variants at the different dilutions (table S3). To evaluate specificity, we assessed the miscalling of 186 AML-related somatic hotspot mutations that are covered by the target panel but are absent from both cell lines (table S3). Espresso outperformed all the other methods in

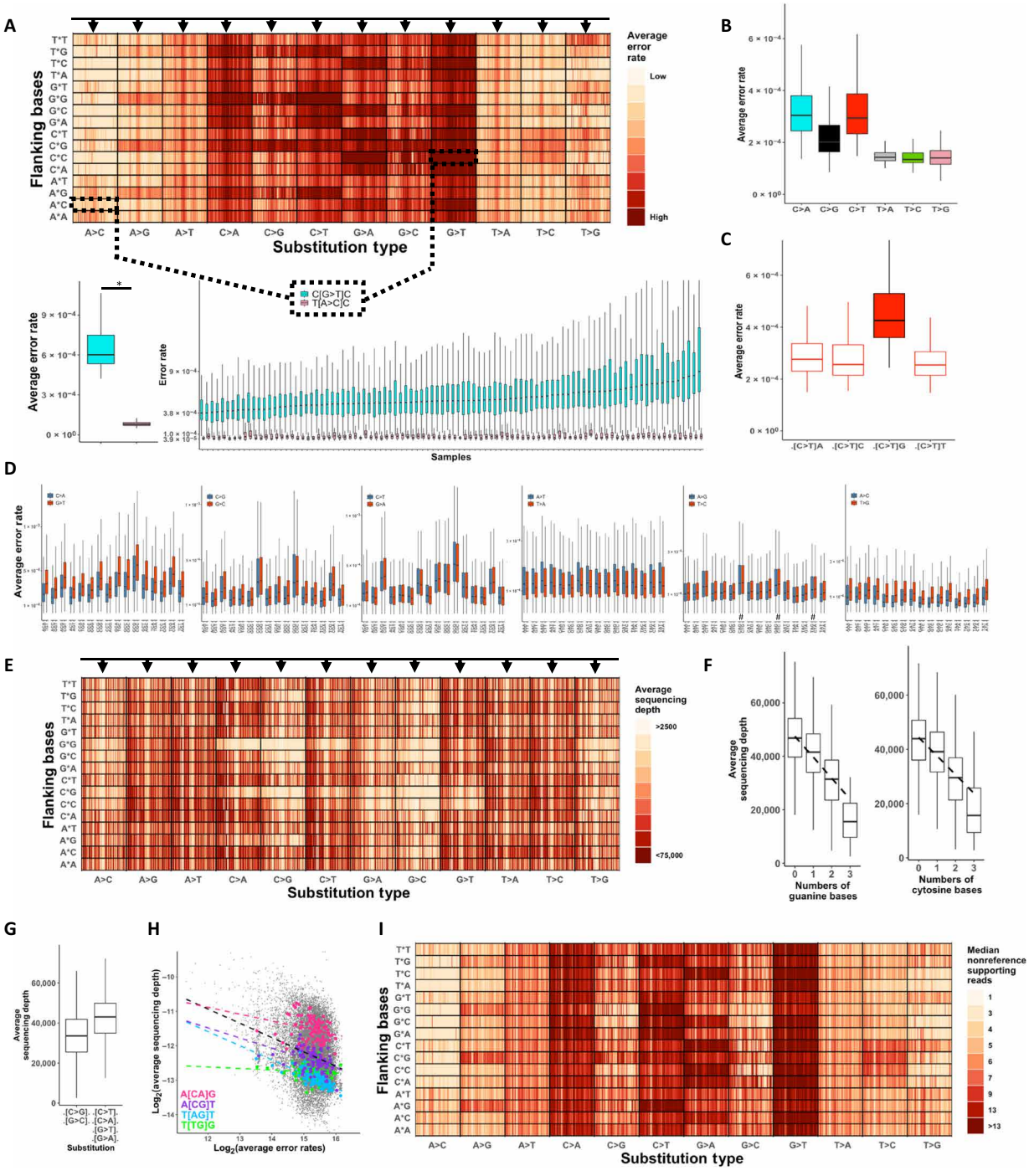

Fig. 2. Error rates significantly differ between trinucleotide sequence contexts. (**A**) Nonreference average error rates at the 192 distinct trinucleotide contexts are shown using the AML-MRD dataset. Vertical lines in each box represent individual samples. Samples' order is kept among distinct contexts. Arrows represent a group of samples with high error rates across multiple contexts. The bottom panels exemplified variation among contextual error rates (*Wilcoxon signed-rank test: $P < 1.8 \times 10^{-17}$) and samples (Mann-Whitney test, samples with the highest and lowest error rates. C[G>T]C: $P < 7.7 \times 10^{-41}$, T[A>C]C: $P < 3.6 \times 10^{-6}$). (**B**). C>T and C>A substitutions are more frequent (Wilcoxon signed-rank test, $P < 1.4 \times 10^{-252}$ for all the comparisons with the other substitution types). (**C**) High error rates at CpG sites (Wilcoxon signed-rank test, $P < 1.1 \times 10^{-64}$ for all comparisons). (**D**) Error rates vary between error contexts and their reciprocals (Wilcoxon rank sum test, $P < 0.05$; #significance was not reached). (**E**) Average sequencing depths. Arrows represent a group of samples with low sequencing depths across multiple contexts. (**F**) Reduced sequencing depth at contexts that include reference cytosine and an increasing number of guanine (Pearson correlation: $r = -0.35$; $P = 2.3 \times 10^{-264}$) and at contexts that include reference guanine with an increasing number of cytosine ($r = -0.29$; $P = 8.6 \times 10^{-179}$). (**G**) Low sequencing depth at contexts with C>G or G>C base substitutions (Wilcoxon signed-rank test: $P = 1.7 \times 10^{-217}$). (**H**) Inverse correlation between depth and error rates (black dashed line, log-log scaled Pearson correlation: $r = -0.27$; $P = 9.7 \times 10^{-308}$). Correlation strengths differ among different error contexts (colored dashed lines). (**I**) The number of nonreference supporting reads at the 192 distinct trinucleotide contexts is shown. The samples' order is identical across (A), (E), and (I).

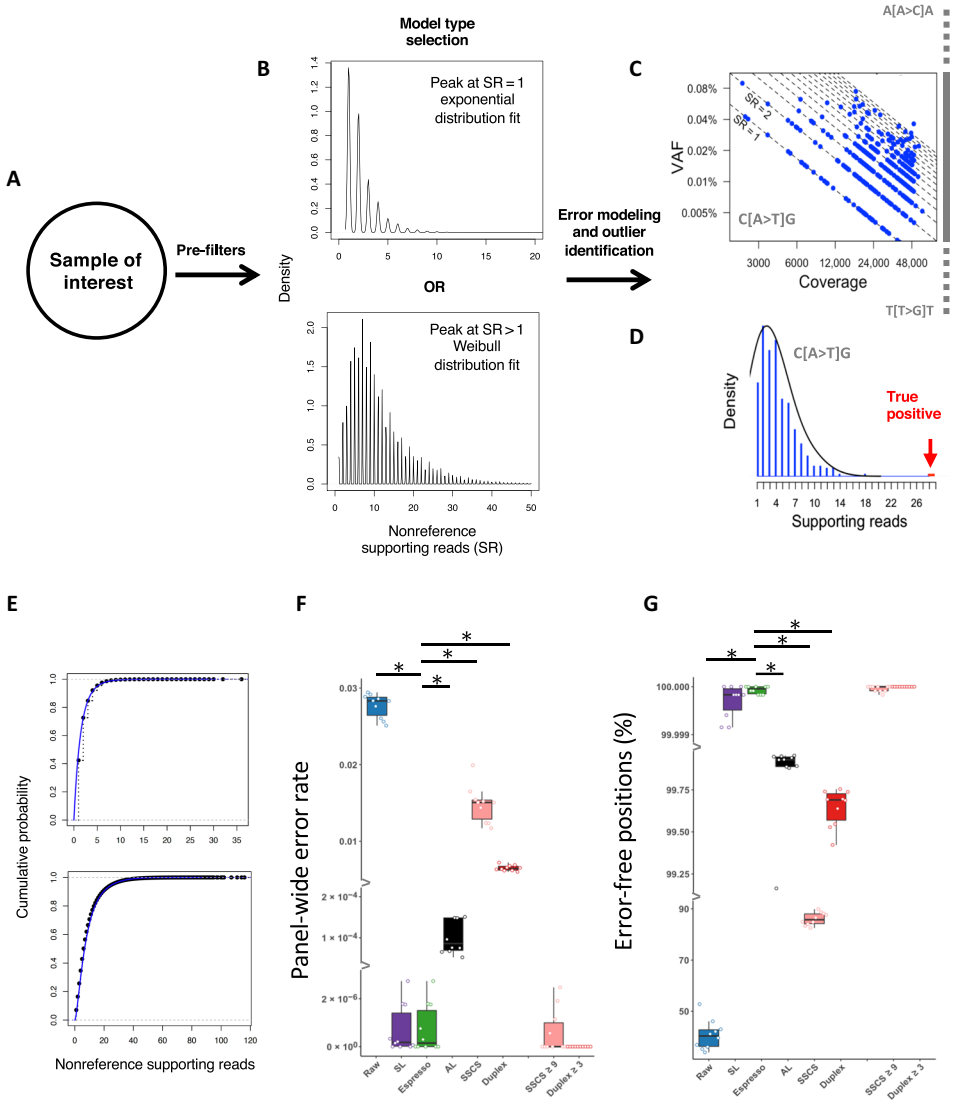

Fig. 3. Integration of intra-sample contextual error modeling for error suppression. Flowchart illustrating the error modeling technique that is implemented by Espresso. (**A**) Following the summarization of the sequencing data to include the dominant alleles at each investigated genomic position, their corresponding read counts, and the average mapping read qualities in each sample of interest, a set of filters is being applied, aiming to deplete potential somatic SNVs and common polymorphism from being included in the error models. (**B**) On the basis of the distribution of the nonreference supporting reads in the enriched error list, Espresso selects between either the exponential or the Weibull probabilistic approaches. (**C**) The nonreference supporting read (SR) counts in each sample are being grouped based on the genomic sequence context to generate 192 context-specific distribution models. (**D**) The models are being reapplied to the entire sample's data for outlier identification. True positives are being determined if they reach statistical significance when compared to their corresponding error distribution. (**E**) The cumulative distribution function graph displays the empirical data (black dots) and the theoretical data (blue line) generated by the 192 models in all the samples included in the CB dataset (top, exponential models) and the AML-MRD dataset (bottom, Weibull models). (**F**) Panel-wide error rates defined as the number of nonreference alleles supporting reads following error suppression, divided by all the reads from the same category (i.e., raw, SSCS, and duplex reads) across the entire 1,264,830-bp panel and (**G**) percentage of error-free positions in the 10 cord blood samples are illustrated. For error suppression, a cutoff P value ≤ 0.05 (Bonferroni-adjusted) was used. SSCS and duplex cutoffs are ≥1 nonreference supporting read unless indicated otherwise. * indicates Wilcoxon signed-rank test: $P < 0.002$.

distinguishing between true and false variants (Fig. 4A). In contrast, duplex sequencing achieved the smallest area under the receiver operator curve (AUC), highlighting the low diagnostic accuracy of this method and, consequently, its limited clinical utility in detecting variants across large hybrid-capture panels.

The use of hybrid-capture NGS panels allows for the detection of mutations at thousands of genomic positions. However, their use also creates unique challenges for true variant identification across so many bases. In addition to high sensitivity and specificity, positive predictive value (PPV) must be prioritized to maximize utility. We assessed PPV in conjunction with sensitivity (i.e., precision-recall analysis). We focused on variants with expected VAF ≤ 0.2%, since accurate variant detection below this threshold is clinically important yet has proven to be a great challenge for existing hybrid-capture NGS platforms (*5*, *30*). Espresso provided a sensitivity of 19.9%, thus achieving the highest number of true-positive, low-VAF alleles at 100% PPV among the tested methods (Fig. 4B). This corresponds to a 6.8-fold improvement as compared to AL, which was the next

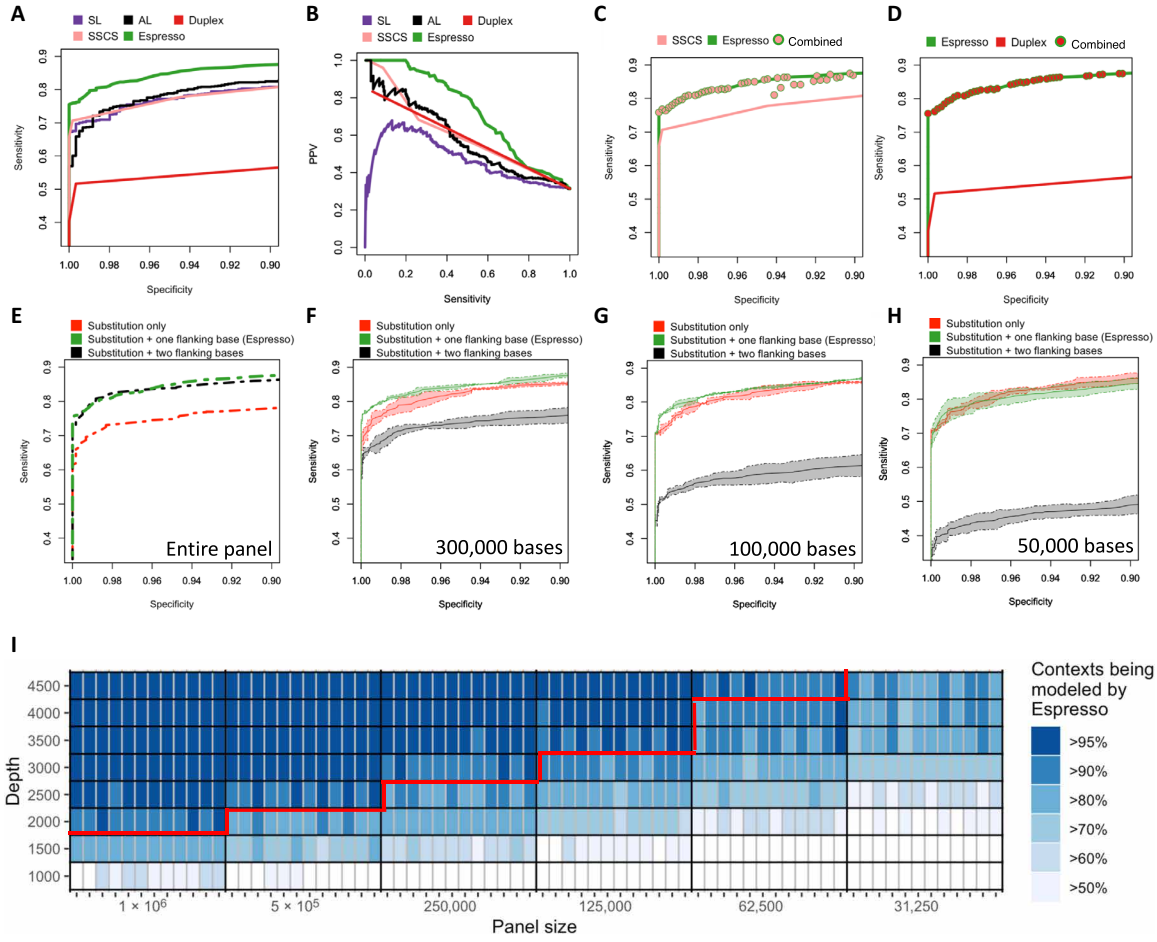

Fig. 4. Statistical measures of performance and constraints of contextual error modeling. (**A**) Espresso demonstrates improved sensitivity versus specificity and (**B**) preferable precision-recall trade-offs as compared with the various indicated methods. The ability of each method to differentiate between 119 positive alleles and 186 negative control variants in a set of serially diluted cell line DNA samples was tested. (**C** and **D**) No substantial benefit of using UMIs to augment Espresso's performance could be determined. Sensitivities and specificities were measured at all the possible combinations of the unique *P* values outputted by Espresso and the unique numbers of SSCS or duplex nonreference supporting reads that were observed in the dataset. The maximum sensitivities at each calculated value of specificity are illustrated. (**E** to **H**) Sensitivity versus specificity trade-offs derived by the reduced and extended contextual error modeling approaches are illustrated in comparison with Espresso. Ninety-five percent confidence intervals (shaded colors) and average values were derived by three random subsets of the data for each one of the indicated in silico decreased panel sizes. (**I**) Heatmap illustrating the percentage of contextual models that can be generated by Espresso when data are being restricted by either panel size reduction or sequencing depth reduction, or both. Data removal was controlled for both the reference and nonreference supporting reads, thus keeping the variant allele frequencies of the nonreference alleles similar to those in the original samples. The red line illustrates such combinations, of which 90% or more of the distinct contextual models could have been generated in every sample in the CL dataset. With datasets that fall below this line, the 12-model contextual error modeling approach can be used in addition to Espresso.

best-performing method to detect low-VAF alleles without scarifying PPV. Notably, SL performed far worse in this analysis than the other methods due to a high number of false-positive calls across various sensitivity thresholds. This result highlights the limited power of noncontextual, sample-level error modeling in detecting mutations with very low read support despite its ability to achieve an extremely high level of error suppression (Fig. 3, F and G). Further supporting this, we compared the false-positive and true-positive calls obtained by Espresso with that of Mutect2 (*16*) at "tumor-only mode." Once more, Espresso demonstrated superior results (table S4).

Previously, the suppression of errors through statistical error modeling was shown to be enhanced by combination with UMI-based approaches (*20*). However, integrating UMI information with Espresso did not confer significant performance improvements (Fig. 4, C and D), suggesting that accurate detection of low-frequency variants can be achieved with Espresso alone. Collectively, the comparative analysis using the CL dataset indicates that the bioinformatic strategy applied here outperformed other methods in the reliable distinction of low-frequency errors from real SNVs.

Impact of panel size and sequencing depth on contextual error modeling

To characterize pragmatic constraints of our method, we compared Espresso with alternative sequence context-based error models. Specifically, we included (i) a simplified 12-model design that accounts only for the 12 possible distinct substitution types without consideration of flanking bases and (ii) an expanded 3072-model design that accounts for the substitution type and for two additional 3′ and 5′ flanking bases. We evaluated the impact of panel size (i.e., number of interrogated bases) and sequencing depth on the performance

of Espresso and the alternative sequence context-based models using the CL dataset.

This comparative analysis exposed critical factors affecting the performance of the alternative models. On the one hand, the performance of the 3072-model approach suffered with reduced panel size (Fig. 4, E to H, and fig. S7A). This is an expected observation that is attributed to the reduction in the number of nonreference alleles being used to populate a relatively high number of models, thus resulting in either model generation failure or an inadequate estimation of the background error noise. In contrast, performance of the 12-model approach was less dependent on panel size since the relatively small number of models was easily populated with nonreference alleles (Fig. 4, E to H, and fig. S7B); however, Espresso consistently outperformed the 12-model approach, presumably because the 12 models were insufficient to account for errors arising within distinct sequence contexts. Moreover, the 12-model approach performed poorly on the largest panel size, possibly as a result of model overfitting from high-VAF errors that escape the initial filtering steps (Materials and Methods). The performance of Espresso was relatively consistent across a broad range of panel sizes from ~1 Mb down to ~50 kb (Fig. 4, E to H, and fig. S7C).

Next, we serially downsampled the CL dataset to simulate various practical scenarios of panel sizes (1 Mb to 32.5 kb) and sequencing depths (4500× to 1000×). At each simulated panel-depth combination, we determined the percentage of trinucleotide contexts that could be modeled directly by Espresso (Fig. 4I). Notably, low represented nonreference alleles that cannot be modeled directly by Espresso would still be analyzed automatically by alternative techniques that are included in the software package (see "Data and materials availability"). Overall, these results illustrate the performance dependencies of Espresso and related sequence context–based models to assist with their implementation in a wide range of sequencing settings.

Detection and monitoring of minimal residual disease

Having demonstrated Espresso's high analytical performance in the CB and CL datasets, we next sought to evaluate its clinical utility. The presence of persistent AML clones that carry genetic abnormalities during or after treatment has been shown to carry crucial prognostic information (*31*). Therefore, we assembled a cohort of 42 patients with AML (AML-MRD; table S5) whose mutations were previously determined at diagnosis (table S3). Forty of the 42 patients had serial samples analyzed by ultra-deep hybrid-capture NGS at two time points during therapy; for the other two patients, single follow-up samples were available.

Since minimal/measurable residual disease (MRD) monitoring may guide clinical decisions (*32*–*34*), in addition to true positives, both false positives and false negatives could have tremendous implications for patient care. We therefore evaluated F1 scores, which represent the harmonic mean of PPV and sensitivity. For comparative performance evaluation, mutations reported at diagnosis were considered as true positives if they were detected in the follow-up samples of the same patient or as false positives if they were detected in other patients. We first applied a cutoff of $\alpha \leq 0.05$ (Bonferroni-adjusted) for the probabilistic methods SL, AL, and Espresso and a heuristic threshold of ≥ 1 nonreference supporting reads for the UMI-based methods SSCS and duplex. Tested on the subset of samples obtained at either the first time point (T_1, closer to diagnosis) or the second time point (T_2, further into treatment), Espresso delivered the highest F1 scores (0.71 at T_1 and 0.74 at T_2) followed by AL and duplex (Fig. 5A). We next applied the optimized SSCS and duplex cutoffs used in the CB analysis (i.e., ≥ 9 and ≥ 3 nonreference supporting reads, respectively). Although F1 scores improved with these parameters, they still fell short due to an increased number of false positives for SSCS ≥ 9 and an increased number of false negatives for duplex ≥ 3 in both the T_1 and the T_2 data subsets as compared with Espresso (Fig. 5B).

Despite the technical differences between the CL and AML-MRD datasets, Espresso once again produced the most preferred balance between sensitivity and specificity (Fig. 5C). We compared Espresso with additional algorithms and saw consistent outcomes. Espresso outperformed Mutect2 (*16*) in both the tumor-only mode and the "panel of normals" mode when samples obtained from 14 healthy adults were used (table S4). Espresso also outperformed deepSNV (*18*), a statistical algorithm that was developed specifically for the accurate detection of SNVs from deep targeted sequencing experiments. The comparison with deepSNV extrapolates beyond the probabilistic approaches being used and illustrates the benefits of other features implemented in our bioinformatic pipeline for the reduction of false-positive calls (fig. S8).

Having established Espresso as the preferred methodology to maximize the accuracy of SNV detection from peripheral blood, we next sought to implement it for the characterization of clonal dynamics in patients with AML. Since the competitive balance among different hematopoietic clones is likely to change during multiple rounds of chemotherapy, we hypothesized that Espresso would enable the identification of resistant clones that were not reported at diagnosis. We therefore extended our analysis to include an additional 147 highly recurrent AML SNVs that are covered by the AML-MRD hybrid-capture panel (table S3). Across all the samples, Espresso identified 92 mutations ($\alpha \leq 0.05$, Bonferroni-adjusted) with the lowest being reported at VAF = 0.0135% (table S6 and fig. S9). These correspond to 59 distinct mutations, out of which 47 (~80%) were present in at least two samples of the same patient (that is, reported at diagnosis and detected in at least one additional time point by Espresso or detected in the two follow-up samples by Espresso). Such a high percentage of validated mutations is an indicator of Espresso's reliable mutation calling. Among these, Espresso has enabled the detection of 22 new putative driver SNVs not reported at diagnosis in 15 patients, including in 3 of the 7 patients (~43%) with no SNVs in the diagnostic report (table S6). Further supporting the validity of the mutations called by Espresso, most of these newly identified mutations were in genes that commonly contribute to positive clonal selection following cytotoxic chemotherapy (*35*–*37*), including *TP53* and *DNMT3A* (Fig. 5D).

Together, our results demonstrate substantial advantages of Espresso over other methods for SNV detection from peripheral blood of patients with AML during the course of therapy. Encouraged by a recent consensus document release from the European LeukemiaNet MRD Working Party (*38*), many studies are now underway to evaluate the prognostic and predictive significance of clonal dynamics in AML and the proposed role of MRD detection as a surrogate endpoint for clinical trials (*39*). Implementation of Espresso in these contexts has the potential for significant clinical utility.

Targeting informative genomic loci for improved practicality of pre-AML genomic screens

Age-related clonal hematopoiesis (ARCH) is a common phenomenon evident by the presence of somatic mutations in hematopoietic

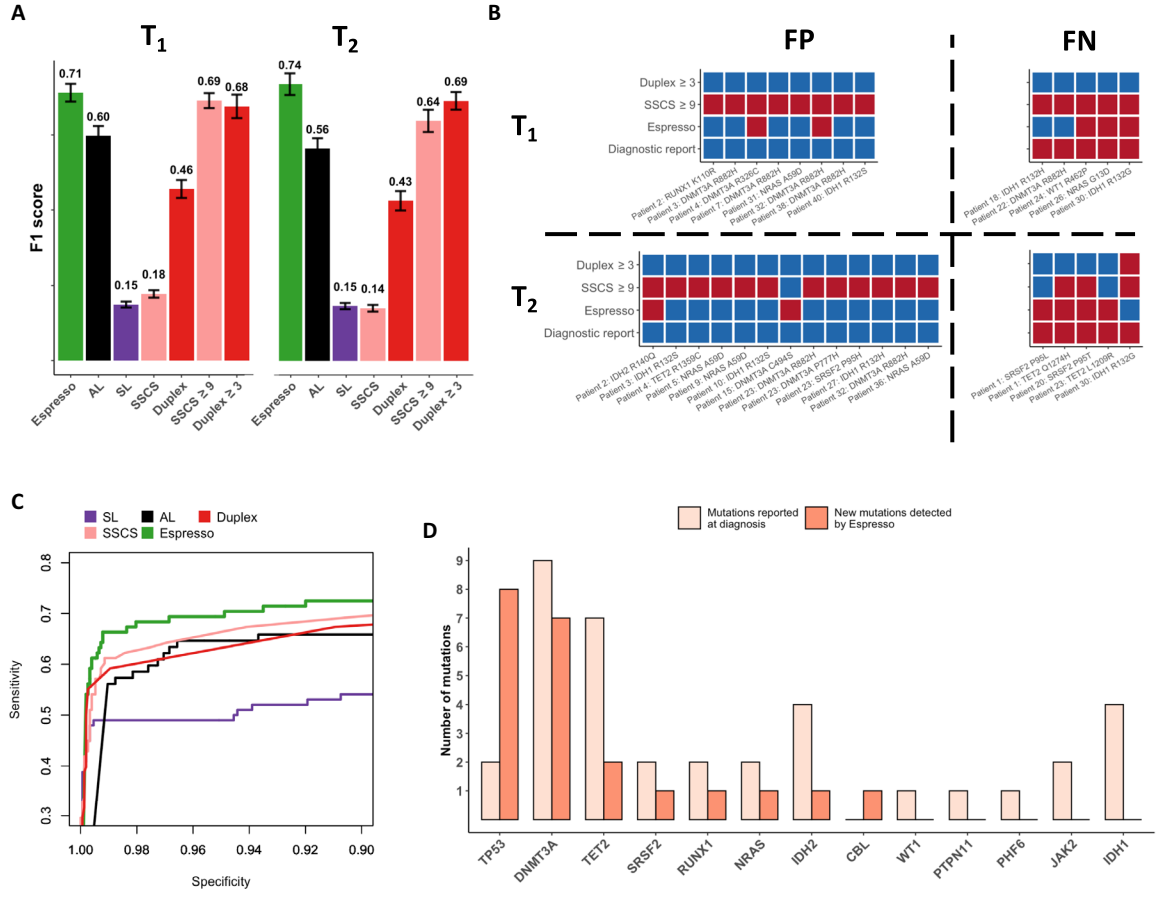

Fig. 5. Measurements of residual disease. (**A**) Espresso provides a preferred balance between precision (PPV) and recall (sensitivity), as determined by the inspection of 78 SNVs reported across 35 of 42 patients at the time of AML diagnosis. Mutations were called in the patients' sample at 21 different iterations. In each iteration, 6 random patients of the 42 were excluded. Median F1 scores and 1 SD are shown for the various methods tested at two time points during the course of treatment (T_1 and T_2, Wilcoxon signed-rank test: $P \leq 6.4 \times 10^{-5}$ for all the comparisons with Espresso). (**B**) The variation in the mutations being called by Espresso ($\alpha \leq 0.05$, Bonferroni-adjusted), SSCS (≥9 nonreference supporting reads), and duplex (≥3 nonreference supporting reads) is illustrated. Red color indicates called mutations, while blue color indicates that mutations were not detected. FP, false positives; FN, false negatives. (**C**) Sensitivity versus specificity as determined by the different tested methods. (**D**) Enrichment of clones, carriers of TP53, and DNMT3A mutations is observed in patients with AML following therapy. The *y* axis represents the number of mutations detected, classified by the affected genes.

stem cells of otherwise healthy individuals that cause a clonal expansion of the stem cells and their progeny (*40*). Recently, our group reported several hundred ARCH-associated mutations spread across 27 genes with various contributions to the risk of AML transformation (*9*). Our study provided a proof of concept for risk prediction of AML. Nevertheless, large population screens using broad sequencing panels remain socioeconomically unattractive because of high costs, the relatively low incidence of AML, and the relatively high incidence of ARCH in the general population.

To address these challenges, we reasoned that interrogating a small number of highly recurrent AML mutations would be a more tractable approach than broad hybrid-capture sequencing. This approach could theoretically result in improved segregation between pre-AML and controls while reducing sequencing costs. The success of this approach relies on the accurate identification of preleukemic mutations in asymptomatic individuals.

We first compiled datasets that would allow comparisons among the distinct methods used in our previous analyses. For this reason, we focused initially on the pre-AML1 dataset, which contains UMIs in the sequencing reads, and the CB dataset, which could be used as a training set for error rate estimation at the AL. Putative driver SNVs (that is, mutations in coding sequences other than synonymous SNVs and mutations at splice sites) identified by each method at the recurrently mutated genomic loci were used to derive random forest classifiers that were trained and tested on their corresponding method's mutation calls (table S7). For the probabilistic methods, $\alpha \leq 0.05$ (Bonferroni-adjusted) was used, and for the UMI-dependent methods, we applied either a threshold of one supporting consensus read or SSCS ≥ 9 and duplex ≥ 3. The Espresso-derived classifier exhibited the highest level of performance for discriminating pre-AML from controls (AUC: 0.74) and reported the highest sensitivity (46.8%) at 100% specificity (Fig. 6A). A reduction in specificity down to 96.3 or 93.7% was needed to achieve the same sensitivity with the SL-derived and SSCS-derived classifiers, respectively. The SSCS-derived model also underperformed the Espresso-derived classifier when the SSCS ≥ 1 cutoff was applied (AUC: 0.66, Fig. 6A, dashed line). The duplex ≥ 3 derived classifier had the poorest performance (AUC: 0.42), owing to poor duplex consensus efficiency (fig. S1B), low duplex coverage (Fig. 1A), and subsequent dropout of mutations not meeting the required cutoff. On the contrary, with a threshold of

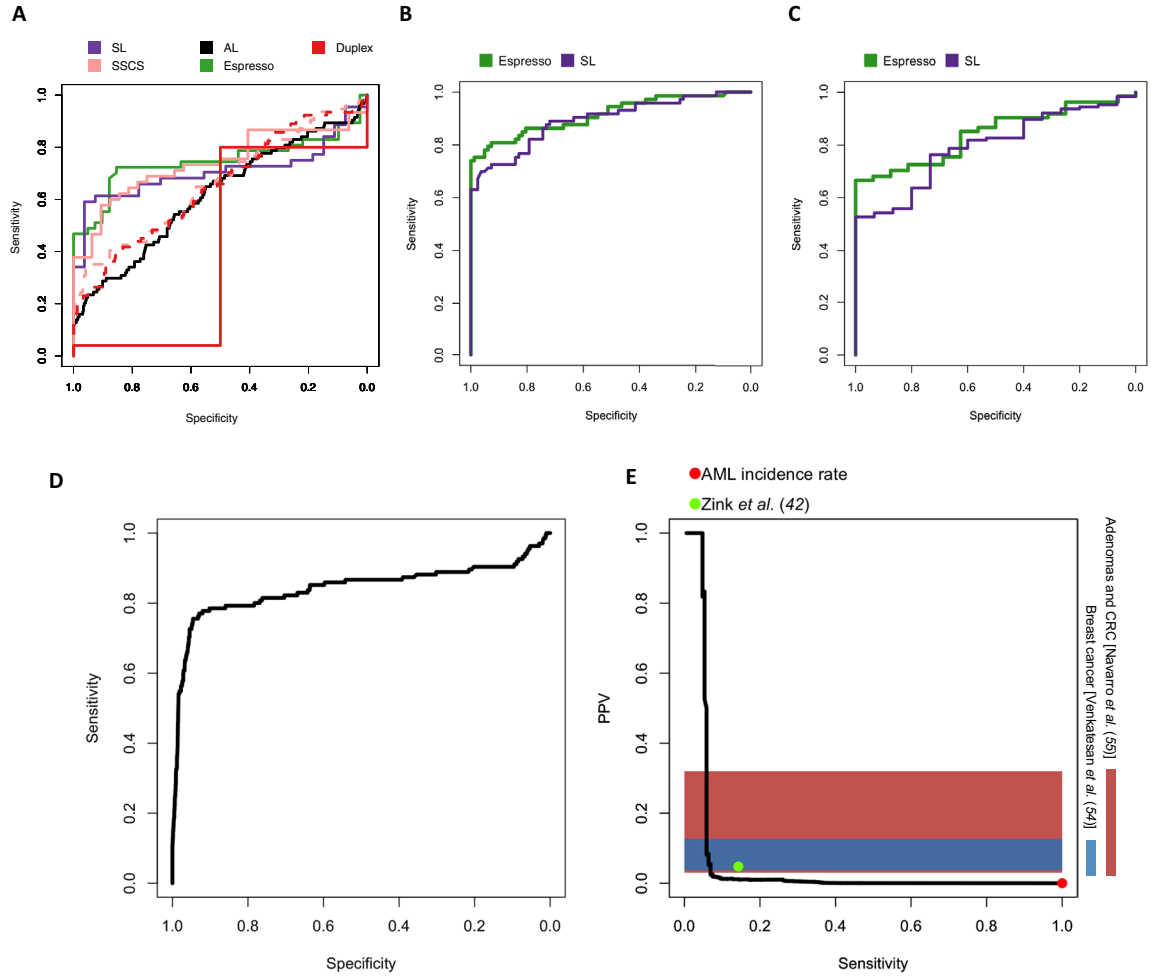

Fig. 6. AML transformation risk prediction using recurrently mutated loci. Classification performance evaluation of pre-AML and control, mutated samples. (**A**) Each classifier was trained and tested on the mutations that were obtained from the classifier's corresponding method. (**B**) Comparison between the Espresso and the SL-derived classifiers. In this iteration, each classifier was trained using its corresponding method's mutation calls and was tested in its accuracy to classify pre-AML cases and controls, including mutated samples identified by the other method as well. (**C**) Comparative performance validation between the Espresso and the SL-derived classifiers to differentiate between pre-AML and control samples obtained from an additional validation dataset (*8*). Information regarding the study participant's age, specific mutations, and their VAFs was obtained directly from the main text. (**D**) Performance estimation using the validation dataset and simulated controls. (**E**) Precision-recall trade-offs are calculated at the individual level (that is, serial samples are accounted for single individuals and individuals without any mutations are also included in the performance measurements). The red dot indicates AML's incidence rate. This is equivalent to a situation where no screen is being conducted at all [PPV = incidence rate = 0.006% (*44*), SN = 100%]. The green dot indicates the model performance using an additional published dataset consisting of 11,262 individuals when the model was set to achieve 100% specificity in the training set. Horizontal color bars represent PPV ranges determined for screening mammography for breast cancer (*54*) and fecal immunochemical test for advanced adenomas and colorectal cancer (CRC) (*55*). Comparison with the genetic risk model performance shows the extent to which sensitivity must be compromised to achieve PPV comparable with these widely applied early detection tests.

one supporting duplex read, a large number of putatively false-positive SNVs were called, resulting in poor classification accuracy (AUC: 0.65, Fig. 6A, dashed line). The AL-derived classifier also performed poorly due to a high number of false-positive SNVs (AUC: 0.62).

There is a low cumulative risk of ARCH progression to hematologic neoplasms (*41*). For this reason, the implementation of a population-based pre-AML genomic screening test would need to achieve exceedingly high specificity and low false-positive rate. We therefore prioritized the Espresso- and SL-derived classifiers for subsequent performance evaluation. Additional mutations that were found by Espresso and SL in the pre-AML2 dataset were included in the analysis (table S7). Each classifier was trained on the mutations found by its corresponding method in both the datasets (pre-AML1 and pre-AML2) and tested on the data that include all the mutations detected by either of the two methods. The Espresso-derived classifier once more provided a better overall sensitivity-specificity balance and a greater sensitivity at 100% specificity (Fig. 6B). Similar trends were observed when both the classifiers were applied to an external validation set consisting of mutations called in 188 pre-AMLs and 181 controls (*8*), with the Espresso-derived classifier again displaying higher discriminatory accuracy (Fig. 6C). Together, the superior classifier performance using mutations called by Espresso illustrates that accurate mutation calling is imperative when designing genetic risk prediction models.

To estimate how well the winning classifier would perform as a population-wide screening test, we spiked the validation set into

>4 million in silico simulated controls (prevalence ~0.005%; Materials and Methods). Despite the small genomic footprint (table S8), the Espresso-derived classifier resulted in accurate identification of the mutated pre-AML samples (AUC: 0.84; Fig. 6D). As an example, when the model was tuned to minimize false-positive calls based on the pre-AML1/pre-AML2 merged training dataset, a sensitivity of 29.3% and a specificity of 99.8% were obtained. Precision-recall analysis revealed the extent to which the Espresso-derived classifier may enrich for individuals at high risk of developing AML as compared with current practice (no screening, i.e., AML incidence rate) (Fig. 6E). Sensitivity was 4.8% at 100% PPV; this small subset detected with no false positives was enriched for highly penetrant *SRSF2*/*IDH2* double-positive individuals with the highest risk for AML development (table S9). Last, we estimated the model performance in an additional published cohort of 11,262 individuals (*42*). In this cohort, when the model was tuned to minimize false positives within the training dataset, a sensitivity of 14.3% and a PPV of 4.8% were obtained (Fig. 6E and table S9).

DISCUSSION

In this study, we described the rationale, technical performance characteristics, and potential clinical utility for Espresso, a novel method to improve hybrid-capture sequencing–based SNV detection. Unlike many other NGS error suppression methods, including the representative published UMI-based and probabilistic model–based approaches tested here, Espresso does not rely on UMIs or a training set of controls for error rate estimations; therefore, Espresso improves practicality by reducing library preparation complexity, assay costs, and analysis time. We observed additional notable advantages of Espresso over alternative methods, and these were consistent across diverse datasets. Specifically, Espresso produced superior error suppression and an improved trade-off between sensitivity and specificity for detection of low-VAF alleles.

These advantages of Espresso were the result of several key features. First, Espresso applies a set of pre-filters to prepare the data for error modeling. Second, Espresso automatically selects between two statistical models to estimate the number of alternative supporting reads rather than the VAFs; thus, in addition to selecting the more appropriate error distribution model, it better accounts for error rate bias resulting from variation in sequencing depth within hybrid-capture NGS datasets. Third, Espresso markedly reduces false-positive calls by considering only the dominant nonreference allele at each interrogated genomic position. Fourth, Espresso leverages a large number of errors that share the same trinucleotide sequence context within the investigated sample; thus, it reduces the potential for misrepresentation of real error rates by relatively small control cohorts.

To explore its potential use in clinical settings, we tested the performance of Espresso to detect SNVs in serial peripheral blood samples from 42 patients with AML who achieved clinical remission. Consistent with the performance in the other investigated datasets, Espresso outperformed all the other tested methods in this setting. Using Espresso, we found resistant subclones enriched for *TP53* and *DNMT3A* mutations that were genetically distinct from the AML clones present at diagnosis. In the future, more extensive cohort studies are needed to determine whether the selection and enrichment of such clones following induction therapy may affect patient outcomes in a non-autonomous fashion, similar to the observations in solid malignancies (*43*). Furthermore, combining accurate detection of persistent mutations together with other independent prognostic markers will be necessary to build clinically relevant models for accurate determination of the risk of relapse.

Our results emphasize the importance of accurate mutation detection for the derivation of classification models in the setting of early detection of AML. Using Espresso, we derived a risk prediction model that is focused on a minimal yet highly informative set of genomic loci that are recurrently mutated in patients with AML. With only 1594 genomic bases being interrogated, our results imply that up to 29.3% of de novo AML cases can be predicted years in advance with a specificity of 99.8%. Although sensitivity may greatly suffer with elevated PPV, considering the incidence rates of AML in the general population (~6:100,000) (*44*), our approach would still provide meaningful patient enrichment. Modest sensitivity may be acceptable when screening the general population as long as specificity and PPV remain high. Further prospective validation studies are required to assess the feasibility, utility, and cost-effectiveness of this targeted approach. Our findings should also be extended to incorporate additional predictive biomarkers. As AML is a blood-borne disease, we envision that epigenetic and metabolomic perturbations within leukocytes may further improve prediction accuracy, thus making AML predictions more clinically useful. Our results indicate that certain biomarker-enriched populations may be at an exceedingly high risk of developing AML. In time, novel therapeutic developments and targeted therapies against blood cells with high-risk mutations may provide the minimal side effects necessary to deliver a favorable risk-benefit ratio that justifies the initiation of early intervention clinical studies.

In summary, we have described, benchmarked, and validated a new practical NGS error suppression technique. We have demonstrated the superiority of Espresso in detecting somatic SNVs as compared with existing state-of-the-art approaches and defined its limitations with respect to sequencing depths and hybrid-capture panel sizes. We used Espresso to derive new biological insights, augmenting our understanding of the genetic mutations that define high-risk malignant transformation and therapy resistance clones in patients with AML. We envision that Espresso will prove useful in guiding clinical decisions and scientific research alike.

MATERIALS AND METHODS

Cohorts description

CB dataset: This dataset is composed of 10 human umbilical cord blood genomic DNA samples obtained from Trillium Hospital (Mississauga, Ontario, Canada) with informed consent in accordance with guidelines approved by the University Health Network Research Ethics Board. Cord blood was processed 24 to 48 hours after delivery. Mononuclear cells were enriched using Ficoll-Paque followed by red blood lysis by ammonium chloride and $CD34^+$ selection before DNA extraction. CL dataset: MOLM13 cell line DNA was mixed with SW48 cell line DNA at relative concentrations of 100, 5, 1, 0.2, 0.04, and 0% and was sequenced in duplicate. Pre-AML1 and pre-AML2 datasets: Detailed information regarding these cohorts is described elsewhere (*9*). Briefly, the pre-AML1 dataset contains peripheral blood genomic DNA samples obtained from a total of 509 individuals upon enrollment into the European Prospective Investigation into Cancer and Nutrition (EPIC) study (*45*) between 1993 and 1998. Together, 414 control individuals who did not develop any hematological disorders during the extended follow-up period and 95 individuals who developed AML were included in this study. The pre-AML2 dataset contains peripheral blood genomic DNA samples obtained from individuals enrolled in

the EPIC-Norfolk longitudinal cohort study between 1994 and 2010. Samples were available from 37 patients with AML and 262 age- and sex-matched controls without a history of cancer or any hematological conditions. Samples taken at multiple time points were available for a fraction of the participants in this cohort. Notably, samples from eight pre-AML patients in the pre-AML2 cohort were separately sequenced in the pre-AML1 dataset (by independent investigators using a different methodology). To avoid statistical misrepresentation of AML predictions, we removed those samples from the pre-AML2 dataset before the derivation of the described genetic risk models. AML-MRD dataset: This dataset is composed of peripheral blood genomic DNA from 42 patients with AML treated at the Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. All 42 patients achieved morphologic leukemia-free state (MLFS) on chemotherapy. Complete count recovery occurred when absolute neutrophil count recovered to $\geq 1 \times 10^9$/liter and platelet count recovered to $\geq 100 \times 10^9$/liter up to 7 days following the bone marrow assessment that confirmed MLFS status. All patients were deidentified with patient IDs. Their demographic and clinical features were captured (table S5). All the samples in this study, including healthy individuals and patients with cancer, were collected with informed consent for research use and were approved by Institutional Review Boards in accordance with the Declaration of Helsinki. Protocols were approved by the following ethics committees: (i) International Agency for Research on Cancer Ethics Committee approval #14-31, (ii) East of England—Cambridgeshire and Hertfordshire Research Ethics Committee reference number 98CN01, and (iii) University Health Network Research Ethics Board # 01-0573.24.

NGS libraries construction and sequencing

Library construction and sequencing were done as previously described (*9*). Briefly, for each sample in the CB, CL, and pre-AML1 datasets, 100 ng of genomic DNA was sheared to 250–base pair (bp) fragments before library construction (KAPA HyperPrep Kit KK8504, Kapa Biosystems) with a Covaris E220 instrument using the recommended settings. After end repair and A-tailing, ligation of UMI-containing adaptors was performed with 100-fold molar excess. Agencourt AMPure XP beads (Beckman Coulter) were used for library cleanup following eight cycles of fragment amplification with 0.5 μM Illumina universal and indexing primers. Targeted hybrid-capture was carried out on pools of three indexed libraries. Five microliters of Cot-I DNA (1 mg ml^{-1}; Invitrogen) and 1 nmol each of xGen Universal Blocking Oligo, TS-p5, and xGen Universal Blocking Oligo, TS-p7 (8 nucleotides) were added to each pool of adaptor-ligated DNA. The mixture was dried using a SpeedVac and then was resuspended in 1.1 μl of water, 3.4 μl of NimbleGen hybridization component A, and 8.5 μl of NimbleGen 2× hybridization buffer. The mixture was heat-denatured at 95°C for 10 min following the addition of 4 μl of xGen Lockdown Probes (3 pmol; xGen AML Cancer Panel v.1.0). Hybridization was conducted at 47°C for 72 hours. Washing and recovery of the captured DNA were initiated with 100 μl of clean streptavidin beads that were added to each capture. Following separation of the libraries and the supernatant using a magnet, 200 μl of 1× Stringent Wash Buffer was added, and the reaction was incubated for 5 min at 65°C. The supernatant containing unbound DNA was removed before repeating the high stringency wash for the second time. The bound DNA was then washed one time with 200 μl of each of the following: 1× Wash Buffer, 1× Wash Buffer II, and 1× Wash Buffer III. The washed DNA on beads was resuspended in 40 μl of nuclease-free water, and this volume was divided into two polymerase chain reaction (PCR) tubes that were subjected to 10 cycles of post-capture amplification (Kapa Biosystems, recommended conditions). Libraries were spiked with 2% PhiX before sequencing. The procedure used for the pre-AML2 dataset is described elsewhere (referred to as the validation cohort) (*9*). For each sample in the AML-MRD dataset, peripheral blood samples were collected during remission in PAXgene Blood DNA Tubes (PreAnalytiX, Hombrechtikon, Switzerland). DNA was extracted according to the manufacturer's instructions. Illumina-compatible libraries were constructed from 100 ng of sheared genomic DNA using the Covaris M220 sonicator (Covaris, Woburn, MA, USA) and the KAPA HyperPrep Kit (#KK8504, Kapa Biosystems, Wilmington, MA, USA). Following end repair and A-tailing, adapter ligation was performed for 16 hours at 4°C using 100-fold molar excess of adapters. Agencourt AMPure XP beads (Beckman Coulter) were used for library cleanup, and ligated fragments were amplified by PCR for 6 cycles using 0.5 μM universal and indexed primers. Following hybrid-capture at 47°C for 72 hours, the captured DNA fragments were enriched with 12 cycles of PCR. Paired-end 2 × 125-bp sequencing was performed on an Illumina HiSeq 2500 instrument with eight libraries multiplexed into each lane.

Bioinformatics pipeline and consensus reads assembly

Paired-end sequencing data from the Illumina platform were converted to FASTQ format. When included, the unique molecular barcode information at each read of the pair was trimmed and was added to the read header. The Burrows-Wheeler aligner (BWA-mem) (*46*) was used for the alignment of the processed FASTQ files to the reference hg19 genome. To eliminate the chance of ambiguous short indel alignment on neighboring SNV miscalls, we removed reads with indels. We further cleaned the data from short and hard clipped reads and any nonunique read alignments. We found that, together, these preprocessing steps can improve SNV detection (fig. S8). Consensus read assembly into read families was done in a similar way to previous reports (*47*, *48*). Specifically, reads that share the same molecular barcode sequence, the genomic position of where each read of the pair maps to the reference, and the CIGAR string were grouped. Families that consisted of at least two reads were used to generate SSCS, and a consensus base was called when there was full agreement. When a consensus base was called, it was assigned with the maximum base quality score observed in its corresponding pre-collapsed reads. Similarly, when two SSCSs with corresponding UMIs on the reciprocal strand were observed, duplex reads were generated. After converting the raw-, SSCS-, and duplex-containing sam files into coordinate-sorted bam files, we used samtools (*49*) version 1.2 and Varscan2 (*14*) version 2.2.8 to summarize the data. The following parameters were used: (i) mpileup parameters: -s -x -BQ0 -q1 -d100000 and (ii) pileup2cns parameters: --min-coverage 10 --min-reads2 1 --min-avg-qual 30 --min-var-freq 0.0001 --p-value 1 --strand-filter 0. These are rather permissive parameters allowing the output of all the dominant alleles in each one of the investigated genomic positions. To allow unbiased performance comparisons, we used this format as an input for all the probabilistic methods (SL, AL, and Espresso) and the UMI-based methods (SSCS and duplex).

Probabilistic models for error correction

With Espresso, we deployed a novel approach to model errors based on their association with either one of the 192 contextual contexts (Fig. 3, A to E). These correspond to 12 base substitution types, four

alternative 5′ bases, and four alternative 3′ bases. To mitigate the impact of outliers and real mutations on overfitting, a set of filters is applied to exclude specific variants from the contextual error models (Supplementary Note and fig. S4). These include the removal of alleles (i) that are observed as germline variants in the general population (*50*, *51*) with minor allele frequency ≥ 0.1%, (ii) with VAF/error rates ≥5%, (iii) that have MapQual<59 and MapQual!=0 [for additional information, please refer to the manual of Varscan2 (*14*)], (iv) that describe recurrent cancer mutations, and (v) that disproportionally persist across multiple samples in the dataset (see the "Flagged alleles" section; Materials and Methods). Notably, to prevent performance comparison bias, we used these filters together with all the probabilistic methods (SL, AL, and Espresso) and the UMI-based methods (SSCS and duplex) tested.

To determine the more appropriate distribution type for error modeling, Espresso first investigates the overall distribution of nonreference supporting reads in a context-independent manner, in the sample's filtered, error-enriched list. On the basis of the observed peak occurrence, either exponential or Weibull distribution models are selected to generate all the contextual models. If the peak corresponds to a single nonreference supporting read, exponential distribution will be used to represent the data; otherwise, if this value is larger than 1, Weibull distribution will be used. Either the "pexp" or "pweibull" R functions are then being used together with the modeled parameters from the fitdistrplus package (either rate or shape and scale) to determine how high any nonreference allele of interest is being represented above its corresponding contextual background. A Bonferroni-corrected P value ≤ 0.05 was used to determine whether any nonreference allele received significantly more supported reads.

For comparative performance analysis, error rate models at the AL were constructed as previously described (*20*). Briefly, if the total number of nonzero allele frequencies seen in the training set used for error modeling was ≤5, we used Gaussian distribution; otherwise, we fit a Weibull distribution to the allele frequencies observed in the training set. Specifically, the "pnorm" or pweibull R functions were used together with the modeled parameters (either mean and SD or shape and scale) to estimate the likelihood that any allele frequency value of interest is above the corresponding modeled distribution derived for the same interrogated position in the corresponding training set. The yielded P values were adjusted by incorporating the fraction of nonzero allele frequencies into the final models [for additional information, please refer to iDES (*20*)]. Training datasets were constructed as follows: (i) The pre-AML1 dataset was used for the CB analysis (Fig. 3) and the CL analysis (Fig. 4). (ii) A training set composed of peripheral blood genomic DNA samples from 14 healthy individuals was sequenced and used in the analysis of the AML-MRD data (Fig. 5). (iii) The CB dataset was used as a training set for the derivation of the AL-based model for AML risk prediction (Fig. 6). To evaluate allele mutated status at the SL, we used Varscan2 (*14*) that computes statistical significance in single samples by Fisher's exact test.

Flagged alleles

While parameters such as specific genomic context, the presence of a repetitive region, and low base or read mapping quality may explain the basis of some errors, these do not always capture artifacts that may persist across multiple samples. We therefore derived a statistical approach to flag recurrently specious alleles. To flag potentially low-frequency artifactual alleles that escaped conventional filtering, we iterated between the 99 and 99.9% nonreference allele frequency quantiles in the entire investigated cohort in increments of 0.1% (user-defined parameters). The 10 derived VAF values were used consecutively to apply Fisher's exact tests, determining whether errors with VAF above the quantile-derived cutoff distribute proportionately among all the observed nonreference alleles in the dataset or being clustered in a low number of alleles across many samples in an unbalanced fashion. Then, if included, we removed recurrent Catalogue of Somatic Mutations in Cancer (COSMIC) (*52*) mutations (that is, SNVs with classification other than synonymous with at least three case reports of hematopoietic and lymphoid tissues; COSMIC version 80) to derive a final list of dataset-specific flagged alleles to be excluded from contextual error modeling.

AML risk prediction

To derive with a list of mutations that are highly associated with leukemic transformation for AML risk prediction model derivation, we interrogated the COSMIC database (*52*) and ranked variants according to their evidence for functional relevance in AML. All the SNVs with classification other than synonymous with at least 10 case reports of "hematopoietic and lymphoid tissues" were considered hotspot variants. For the future implementation of our findings, we reasoned that any hybrid-capture probe design and short sequencing reads would efficiently encompass at least several genomic bases surrounding these hotspots. Therefore, we extended the variant calls to capture mutations with a putative deleterious effect that are within five–amino acid distance surrounding each hotspot variant. Genomic loci that were found to be mutated in the training cohort (pre-AML1 and pre-AML2) were used for the final model derivation (table S8). Notably, we discarded genomic loci with mutations in *KIT*, *KRAS*, and *PHF6* as these were found solely in the training set's controls. Such enrichment surely does not correlate with real-life evidence and can bias classification. We then used a random forest algorithm via the R package randomForest. Mutations were grouped by genes, and their VAFs were used to train the model together with the age of the individuals at sampling and the number of the mutations that they carry. If more than one mutation was detected in the same gene, the highest VAF was used. The number of features used for each one of the 5000 generated trees was two.

Simulated controls

To simulate a large population screen, we used the mutations detected by Espresso in the controls from the pre-AML1 and pre-AML2 (termed merged dataset here). We first calculated the frequency of controls that carry at least one mutation at the following age groups: 20 to 49, 50 to 64, 65 to 74, and >75 years old. For these age groups, we obtained the incidence rates of AML through the Surveillance, Epidemiology, and End Results Program (*53*). By assuming similar age distribution for the validation cohort (*8*) and the individuals interrogated in the merged dataset and knowing the number of pre-AML cases interrogated in the validation cohort (n = 188), we were able to estimate the number of simulated controls needed to mimic real incidence rates for each age group. Overall, 4,033,904 controls were simulated.

The frequency of ARCH and the number of mutations that each individual carries within each control age group from the merged dataset helped us to estimate how many of the simulated individuals are expected to carry mutations in the relevant genomic loci (table S8). Overall, 5.05, 7.69, 10.70, and 19.09% of the individuals within the age range of 20 to 49, for 50 to 64, for 65 to 74, and ≥75 years,

respectively, were simulated to have ARCH. A total of 285,629 individuals (~7%) were simulated to carry one mutation, 934 with two mutations (~0.02%), and 156 with three mutations (~0.004%). We next assigned the specific mutations to the simulated individuals based on their association with each age group. For example, for the 149,423 simulated mutated controls with a simulated age of 50 to 64, we populated a list of 149,423 specific mutations that were detected in control individuals in the same age group or in younger age groups in the merged dataset. We also allowed 10% of the mutations detected in the merged dataset in one age group older to be randomly included. Last, we aimed to assign VAF to the simulated mutations. We observed that the VAF of the detected mutations in the merged dataset did not significantly correlate with age [$R_{(\text{Pearson})} = 0.20$; $P = 0.07$] and that a lognormal distribution accurately captures the VAF distribution among all the detected mutations. We therefore used the "rlnorm" R function to simulate VAFs. This resulted with a median VAF of 1.45% and a mean VAF of 2.45% for the simulated controls; 37.46% of the simulated VAFs received a value of VAF $\geq$ 2%. As intended, these values are highly comparable with those of the mutations found in the merge dataset's controls (table S7).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/50/eabe3722/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. Cancer Genome Atlas Research Network, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
2. The International Cancer Genome Consortium, International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
3. M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K. S. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortés-Ciriano, D. C. Zhou, W.-W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang; MC3 Working Group; Cancer Genome Atlas Research Network, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 1034–1035.e18 (2018).
4. Y. Shu, X. Wu, X. Tong, X. Wang, Z. Chang, Y. Mao, X. Chen, J. Sun, Z. Wang, Z. Hong, L. Zhu, C. Zhu, J. Chen, Y. Liang, H. Shao, Y. W. Shao, Circulating tumor DNA mutation profiling by targeted next generation sequencing provides guidance for personalized treatments in multiple cancer types. *Sci. Rep.* **7**, 583 (2017).
5. T. A. Clark, J. H. Chung, M. Kennedy, J. D. Hughes, N. Chennagiri, D. S. Lieber, B. Fendler, L. Young, M. Zhao, M. Coyne, V. Breese, G. Young, A. Donahue, D. Pavlick, A. Tsiros, T. Brennan, S. Zhong, T. Mughal, M. Bailey, J. He, S. Roels, G. M. Frampton, J. M. Spoerke, S. Gendreau, M. Lackner, E. Schleifman, E. Peters, J. S. Ross, S. M. Ali, V. A. Miller, J. P. Gregg, P. J. Stephens, A. Welsh, G. A. Otto, D. Lipson, Analytical validation of a hybrid capture–based next-generation sequencing clinical assay for genomic profiling of cell-free circulating tumor DNA. *J. Mol. Diagn.* **20**, 686–702 (2018).
6. J. Phallen, M. Sausen, V. Adleff, A. Leal, C. Hruban, J. White, V. Anagnostou, J. Fiksel, S. Cristiano, E. Papp, S. Speir, T. Reinert, M.-B. W. Orntoft, B. D. Woodward, D. Murphy, S. Parpart-Li, D. Riley, M. Nesselbush, N. Sengamalay, A. Georgiadis, Q. K. Li, M. R. Madsen, F. V. Mortensen, J. Huiskens, C. Punt, N. van Grieken, R. Fijneman, G. Meijer, H. Husain, R. B. Scharpf, L. A. Diaz Jr., S. Jones, S. Angiuoli, T. Ørntoft, H. J. Nielsen, C. L. Andersen, V. E. Velculescu, Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* **9**, eaan2415 (2017).
7. C. P. Paweletz, A. G. Sacher, C. K. Raymond, R. S. Alden, A. O'Connell, S. L. Mach, Y. Kuang, L. Gandhi, P. Kirschmeier, J. M. English, L. P. Lim, P. A. Jänne, G. R. Oxnard, Bias-corrected targeted next-generation sequencing for rapid, multiplexed detection of actionable alterations in cell-free DNA from advanced lung cancer patients. *Cancer Res.* **22**, 915–922 (2016).
8. P. Desai, N. Mencia-Trinchant, O. Savenkov, M. S. Simon, G. Cheang, S. Lee, M. Samuel, E. K. Ritchie, M. L. Guzman, K. V. Ballman, G. J. Roboz, D. C. Hassane, Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.* **24**, 1015–1023 (2018).
9. S. Abelson, G. Collord, W. Stanley, K. Ng, O. Weissbrod, N. M. Cohen, E. Niemeyer, N. Barda, P. C. Zuzarte, L. Heisler, Y. Sundaravadanam, R. Luben, S. Hayat, T. T. Wang, Z. Zhao, I. Cirlan, T. J. Pugh, D. Soave, K. Ng, C. Latimer, C. Hardy, K. Raine, D. Jones, D. Hoult, A. Britten, J. D. Mcpherson, M. Johansson, F. Mbabaali, J. Eagles, J. K. Miller, D. Pasternack, L. Timms, P. Krzyzanowski, P. Awadalla, R. Costa, E. Segal, S. V. Bratman, P. Beer, S. Behjati, I. Martincorena, J. C. Y. Wang, K. M. Bowles, J. R. Quirós, A. Karakatsani, C. La Vecchia, A. Trichopoulou, E. Salamanca-Fernández, J. M. Huerta, A. Barricarte, R. C. Travis, R. Tumino, G. Masala, H. Boeing, S. Panico, R. Kaaks, A. Krämer, S. Sieri, E. Riboli, P. Vineis, M. Foll, J. McKay, S. Polidoro, N. Sala, K. T. Khaw, R. Vermeulen, P. J. Campbell, E. Papaemmanuil, M. D. Minden, A. Tanay, R. D. Balicer, N. J. Wareham, M. Gerstung, J. E. Dick, P. Brennan, G. S. Vassiliou, L. I. Shlush, Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
10. A. M. Newman, S. V. Bratman, J. To, J. F. Wynne, N. C. W. Eclov, L. A. Modlin, C. L. Liu, J. W. Neal, H. A. Wakelee, R. E. Merritt, J. B. Shrager, B. W. Loo Jr., A. A. Alizadeh, M. Diehn, An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
11. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9530–9535 (2011).
12. M. W. Schmitt, J. B. Hiatt, L. A. Loeb, S. R. Kennedy, E. J. Fox, J. J. Salk, Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci.* **109**, 14508–14513 (2012).
13. C. S. Kim, S. Mohan, M. Ayub, D. G. Rothwell, C. Dive, G. Brady, C. Miller, In silico error correction improves cfDNA mutation calling. *Bioinformatics* **35**, 2380–2385 (2018).
14. D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, R. K. Wilson, VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
15. C. Kockan, F. Hach, I. Sarrafi, R. H. Bell, B. McConeghy, K. Beja, A. Haegert, A. W. Wyatt, S. V. Volik, K. N. Chi, C. C. Collins, S. C. Sahinalp, SiNVICT: Ultra-sensitive detection of single nucleotide variants and indels in circulating tumour DNA. *Bioinformatics* **33**, 26–34 (2016).
16. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, G. Getz, Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
17. C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray, R. K. Cheetham, Strelka: Accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
18. M. Gerstung, C. Beisel, M. Rechsteiner, P. Wild, P. Schraml, H. Moch, N. Beerenwinkel, Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
19. Y. Shiraishi, Y. Sato, K. Chiba, Y. Okuno, Y. Nagata, K. Yoshida, N. Shiba, Y. Hayashi, H. Kume, Y. Homma, M. Sanada, S. Ogawa, S. Miyano, An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
20. 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, A. A. Alizadeh, Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **34**, 547–555 (2016).
21. M. Gerstung, E. Papaemmanuil, P. J. Campbell, Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).
22. D. E. Larson, C. C. Harris, K. Chen, D. C. Koboldt, T. E. Abbott, D. J. Dooling, T. J. Ley, E. R. Mardis, R. K. Wilson, L. Ding, SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2011).
23. T. Moriyama, S. Imoto, S. Hayashi, Y. Shiraishi, S. Miyano, R. Yamaguchi, A Bayesian model integration for mutation calling through data partitioning. *Bioinformatics* **35**, 4247–4254 (2019).
24. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
25. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. J. R. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Børresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjörd, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinsk, N. Jäger, D. T. W. Jones, D. Jonas, S. Knappskog, M. Koo, S. R. Lakhani, C. López-Otín, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. J. Tutt, R. Valdés-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson,

S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, M. R. Stratton, Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
26. R. C. Poulos, J. Olivier, J. W. H. Wong, The interaction between cytosine methylation and processes of DNA replication and repair shape the mutational landscape of cancer genomes. *Nucleic Acids Res.* **45**, 7786–7795 (2017).
27. 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, 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.* **41**, e67 (2013).
28. D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, A. Gnirke, Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
29. Y. Benjamini, T. P. Speed, Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
30. J. I. Odegaard, J. J. Vincent, S. Mortimer, J. V. Vowles, B. C. Ulrich, K. C. Banks, S. R. Fairclough, O. A. Zill, M. Sikora, R. Mokhtari, D. Abdueva, R. J. Nagy, C. E. Lee, L. A. Kiedrowski, C. P. Paweletz, H. Eltoukhy, R. B. Lanman, D. I. Chudova, A. A. Talasaz, Validation of a plasma-based comprehensive cancer genotyping assay utilizing orthogonal tissue- and plasma-based methodologies. *Clin. Cancer Res.* **24**, 3539–3549 (2018).
31. M. Jongen-Lavrencic, T. Grob, D. Hanekamp, F. G. Kavelaars, A. al Hinai, A. Zeilemaker, C. A. J. Erpelinck-Verschueren, P. L. Gradowska, R. Meijer, J. Cloos, B. J. Biemond, C. Graux, M. van Marwijk Kooy, M. G. Manz, T. Pabst, J. R. Passweg, V. Havelange, G. J. Ossenkoppele, M. A. Sanders, G. J. Schuurhuis, B. Löwenberg, P. J. M. Valk, Molecular minimal residual disease in acute myeloid leukemia. *N. Engl. J. Med.* **378**, 1189–1199 (2018).
32. H.-H. Zhu, X.-H. Zhang, Y.-Z. Qin, D.-H. Liu, H. Jiang, H. Chen, Q. Jiang, L.-P. Xu, J. Lu, W. Han, L. Bao, Y. Wang, Y.-H. Chen, J.-Z. Wang, F.-R. Wang, Y.-Y. Lai, J.-Y. Chai, L.-R. Wang, Y.-R. Liu, K.-Y. Liu, B. Jiang, MRD-directed risk stratification treatment may improve outcomes of t(8;21) AML in the first complete remission: Results from the AML05 multicenter trial. *Blood* **121**, 4056–4062 (2013).
33. U. Platzbecker, J. M. Middeke, K. Sockel, R. Herbst, D. Wolf, C. D. Baldus, U. Oelschlägel, A. Mütherig, L. Fransecky, R. Noppeney, G. Bug, K. S. Götze, A. Krämer, T. Bochtler, M. Stelljes, C. Groth, A. Schubert, M. Mende, F. Stölzel, C. Borkmann, A. S. Kubasch, M. von Bonin, H. Serve, M. Hänel, U. Dührsen, J. Schetelig, C. Röllig, M. Kramer, G. Ehninger, M. Bornhäuser, C. Thiede, Measurable residual disease-guided treatment with azacitidine to prevent haematological relapse in patients with myelodysplastic syndrome and acute myeloid leukaemia (RELAZA2): An open-label, multicentre, phase 2 trial. *Lancet Oncol.* **19**, 1668–1679 (2018).
34. M. Balsat, A. Renneville, X. Thomas, S. de Botton, D. Caillot, A. Marceau, E. Lemasle, J.-P. Marolleau, O. Nibourel, C. Berthon, E. Raffoux, A. Pigneux, C. Rodriguez, N. Vey, J.-M. Cayuela, S. Hayette, T. Braun, M. M. Coudé, C. Terre, K. Celli-Lebras, H. Dombret, C. Preudhomme, N. Boissel, Postinduction minimal residual disease predicts outcome and benefit from allogeneic stem cell transplantation in acute myeloid leukemia with *NPM1* mutation: A study by the acute leukemia french association group. *J. Clin. Oncol.* **35**, 185–193 (2016).
35. T. N. Wong, G. Ramsingh, A. L. Young, C. A. Miller, W. Touma, J. S. Welch, T. L. Lamprecht, D. Shen, J. Hundal, R. S. Fulton, S. Heath, J. D. Baty, J. M. Klco, L. Ding, E. R. Mardis, P. Westervelt, J. F. Dipersio, M. J. Walter, T. A. Graubert, T. J. Ley, T. E. Druley, D. C. Link, R. K. Wilson, Role of *TP53* mutations in the origin and evolution of therapy-related acute myeloid leukaemia. *Nature* **518**, 552–555 (2015).
36. T. N. Wong, C. A. Miller, M. R. M. Jotte, N. Bagegni, J. D. Baty, A. P. Schmidt, A. F. Cashen, E. J. Duncavage, N. M. Helton, M. Fiala, R. S. Fulton, S. E. Heath, M. Janke, K. Luber, P. Westervelt, R. Vij, J. F. DiPersio, J. S. Welch, T. A. Graubert, M. J. Walter, T. J. Ley, D. C. Link, Cellular stressors contribute to the expansion of hematopoietic clones of varying leukemic potential. *Nat. Commun.* **9**, 455 (2018).
37. O. A. Guryanova, K. Shank, B. Spitzer, L. Luciani, R. P. Koche, F. E. Garrett-Bakelman, C. Ganzel, B. H. Durham, A. Mohanty, G. Hoermann, S. A. Rivera, A. G. Chramiec, E. Pronier, L. Bastian, M. D. Keller, D. Tovbin, E. Loizou, A. R. Weinstein, A. R. Gonzalez, Y. K. Lieu, J. M. Rowe, F. Pastore, A. S. McKenney, A. V. Krivtsov, W. R. Sperr, J. R. Cross, C. E. Mason, M. S. Tallman, M. E. Arcila, O. Abdel-Wahab, S. A. Armstrong, S. Kubicek, P. B. Staber, M. Gönen, E. M. Paietta, A. M. Melnick, S. D. Nimer, S. Mukherjee, R. L. Levine, DNMT3A mutations promote anthracycline resistance in acute myeloid leukemia via impaired nucleosome remodeling. *Nat. Med.* **22**, 1488–1495 (2016).
38. G. J. Schuurhuis, M. Heuser, S. Freeman, M.-C. Béné, F. Buccisano, J. Cloos, D. Grimwade, T. Haferlach, R. K. Hills, C. S. Hourigan, J. L. Jorgensen, W. Kern, F. Lacombe, L. Maurillo, C. Preudhomme, B. A. van der Reijden, C. Thiede, A. Venditti, P. Vyas, B. L. Wood, R. B. Walter, K. Döhner, G. J. Roboz, G. J. Ossenkoppele, Minimal/measurable residual disease in AML: A consensus document from the European LeukemiaNet MRD Working Party. *Blood* **131**, 1275–1291 (2018).
39. F. Buccisano, L. Maurillo, M. I. Del Principe, A. Di Veroli, E. De Bellis, A. Biagi, A. Zizzari, V. Rossi, V. Rapisarda, S. Amadori, M. T. Voso, F. Lo-Coco, W. Arcese, A. Venditti, Minimal residual disease as a biomarker for outcome prediction and therapy optimization in acute myeloid leukemia. *Expert Rev. Hematol.* **11**, 307–313 (2018).
40. L. I. Shlush, Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
41. D. P. Steensma, Clinical consequences of clonal hematopoiesis of indeterminate potential. *Blood Adv.* **2**, 3404–3410 (2018).
42. F. Zink, S. N. Stacey, G. L. Norddahl, M. L. Frigge, O. T. Magnusson, I. Jonsdottir, T. E. Thorgeirsson, A. Sigurdsson, S. A. Gudjonsson, J. Gudmundsson, J. G. Jonasson, L. Tryggvadottir, T. Jonsson, A. Helgason, A. Gylfason, P. Sulem, T. Rafnar, U. Thorsteinsdottir, D. F. Gudbjartsson, G. Masson, A. Kong, K. Stefansson, Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
43. C. C. Coombs, A. Zehir, S. M. Devlin, A. Kishtagari, A. Syed, P. Jonsson, D. M. Hyman, D. B. Solit, M. E. Robson, J. Baselga, M. E. Arcila, M. Ladanyi, M. S. Tallman, R. L. Levine, M. F. Berger, Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. *Cell Stem Cell* **21**, 374–382.e4 (2017).
44. T. Hao, M. Li-Talley, A. Buck, W. Chen, An emerging trend of rapid increase of leukemia but not all cancers in the aging population in the United States. *Sci. Rep.* **9**, 12070 (2019).
45. E. Riboli, K. J. Hunt, N. Slimani, P. Ferrari, T. Norat, M. Fahey, U. R. Charrondière, B. Hémon, C. Casagrande, J. Vignat, K. Overvad, A. Tjønneland, F. Clavel-Chapelon, A. Thiébaut, J. Wahrendorf, H. Boeing, D. Trichopoulos, A. Trichopoulou, P. Vineis, D. Palli, H. B. Bueno-De-Mesquita, P. H. M. Peeters, E. Lund, D. Engeset, C. A. González, A. Barricarte, G. Berglund, G. Hallmans, N. E. Day, T. J. Key, R. Kaaks, R. Saracci, European prospective investigation into cancer and nutrition (EPIC): Study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).
46. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
47. 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, L. A. Loeb, Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
48. T. T. Wang, S. Abelson, J. Zou, T. Li, Z. Zhao, J. E. Dick, L. I. Shlush, T. J. Pugh, S. V. Bratman, High efficiency error suppression for accurate detection of low-frequency variants. *Nucleic Acids Res.* **47**, e87 (2019).
49. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
50. The 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
51. G. Glusman, J. Caballero, D. E. Mauldin, L. Hood, J. C. Roach, Kaviar: An accessible system for testing SNV novelty. *Bioinformatics* **27**, 3216–3217 (2011).
52. S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C. YinKok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, P. J. Campbell, COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
53. SEER Cancer Stat Facts—National Cancer Institute (2017); https://seer.cancer.gov/statfacts/.
54. A. Venkatesan, K. Kerlikowske, P. Chu, R. Smith-Bindman, E. A. Sickles, Positive predictive value of specific mammographic findings according to reader and patient variables. *Radiology* **250**, 648–657 (2009).
55. M. Navarro, A. Nicolas, A. Ferrandez, A. Lanas, Colorectal cancer population screening programs worldwide in 2016: An update. *World J. Gastroenterol.* **23**, 3632–3642 (2017).

Acknowledgments: We acknowledge the support from the Princess Margaret Cancer Foundation. We thank the Genome Technologies team at the Ontario Institute for Cancer Research as well as the Princess Margaret Genomics Centre for technical expertise with sequencing, data management, and QC analysis. We acknowledge the dedicated work of G. Vassiliou from the Wellcome Trust Sanger Institute for supervising sample acquisition from the EPIC-Norfolk longitudinal cohort and for the management of the sequencing data, its creation, and its sharing. **Funding:** S.A. received support from the Benjamin Pearl fellowship from the McEwen Centre for Regenerative Medicine and is funded by the Ontario Institute for Cancer Research. A.G.X.Z. was supported by the CIHR Vanier Scholarship. S.V.B., S.M.C., and T.J.P. were supported by the Gattuso-Slaight Personalized Cancer Medicine Fund at the Princess Margaret Cancer Centre. J.E.D. was supported by funds from the Princess Margaret Cancer Centre Foundation, Ontario Institute for Cancer Research, with funding from the Province of Ontario, Canadian Institutes for Health Research, Canadian Cancer Society Research Institute, Terry Fox Foundation, Genome Canada through the Ontario Genomics Institute, and the Canada Research Chair. **Author contributions:** S.A. developed the concept and the code, led and supervised the analysis of the data, and wrote the manuscript. A.G.X.Z. created the R package and contributed to the development of the analytical pipeline. I.N.-M.

contributed to the development of the R package and the correspondence with the reviewers. S.W.K.N. provided support for genetic predictive model derivation. T.T.W. and T.J.P. provided bioinformatics support and contributed to the analysis of the data. M.D.M., T.M., and S.M.C. enabled sample acquisition and clinical data curation and provided clinical expertise. L.I.S. and P.A. contributed to the study design and/or in obtaining data and materials. J.E.D. and S.V.B. supervised all aspects of the study and wrote the manuscript. All authors commented on the manuscript at all stages. **Competing interests:** S.V.B. is co-inventor on patents, distinct from the subject matter of this manuscript, relating to circulating tumor DNA analysis, one of which has been licensed to Roche Molecular Diagnostics. S.V.B. provides consultation for DNAMx Inc. and has received research funding from Nektar Therapeutics. T.J.P. provides consultation for Merck, Chrysalis Biomedical Advisors, the Canadian Pension Plan Investment Board, and Illumina (compensated) and receives research support (institutional) from Roche/Genentech. All other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All the sequencing files used in this manuscript can be downloaded from the European Genome-phenome Archive (EGA) or the NCBI Sequence Read Archive under the following accession numbers: (i) pre-AML1, EGAD00001003583; (ii) pre-AML2, EGAD00001003703; (iii) CB, EGAD00001005261; (iv) AML-MRD, EGAD00001005270; and (v) CL, SRP141184. Software and supplemental code are available on GitHub at www.github.com/abelson-lab/Espresso and www.github.com/abelson-lab/Espresso_paper, respectively. Additional data related to this paper may be requested from the authors.

Submitted 17 August 2020
Accepted 23 October 2020
Published 9 December 2020
10.1126/sciadv.abe3722

Citation: S. Abelson, A. G. X. Zeng, I. Nofech-Mozes, T. T. Wang, S. W. K. Ng, M. D. Minden, T. J. Pugh, P. Awadalla, L. I. Shlush, T. Murphy, S. M. Chan, J. E. Dick, S. V. Bratman, Integration of intra-sample contextual error modeling for improved detection of somatic mutations from deep sequencing. *Sci. Adv.* **6**, eabe3722 (2020).