Supplementary Note

Barcode collisions. In designing our 4bp physical barcodes, we considered the possibility that distinct parental molecules could accidentally receive the same exogenous UID (generally known as the "birthday problem"[1]). Such "barcode collisions" can only happen when different DNA molecules share identical start/end coordinates. We previously observed redundant start/end coordinates in <50% of cfDNA and <10% of acoustically shorn gDNA molecules in prior work (Supplementary Fig. 4a in Newman et al., 2014[2]). Therefore, the majority of recovered molecules were expected to be unique and unaffected by barcode collisions. Moreover, given 256 possible physical UIDs (i.e., 4^4 combinations), the probability of assigning the same UID to two distinct molecules was determined to be only 0.39% (*pbirthday* function in R). Therefore, we predicted that our barcode strategy would have sufficient complexity to avoid performance degradation due to barcode collisions.

To validate our design assumptions, we performed two dedicated analyses. First, using cfDNA samples spanning a range of clinically relevant haploid genome equivalents (hGEs), we estimated the fraction of barcoded molecules that would likely be lost due to the "birthday problem". We determined (i) the distribution of distinct molecules with identical start/end coordinates (**Supplementary Fig. 2a**), and (ii) the number of affected molecules for each bin size *n*. Specifically, given k = 256 possible physical UIDs and a bin size of *n* molecules with identical start/end coordinates, we calculated the number of expected barcode collisions with the following formula, which is commonly used to determine the number of expected collisions in a hash table[3]:

- $E(barcode\ collisions) = n - k + k\left(\frac{k-1}{k}\right)^n$

While the resulting quantity C_n denotes the number of expected collisions within bin size *n*, it does not capture the total number of affected barcodes, which is at most twice this quantity. Therefore, given *N* distinct bin sizes, we conservatively calculated the total number of molecules affected by barcode collisions:

- $Number\ of\ affected\ molecules = \sum_{n=2}^{N} 2C_n$

Based on this approach, we estimated a minor loss of between 0.15% and 0.56% of all recovered molecules (**Supplementary Fig. 2a–c**).

To confirm these estimates, we performed a second analysis to directly measure the rate of such collisions in our own data, by relying on known germline sequence variants within the ligated cfDNA inserts. Specifically, we evaluated UID families with identical start/end coordinates and capturing a cfDNA fragment harboring a previously identified germline SNP known to be heterozygous (i.e., A/B alleles). By focusing on UID families harboring at least one member supporting the alternate (non-reference) allele (i.e., allele B), we were able to determine which UID families had support for both maternal (B) and paternal (A) alleles. Such families, we hypothesized, were more likely to have occurred due to barcode collisions than due to sequencing errors. In support of this hypothesis, UID families with an erroneous reference allele (A) were >13-fold more common than those with another discordant non-reference allele (not A, not B). Importantly, this approach revealed losses of between 0.1% and 0.44% of recovered molecules (**Supplementary Fig. 2d**), similar to our previous estimates on the same cfDNA samples (**Supplementary Fig. 2c**).

Based on these data, the birthday problem is unlikely to significantly impact our short barcodes, because (1) the loss of 0.1-0.5% of molecules to barcode collisions is minor for cfDNA input masses that are clinically obtainable, and because (2) given the generally low fractions of ctDNA observed in patient samples, such collisions are likely to mostly involve wild-type molecules when they occur.

Advantages of short barcodes. Barcodes can either be (1) made enzymatically by polymerase extension over a degenerate synthetic template (as in Kennedy et al. 2014[4]), or (2) synthesized in a single unit with the adapter sequences. The first strategy relies on the diversity of synthesis of degenerate N-mers, which may contain biases[5], and the processivity of polymerases, which can vary as a function of such templates. The second strategy offers the advantage of defined composition with desired diversity through the balanced mixing of individually synthesized oligos, but becomes increasingly expensive with greater barcode complexity. We initially tested approach #1 but found poor efficiency of cfDNA input recovery. This led us to develop option #2, and these adapters with commercially synthesized barcodes yielded superior hGE recovery, potentially due to improved ligation efficiency due to intact adapter ends. Other considerations included maximizing informative sequencing space and minimizing the alteration of the adapter sequences to preserve proper annealing. After analyzing the expected number of input molecules with identical start/end coordinates, we found that 4-base barcodes can generate sufficient diversity to differentiate the vast majority of molecules for clinically relevant input amounts (see *Barcode ambiguity* above). Thus, short barcodes were used in order to maximize sequencing space and molecule recovery while minimizing adapter cost.

Assay efficiency. To determine the efficiency of the CAPP-Seq library protocol, especially as it relates to error correction and duplex strand recovery, we performed a dedicated experiment, shown schematically in **Supplementary Fig. 3a**. Starting with 32ng of input cfDNA from a healthy donor (=10,560 haploid genome equivalents, or hGEs, assuming 330 hGEs / ng), we performed CAPP-Seq library prep to completion. We then split the post-capture library in two, and sequenced each of these on a separate flow-cell lane (i.e., Lane 1 and 2 in **Supplementary Fig. 3a**). Using shared genomic start/end positions and UIDs, we were able to determine which original molecules were sequenced on both lanes and which were not. To estimate the total number of post-capture molecules, we then used a well-established "mark and recapture" approach from ecology, termed the Lincoln-Petersen method[6]. If L_1 denotes number of distinct (i.e., non-duplicated) molecules recovered in lane 1, L_2 denotes the number of distinct molecules recovered in lane 2, and $L_{1,2}$ denotes the number of distinct molecules recovered in both lanes, then the number of post-capture molecules M can be estimated with the following formula[6]:

- $\widehat{M} = \frac{L_1 L_2}{L_{1,2}}$

The mark and recapture method estimates suggested that ~50-60% of total hGEs that entered the library preparation process made it through post-capture PCR (**Supplementary Fig. 3b**, left). This estimate remained stable even when subsampling down by an order of magnitude (**Supplementary Fig. 3b**, left). In addition, our result was similar to a prior estimate of CAPP-Seq library efficiency based on mass input, number of PCR cycles, and PCR efficiency (Supplementary Fig. 4b in Newman et al. 2014[2]).

We then examined the efficiency of recovering duplex molecules. Regardless of overall sequencing amounts, we estimated a post-capture recovery rate of ~12% (**Supplementary Fig. 3b**, right). This result agreed with expectations of seeing independent single-stranded molecules using sampling models (described below). In fact, post-sequencing duplex recovery rates were also highly predictable (**Supplementary Fig. 4d**), suggesting a lack of significant biases in duplex capture.

Duplex recovery. Duplex sequencing provides exceptional error suppression, but is limited by inefficient recovery of double-stranded (DS) molecules[4]. Since all input molecules are denatured into single-stranded (SS) molecules prior to PCR amplification (followed by hybrid capture and additional PCR for CAPP-Seq), we hypothesized that the duplex recovery rates observed in our data should be consistent with statistical sampling models. If not, this would suggest a potential bias in our assay. In order to test this hypothesis, we used the binomial distribution, which determines the probability of drawing a success (i.e., duplex molecule) with replacement. Although no replacement occurs in actual sequencing data, the probability of finding a matching pair of SS molecules drawn with replacement can be viewed as an approximation to finding a duplex molecule. We therefore estimated the expected number of pairs (i.e., duplex molecules) within a pool of *m* single-stranded hGEs drawn from *n* distinct input hGEs, using the following formula:

- $E(duplex\ molecules) = n(1 - \sum_{i=0}^{1} P\left[Binom\left(m, \frac{1}{n}\right) = i\right])$

Here, the number of distinct input molecules *n* is multiplied by the probability of observing a given molecule ≥2 times, yielding an expectation for the number of recovered duplex molecules. To evaluate this model, we predicted duplex recovery for 201 cfDNA samples sequenced with appropriate barcodes (i.e., insert UIDs; **Supplementary Table 2**). For each sample, *n* was set to the number of input hGEs (input mass in ng x 330 hGEs / ng) and *m* was set to the median on-target depth following de-duplication. Importantly, knowledge of duplex support was deliberately excluded from the latter, such that *m* was only based on single-stranded coverage. We observed a significant linear relationship between predicted and observed duplex-supported molecules (R^2 = 0.79, P<0.0001), with a slope of 1.09 and an intercept of 13 (**Supplementary Fig. 4d**). Thus, probabilistic modeling can accurately predict duplex recovery in our data, suggesting that strand loss is primarily related to subsampling, not to specific technical limitations of our adapter design or library protocol.

Over-sequencing versus barcode recovery. To explore how sequencing depth relates to the recovery of single-stranded consensus sequence (SSCS) and double-stranded consensus sequence (DCS) molecule recovery for a range of input masses, we used a metric that relates the number of raw on-target sequence reads to DNA input mass. Specifically, we calculated the median panel coverage before duplication removal and divided this number by the number of input hGEs. The resulting quantity, termed "Fold over-sequencing relative to input hGEs," provided a convenient metric that is independent of a specific input mass or sequencing amount. We then applied this metric to assess both the number of reads required to build a SSCS molecule (**Supplementary Fig. 4a**) and the number of SSCS molecules required to build a DCS molecule (**Supplementary Fig. 4b**). SSCS yields were highly correlated with level of over-sequencing (R^2 = 0.73). Importantly, this relation remained significantly concordant (P<0.0001) when considering our most common cfDNA input mass (32 ng) separately

from smaller input masses (<32ng) (**Supplementary Fig. 4a**). DCS yields were reasonably predictable by a power function (**Supplementary Fig. 4b**).

To compare our DCS yields to the literature, we performed an analysis originally described in Kennedy et al. (2014)[4] for determining the optimal amount of over-sequencing for duplex recovery. Specifically, we examined the total number of on-target reads required to build a single DCS molecule ("DCS efficiency") as a function of peak family size. The latter is defined as the mode of all SSCS family sizes (i.e., the number of reads per SSCS family) (e.g., Fig 5a in Kennedy et al.[4]). Using a highly over-sequenced 32 ng input sample of cfDNA, we performed *in silico* down-sampling from ~5000 down to ~670 hGEs, and then computed DCS efficiency and peak family sizes at defined intervals. We found an optimal DCS efficiency at peak sizes of between 1 and 2 (**Supplementary Fig. 4c**). This maximal DCS efficiency was achieved at mean SSCS family sizes of ~2 to 2.7. Notably, we observed a peak DCS efficiency that was ~4.5-fold higher than that reported by Kennedy et al. and our optimal peak size was lower, likely owing to fundamental molecular biology differences in our approaches.

Background errors in independent cfDNA sequence data. We observed a high ratio of G>T to C>A changes in cfDNA sequence data with respect to the plus strand of the reference genome (**Fig. 2b**). Based on experimental evidence, we hypothesized that this imbalance could be explained by (i) oxidative damage occurring during target enrichment and (ii) capture baits that exclusively target the plus strand (**Supplementary Fig. 6d**). We also hypothesized that this damage was due to 8-oxoG, however, unlike oxidized oil-mediated 8-oxoG[7] and sonication-induced 8-oxoG[8], the specific mechanism for G>T/C>A damage in our data remains unclear. To examine the reproducibility of our capture-based model using independent sequencing data, we analyzed two studies that applied hybrid capture methods to cfDNA and that did not perform sonication.

The first study, by De Mattos-Arruda and colleagues, used the same capture reagent as CAPP-Seq (NimbleGen SeqCap), but applied a different library preparation protocol[9]. In a representative subset of their data (4 plasma samples and 1 cerebrospinal fluid (CSF) sample), we observed a significantly higher ratio of G>T to C>A errors (**Supplementary Fig. 6e, left**), consistent with our model. In another set of samples, the authors applied whole exome profiling to cfDNA samples using the Nextera Rapid Capture Exome kit (37Mb) (Illumina). Examining a randomly chosen subset of 2 patients, we also observed an imbalance in G>T to C>A errors, however the ratio of errors was reversed (**Supplementary Fig. 6e, right**). This striking reversal in the polarity of the G>T versus C>A error bias corresponds to the strand polarity of the capture reagents: NimbleGen's SeqCap targets the (+) strand, while Illumina's Nextera Exome Kit targets the (–) strand (A. Aravanis, Illumina, personal communication).

The second study, by Butler and colleagues, applied the Agilent SureSelect Human All Exon v4 UTR reagent to whole-exome profiling of cfDNA[10]. Interestingly, these data also showed a higher ratio of C>A to G>T errors (**Supplementary Fig. 6e, right**), again suggesting that the (–) strand was being captured by Agilent SureSelect baits, which we later confirmed[11].

Collectively, these data strongly support our hybrid-capture-based model of oxidative damage, highlighting the need for methods, such as background polishing, to address the corresponding sequence artifacts.

Separately, we tested whether independent cfDNA datasets also exhibit stereotypical background errors across multiple genomic locations. By analyzing 5 cfDNA samples from De Mattos-Arruda and colleagues[9], we found widespread recurrent errors for nearly all base substitution types (**Supplementary Fig. 5a, top**). Moreover, these error profiles were comparable between cfDNA obtained from plasma and from CSF (**Supplementary Fig. 5a, top**). Striking similarities in background patterns were also observed when these independent data were intersected with our NSCLC tumor genotyping panel (**Supplementary Fig. 5a, bottom**). Therefore, recurrent background errors in capture-based cfDNA sequence data are likely a general phenomenon – they can be found independently of cfDNA origin, oligonucleotide synthesis batch, and library protocol.

Additional details related to NSCLC selector design. CAPP-Seq selectors are generally designed to cover as many patients and mutations per patient as possible with minimal genomic space (**Fig. 1a**). To prioritize inclusion of genomic regions, we used an approach that leverages a "recurrence index" (RI) metric, defined as the percent of patients (in a given cohort) that harbor mutations (e.g., SNVs/indels) in a given kilobase of genomic sequence. A similar strategy was used previously[2], with exons as the primary genomic unit and without considering indels.

Mutation annotation format (MAF) files were obtained from TCGA whole exome sequencing studies of 606 lung adenocarcinoma[12] (LUAD) tumors (v2.4) and 178 lung squamous cell carcinoma[13] (SCC) tumors (v2.3). MAF files were pre-filtered using UCSC genome browser feature tracks to eliminate variants in (i) repeat-rich genomic regions (RepeatMasker, simple repeats, microsatellites, interrupted repeats and segmental duplications, all downloaded October 19, 2013) and (ii) intervals with low mapping rates or low k-mer uniqueness (wgEncodeCrgMapability 100mer track, wgEncodeDukeMapability 35mer track).

Using filtered MAF data as input, we restricted our search space to known lesions flanked by a user-defined buffer (by default, 1bp), with a minimum tile size of 100bp. Since only a subset of an exon may contain known somatic mutations, this approach saves sequencing space. Selected regions were subsequently ranked by decreasing RI, and those in the top 10 percent of both RI and the number of patients per region were included. This process was then iteratively repeated using relaxed percentile filters (e.g., to permit the top 1/3 regions) and regions that maximally increased the median number of mutations per patient were added. Selector growth terminated when the desired size was reached (175 kb to yield 8 mutations in the average NSCLC patient, **Supplementary Table 1**), or when all genomic regions satisfying these filters were exhausted.

Additional details related to selector-wide genotyping. A schematic of the SNV genotyping approach is provided in **Supplementary Fig. 10b**. Given that base substitution classes have disparate background distributions (**Figs. 2b, 3a**), we sought to control the false positive rate (FPR) for each class separately. Toward that end, we modeled the cumulative distribution of background errors for each base substitution class, excluding candidate variants with >5 supporting reads to minimize the confounding influence of true variants. We found that power series and exponential functions fit the observed data well (**Supplementary Fig. 10a**), and for each class, we selected the function that best captured the data using linear regression in log-linear space. To increase sensitivity, we modeled candidate variants with and without strand support (plus and minus oriented reads) separately, for a total of 24 base substitution

models per input sample (2 × 12 substitution classes). Such models readily illustrate the impact of background polishing on substitution-specific error rates (**Supplementary Fig. 10a**). Each of the 24 functions was independently solved to identify the minimum number of supporting reads *t* needed to yield *y* false positive calls (i.e., cumulative errors). To minimize the FPR, we used $y = 0$ in this work (**Supplementary Fig. 10a**).

To identity SNVs, base substitution thresholds were further adjusted for each candidate variant based on considerations of local error rate *e* and position-specific sequencing depth *d* (**Supplementary Fig. 10b**). Since discrete genomic intervals often exhibit differences in background error rates (e.g., misalignment due to repeat content), we explicitly analyzed the error rate of each gene *e*, defined as the number of positions harboring non-reference bases divided by the number of sequenced bases. If a given gene *g* was within the top 25 percent of selector-wide gene-level error rates, then the base substitution threshold *t* for each candidate variant in *g* was up-weighted:

- $t \leftarrow t \times w$, where $w = \min\{q^2, 5\}$ and $q = e$ divided by the 75th percentile of the error rates of all evaluable genes

Subsequently, if the sequencing depth *d* of a given candidate variant was less than the median selector-wide sequencing depth d^{med}, *t* was down-weighted:

- $t \leftarrow t/w^*$, where $w^* = \ln(d^{med}/d)$

Variants that satisfied *t* were saved as candidate SNVs.

Next, we applied a heuristic filter to detect and remove remaining background alleles within the list of candidate SNVs *L* (provided $|L|$ was ≥4) (**Supplementary Fig. 10b**). Upon ranking *L* by increasing AFs, an iterator *i* was used to traverse the list. For each *i*, *L* was split into two parts, SNVs with an AF below L_i and SNVs with an AF ≥L_i. A two-sided *F*-test was employed to statistically evaluate the difference in variance between the two lists, yielding a p-value. The SNV list *L* was then traversed in order of increasing AFs to identify the index i^* of the first p-value corresponding to a local minimum. Such a minimum, if detected, indicates a potential inflection point between noise (lower tail) and signal (higher AFs). If the p-value corresponding to i^* was below 0.05 and if L_i was at least 10% greater than L_{i-1}, we subsequently evaluated the difference between L_i and the distribution of potential background events, L_1 to L_{i-1}, using a one-sided *z* test (justified given normality observed for SNV AFs). If the corresponding p-value was <0.01, the candidate SNV list was split and the lower tail (L_1 to L_{i-1}) was removed. In empirical analyses, this procedure was found to improve specificity (data not shown), suggesting it can effectively detect and remove residual background variants.

For the analyses in this work, we required a minimum position-specific depth of 20 hGEs for tumors and 1,000 hGEs for cfDNA. To incorporate paired germline samples, we eliminated candidate variant calls if present in paired germline with ≥1% AF, ≥4 supporting reads, and in a position with ≥10x total depth.

To evaluate the technical performance of our approach, we created an *in silico* dilution series in which a control cfDNA sample with median depth of 4,046 hGEs was manipulated to introduce 100 uniformly distributed homozygous SNVs (**Supplementary Table 2**). Each synthetic numerator was then added to the original cfDNA sample in 5% and 0.5% proportions. To emulate the median length of cfDNA, thereby maintaining its distribution in sequencing data, genomic regions were randomly spiked in 170bp contiguous segments. Robust performance was observed (**Supplementary Fig. 10c**). Separately, in comparison to the approach we previously employed for tumor genotyping[2], we found that the adaptive method exhibited higher sensitivity and

specificity for somatic genotyping of tumors, whose variant calls were tested within a ctDNA monitoring framework (same analysis as in **Fig. 5b**; data not shown).

Supplementary Note References

1. Diaconis, P. & Mosteller, F. Methods for Studying Coincidences. in *Selected Papers of Frederick Mosteller* (eds. Fienberg, S. & Hoaglin, D.) 605-622 (Springer New York, 2006).
2. Newman, A.M.*, et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**, 548-554 (2014).
3. Stein, C., Drysdale, R. & Bogart, K. *Discrete Mathematics for Computer Scientists*, (Addison-Wesley Publishing Company, 2010).
4. Kennedy, S.R.*, et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606 (2014).
5. Palfrey, D., Picardo, M. & Hine, A.V. A new randomization assay reveals unexpected elements of sequence bias in model 'randomized' gene libraries: implications for biopanning. *Gene* **251**, 91-99 (2000).
6. Seber, G.A.F. *The Estimation of Animal Abundance: And Related Parameters*, (Macmillan Publishing Company, 1982).
7. Ichinose, T.*, et al.* Liver carcinogenesis and formation of 8-hydroxy-deoxyguanosine in C3H/HeN mice by oxidized dietary oils containing carcinogenic dicarbonyl compounds. *Food Chem Toxicol* **42**, 1795-1803 (2004).
8. Costello, M.*, et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Research* **41**, e67 (2013).
9. De Mattos-Arruda, L.*, et al.* Cerebrospinal fluid-derived circulating tumour DNA better represents the genomic alterations of brain tumours than plasma. *Nat Commun* **6**, 8839 (2015).
10. Butler, T.M.*, et al.* Exome Sequencing of Cell-Free DNA from Metastatic Cancer Patients Identifies Clinically Actionable Mutations Distinct from Primary Disease. *PLoS One* **10**, e0136407 (2015).
11. Gnirke, A.*, et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189 (2009).
12. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550 (2014).
13. Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).