**SUPPLEMENTARY METHODS**

**High-Intensity Sequencing Reveals the Sources of Plasma Circulating Cell-Free DNA Variants**

Razavi et al.

**cfDNA analysis pipeline**

### **Preprocessing and alignment**

Raw bcl files from the sequencer were demultiplexed and converted to fastq format using bcl2fastq version 2.18.0.6 (Illumina). Settings in the sample sheets were configured so that bcl2fastq trimmed the first 7 bp at the 5’ end of the reads containing the 6 bp UMI along with the ‘T’ base, following which, the UMI sequences were added to the read name. The fastqs were further processed to trim 3’ overhangs based on pairwise alignment of the paired-end reads. The reads were then aligned using BWA MEM (*56*) to a version of the human reference genome created by combining the UCSC hg19 assembly with decoy sequences.

### **Error correction and read pair stitching**

Error correction was performed by first identifying reads originating from the same source DNA fragment and building a consensus sequence from those reads. Aligned reads were bagged using mapping coordinates (5’ end of the paired-end reads) and UMIs added during library construction step. Each read pair within a bag represented a copy of either the reference or non-reference strand of the initial DNA fragment. A read-pair was derived from the reference (or the non-reference) strand of the initial fragment if the read pair was mapped in F1R2 (or F2R1) orientations. In bagging the reads, a small amount of *slack* (1 bp) was allowed in the position of the 5’ end of the reads, suggesting that read pairs in which the 5’ end of one of the two reads is off by 1 bp could still go into the same bag, provided that they shared the same UMI pair with the other read pairs. Additionally, up to two mismatches were allowed within the combined UMI length (12 bp) to account for sequencing errors. Bagging was performed in a greedy fashion by picking a founder pair and adding all read pairs that satisfied the 1 bp slack and UMI error requirements with the founder pair.

Reads within each bag were error-corrected by building a reference-free consensus of each set of reads in the bag. A proto-consensus was initialized by copying one of the reads in the set. This proto-consensus was adjusted by adding the rest of the reads in the bag one at a time and updating the proto-consensus to the pairwise alignment between the proto-consensus and the newly added read. At the end of this process, the proto-consensus pair was converted to a collapsed error-corrected pair by sliding through the proto-consensus one position at a time and selecting a consensus base and nominal quality score for that position based on the sum of quality scores for each base at that position. A collapsed read was called duplex-collapsed if both strands of the initial molecule were represented in the bag. The read was called non-duplex collapsed if only one of the two strands was represented in the bag. Duplex-collapsed reads had lower error rates because of a better chance of correcting PCR and sequencing errors when both strands were represented in the bag. Bags with a single read pair were omitted from the output of the error correction step, as error correction was not possible on those reads.

After the initial error correction, collapsed reads were re-aligned to the reference genome. Collapsed read pairs with sufficient implied overlap were stitched together. This step builds a single contiguous sequence for the entire fragment, which provides better context for downstream assembly. In addition, this allows for further error correction when there is a discrepancy between the base calls from the two paired end reads within the overlap between the two reads. The paired-end collapsed reads that did not have sufficient overlap were unaffected by stitching.

### **Candidate variant generation by de novo assembly**

A *de novo* assembly was constructed for each contiguous target region based on the error corrected read pairs. The assembly provided a compact, reference-free representation of all the haplotypes observed in the reads. For each target region, an optimal *k*-mer size *k* ∈ (48, 64, 96) was dynamically selected based on the largest perfect repeat 𝜆 within that region. The selected *k* was the smallest *k* such that *k* ≥ 𝜆. Once *k* was selected, a De Bruijn graph was assembled based on *k*-mers from all reads overlapping the target region. The error-corrected reads were first localized to target regions by alignment to the reference genome sequence. The reads in each target region were assembled into a de Bruijn graph, the reads aligned to the graph, and a read support vector was populated for each *k*-mer in the graph based on the lowest quality base in each occurrence of that *k*-mer in an error-corrected read. The graph was iteratively pruned by applying heuristics to remove edges with weak support. Variant calling was performed by aligning the edges of the pruned graph to a reference and listing genomic coordinates where an edge was different from the reference. By extending the edge on either side with the edges with the highest support upstream and downstream, it was possible to provide a longer context for the alignment than what was possible based on a single read, which helped avoid alignment-related artifacts in variant calling.

***Reproducibility of the targeted DNA assay***

Two similar targeted DNA assay protocols (V1 and V2) for plasma cfDNA and matching WBC gDNA samples were employed in this study. The main differences were the UMI sequences in the library adapters and the reaction volumes in hybridization enrichment process, neither of which would be expected to influence results. To ensure assay performance equivalence, six patient samples with a large cfDNA yield were selected (**Supplementary Table 6**), allowing for reprocessing with both assay protocols, as well as droplet digital PCR (ddPCR). Bio-Rad ddPCR was used to measure canonical hotspot variants in five of the six patient samples, revealing good agreement with measurements using the targeted DNA assays (**Fig. 1b**). **Fig. 1c** and **Supplementary Fig. 3** show the measured allele fraction of variants called either using the V1 or V2 targeted DNA assay protocol. Measured VAFs between the two technical replicates for samples from five patients (**Fig. 1c**) showed a strong agreement (*R2* = 0.9997) as well as for one sample from a hypermutated case (**Supplementary Fig. 3,** *R2* = 0.9972).

### **Machine learning error model**

Quality scores were assigned to all called variants using a pair of probabilistic error models. For SNVs and MNVs, a hierarchical Bayesian model with one set of noise parameters per genomic position in the panel and alternative allele was trained using baseline control samples from the technical validation study. Candidate variants generated from the De Bruijn graph were scored relative to the collapsed pileup information without base alignment quality. The model assumed that each position and alternative allele in the genome, indexed by p, had a baseline noise rate dp 𝜆p, where dp was the total depth at this position. This noise rate is defined as the expected alternate allele count at each position and alternative allele, assuming yp~ Pois(dp 𝜆p). Each 𝜆p was also assumed to be drawn from a mixture of Gamma distributions, the parameters of which varied based on a discrete covariate. This discrete covariate encoded trinucleotide context, whether this position was uniquely alignable (based on ENCODE tracks), and whether the position fell within a known segmental duplication. The full hierarchical model was trained by drawing from the joint posterior of all position-level and global parameters using a Markov chain Monte-Carlo sampler. Metropolis-Hastings within Gibbs was used to update the global parameters simultaneously, accounting for their posterior correlation during sampling. 2000 iterations of burn-in and 2000 sampling iterations were used, with four chains compared to assess convergence based on the Gelman-Rubin diagnostic criterion.

Following posterior sampling, the marginal posterior of each 𝜆p was summarized by its mean and variance to obtain a Gamma approximation. This Gamma approximation was then translated into a negative Binomial posterior predictive distribution for the observed number of alternative allele counts at each position and alternative allele. For each position, dispersion re-estimation was subsequently performed for each position and alternative allele to address deviations from the Poisson assumption. A maximum likelihood estimate of the negative Binomial dispersion parameter was obtained for each position and alternative allele, holding the rate fixed based on the posterior predictive approximation. The maximum of the original dispersion parameter from the posterior predictive distribution and the maximum likelihood was retained. This re-estimated posterior predictive approximation for each position and alternative allele provided the basis for all quality scores. Formally, this corresponded to assuming an approximate posterior predictive distribution yp~ NB(mpdp, rp), where mp and rp were the rate and dispersion parameters as discussed above. For each observed alternative allele, the quality score was defined as -10 \* log10(P(ADp>= yp)), where the probability was computed based on the re-estimated posterior predictive approximation described above. For observed MNVs, a negative Binomial approximate posterior predictive distribution was computed matching the mean and average variance of its component SNVs, followed by the quality score calculation as above.

For indels, a similar Bayesian hierarchical model was employed (**Supplementary Fig. 10**). Each anchor position p and indel length l (positive for insertions, negative for deletions) in the panel was assumed to have a baseline noise rate dp 𝜆pl. Additionally, 𝜆pl = 𝜆pωlp. 𝜆p was assumed to be drawn from a Gamma distribution with parameters unique to the combination of proximity to segmental duplications, lack of unique alignability, and repeats as identified by RepeatMasker. 𝛚p (a probability vector) was assumed to be drawn from a Dirichlet distribution with parameters varied based on the same factors. Training was done via a similar Metropolis-Hastings within Gibbs algorithm, with analogous numbers of iterations and convergence diagnostics. The same dispersion approximation and re-estimation was applied to each 𝜆pl, using the marginal posterior obtained from the samples. Quality scores for observed indels were computed as described above using these approximate posterior predictive distributions. The estimation processes described above were carried out separately for cfDNA and WBC samples, resulting in four models in total (SNV and indel models for each of cfDNA and WBC).

### **Edge artifact scoring**

DNA damage processes similar to those observed in ancient DNA samples generated artifacts in samples used in this study. These “edge artifacts” were frequently near the ends of fragments. This error source was controlled by using a discrete mixture model. Specifically, for each putative artifact, the following metrics were considered: (1) the median distance of the putative artifact from the fragment end, (2) the Phred-scaled p-value (from the noise model) of the putative artifact, (3) the concentration towards fragment ends based on a Wilcoxon rank-sum test, and (4) the observed putative artifact allele fraction. Each metric was separately discretized to a resolution of 5 for the median distance and Phred-scaled p-value, and to a resolution of 0.01 for allele fractions. The model assumed that the observed distribution of these features for “at-risk” G>A and C>T variants consisted of a mixture of edge artifact and null (non-artifact) distributions. The proportion of SNVs originating from each mixture component was assumed to vary across samples. For training, SNVs, except for G>A and C>T, were considered as negative controls, assuming they come from the null distribution. Similarly, for training the model, variants considered at risk for artifactual origin were those with a median edge distance below 25 bp and a median absolute deviation in edge distance below 25 bp. Additionally, it was assumed that the discrete mixture distribution for the artifact component factorized across features. The null mixture component was estimated directly from the observed joint distribution of the given features in the negative control SNVs. The alternative mixture component was estimated via maximum likelihood, holding the null distribution estimate fixed. The model was trained using SNVs from held-out healthy controls and samples from the MSK site not used in this study.

The resulting mixture model was applied to each sample. On a per-sample basis, the proportion of at-risk candidate variants (G>A and C>T SNPs passing error model) was estimated via maximum likelihood, fixing the estimates of mixture components. These estimated proportions varied in training samples, from less than 1% to over 98%. This estimate was then applied to compute, for each such variant, the posterior probability that it originated from the edge artifact mixture component. SNVs with posterior probability greater than 1% were flagged as potential edge artifacts and removed from further analysis.

### **Joint variant analysis using the machine learning error model**

The noise models described above were used to assess the support for non-WBC variant origins. Starting from the approximate posterior predictive distributions estimated above, the observed alternative depth for each putative variant was assumed to correspond to the sum of negative Binomial noise, with parameters based on the noise model, and Poisson-distributed signal. Formally, assume ADcfDNA,p~ Pois(AFcfDNA,pdp) + NB(mpdp, rp) for cfDNA, with an analogous model for the WBC observation. From this assumption, the likelihood of the observed pair of cfDNA and WBC alternative depths was computed, conditional on the cfDNA and WBC total depths, noise parameters, and hypothesized alternative allele fractions. This provided a joint likelihood function parameterized by the pair of cfDNA and WBC alternative allele fractions (parameters). Combining this likelihood with a hypothesized joint uniform prior on the alternative allele fractions, the posterior probability P(AFcfDNA,p>= k \* AFWBC,p) was computed. Joint calling amounted to setting a threshold on this probability, and selecting a functional form for k.

The threshold on this posterior probability and functional form for k were jointly determined from control data and structured cross-validation across tumor types. The functional form of k() was restricted to a piecewise linear function of the observed WBC allele fraction. For allele fractions greater than 1/3, k was set to 3 to address the appearance of germline heterozygous variants within regions of altered copy number. A range of values for k(0) from 1 to 3 was evaluated, setting the slope as required to maintain continuity of the resulting function. Next, control of false positives was required, defined as minimizing deviation from target false positive range of 1 variant per sample in held-out healthy control samples with a binomial 95% upper confidence bound of at most 3 variants per sample. Within the set of k(0) and probability threshold combinations admitted by these criteria, cross-validation across tumor types was employed to select final parameters. For each tumor type, a training set of samples was created consisting of all other tumor types (e.g. the training set for breast cancer samples consisted of NSCLC and CRPC samples). The combination of k(0) and probability threshold that maximized sensitivity in the training set from the admissible set defined above was then selected. The selected parameters were then employed on the test set, consisting of the original samples (e.g. breast cancer samples in the preceding example).