An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage

Aaron M. Newman*, Scott V. Bratman*, Jacqueline To, Jacob F. Wynne, Neville C. W. Eclov, Leslie A. Modlin, Chih Long Liu, Joel W. Neal, Heather A. Wakelee, Robert E. Merritt, Joseph B. Shrager, Billy W. Loo, Jr., Ash A. Alizadeh[#], and Maximilian Diehn[#]

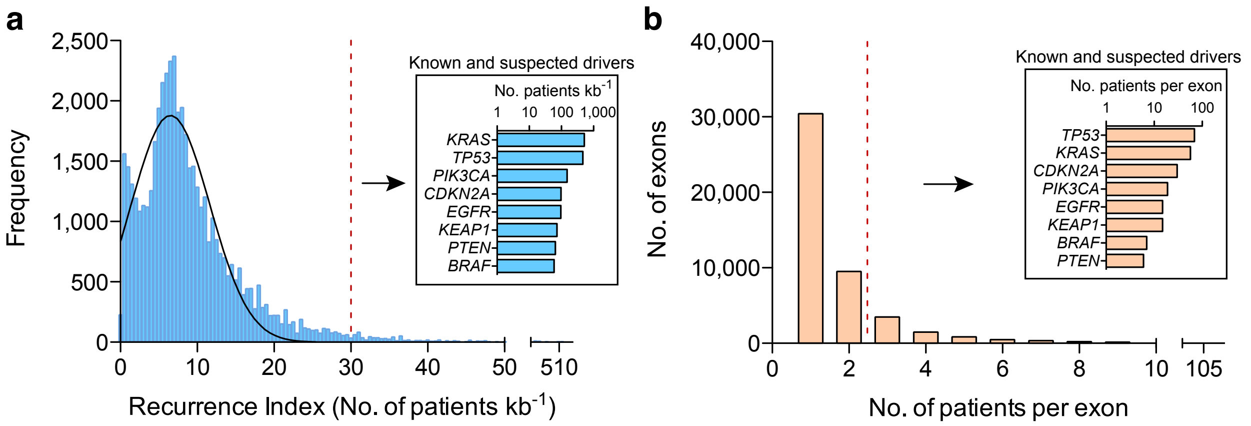
Supplementary Information

Journal: Nature Medicine

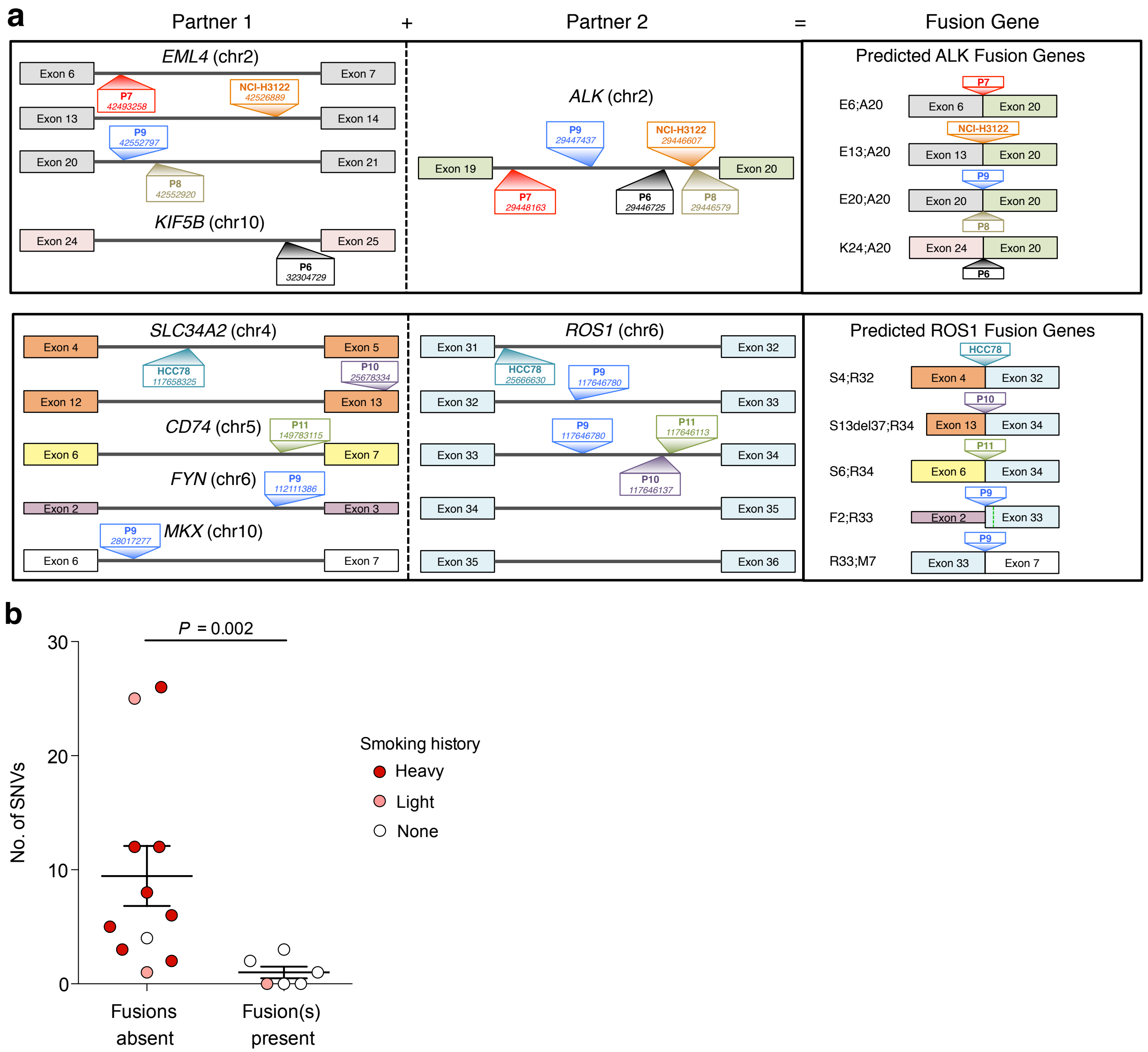
Article Title: **An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage**

Corresponding authors: Maximilian Diehn, MD/PhD, and Ash A. Alizadeh, MD/PhD

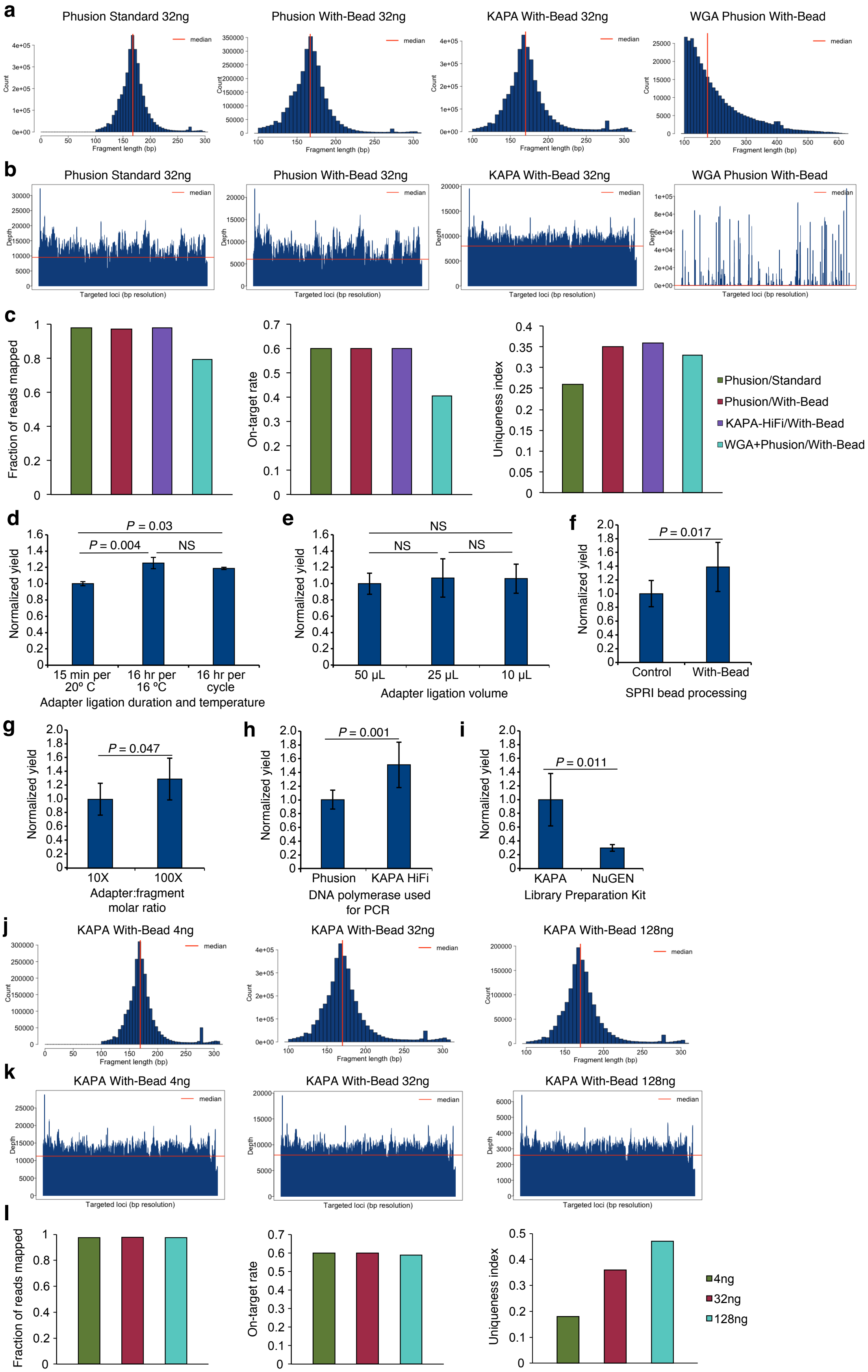
SI Guide



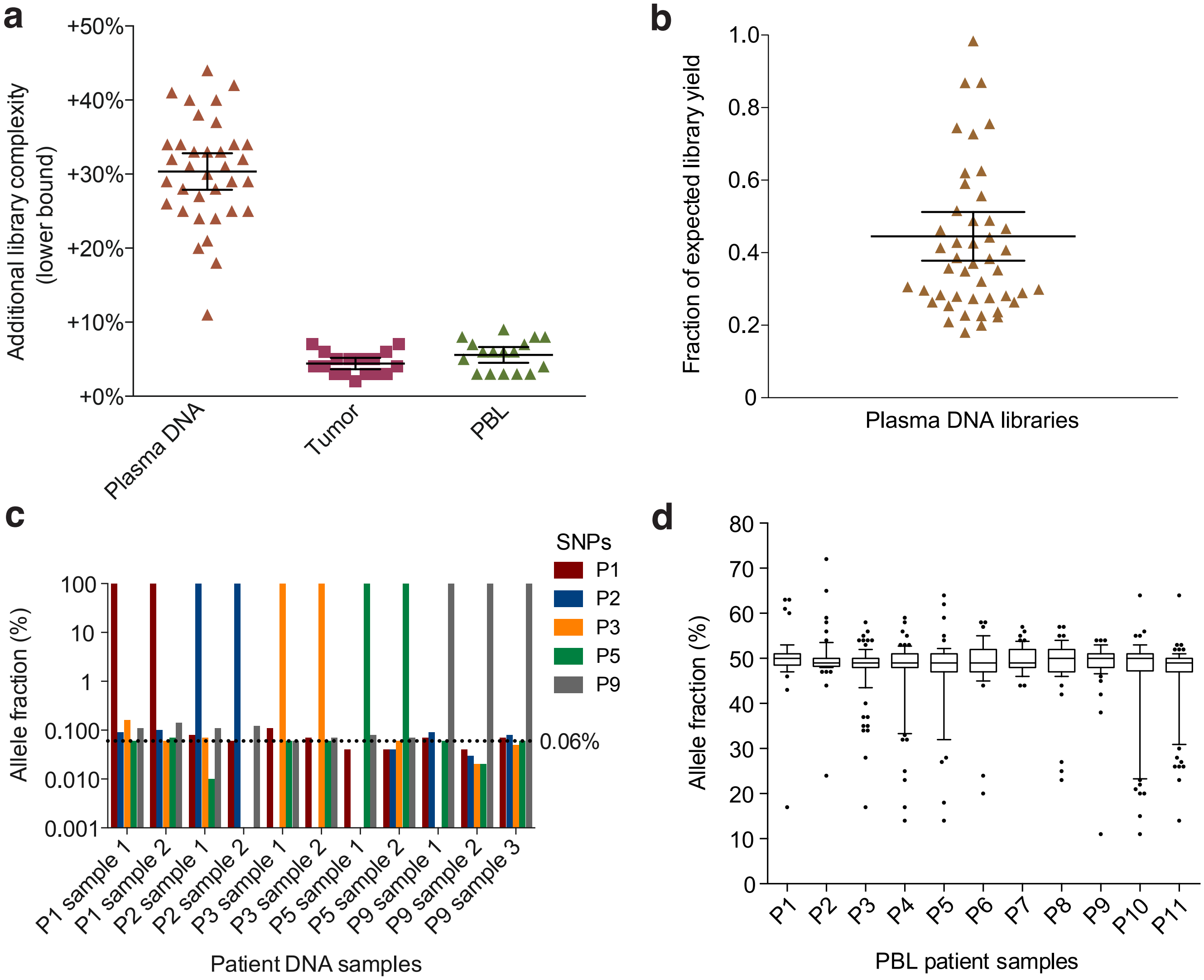
Supplementary Figure 1: Statistical enrichment of recurrently mutated NSCLC exons captures known drivers. We employed two metrics to prioritize exons with recurrent mutations for inclusion in the CAPP-Seq NSCLC selector. The first, termed Recurrence Index (RI), is defined as the number of unique patients (i.e., tumors) with somatic mutations per kilobase of a given genomic unit (here, exon) and the second is defined as the number of unique patients (i.e., tumors) with mutations in a given genomic unit (here, exon). We analyzed exons containing at least one non-silent SNV identified by TCGA (*n* = 47,769) in a combined cohort of 407 lung adenocarcinoma (LUAD) and squamous cell carcinoma (SCC) patients. (**a**) Known and suspected NSCLC drivers are highly enriched at RI ≥30 (inset), comprising 1.8% (*n* = 861) of analyzed exons. (**b**) Known and suspected NSCLC drivers are highly enriched at ≥3 patients with mutations per exon (inset), encompassing 16% of analyzed exons.



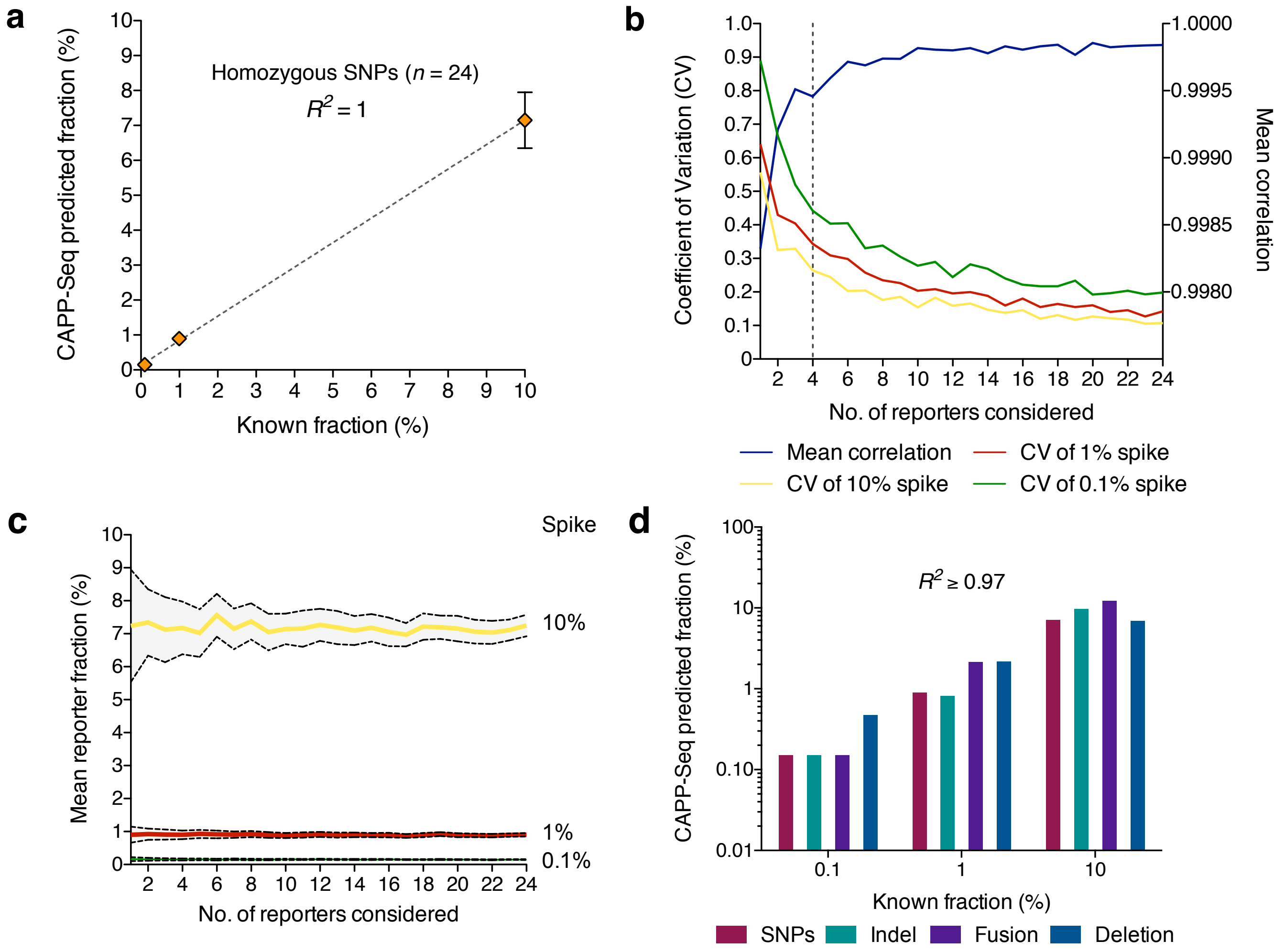
Supplementary Figure 2: Fusion discovery in NSCLC cell lines and tumor samples. (**a**) Base-pair resolution breakpoint mapping for all patients and cell lines enumerated by FACTERA. Gene fusions involving ALK (top) and ROS1 (bottom) are graphically depicted. Schematics in the left and middle panels indicate the exact genomic positions (hg19 NCBI Build 37.1/GRCh37) of the breakpoints in *ALK*, *ROS1*, *EML4*, *KIF5B*, *SLC34A2*, *CD74*, *MKX*, and *FYN*. Right panels depict exons flanking the predicted gene fusions with notation indicating the 5' fusion partner gene and last fused exon followed by the 3' fusion partner gene and first fused exon. For example, in S13del37;R34 exons 1–13 of *SLC34A2* (excluding the 3' 37 nucleotides of exon 13) are fused to exons 34–43 of *ROS1*. Exons in *FYN* are from its 5' UTR and precede the first coding exon. The green dotted line in the predicted *FYN-ROS1* fusion indicates the first in-frame methionine in *ROS1* exon 33, which preserves an open reading frame encoding the ROS1 kinase domain. All rearrangements were each independently confirmed by PCR, FISH, and/or Sanger-sequencing. (**b**) Presence of fusions is inversely related to the number of SNVs detected by CAPP-Seq. For each patient listed in **Table 1** the number of identified SNVs versus the presence (n = 11) or absence (n = 6) of detected genomic fusions is plotted. Statistical significance was determined using a two-sided Wilcoxon rank sum test, and summarized values are presented as means ± s.e.m.



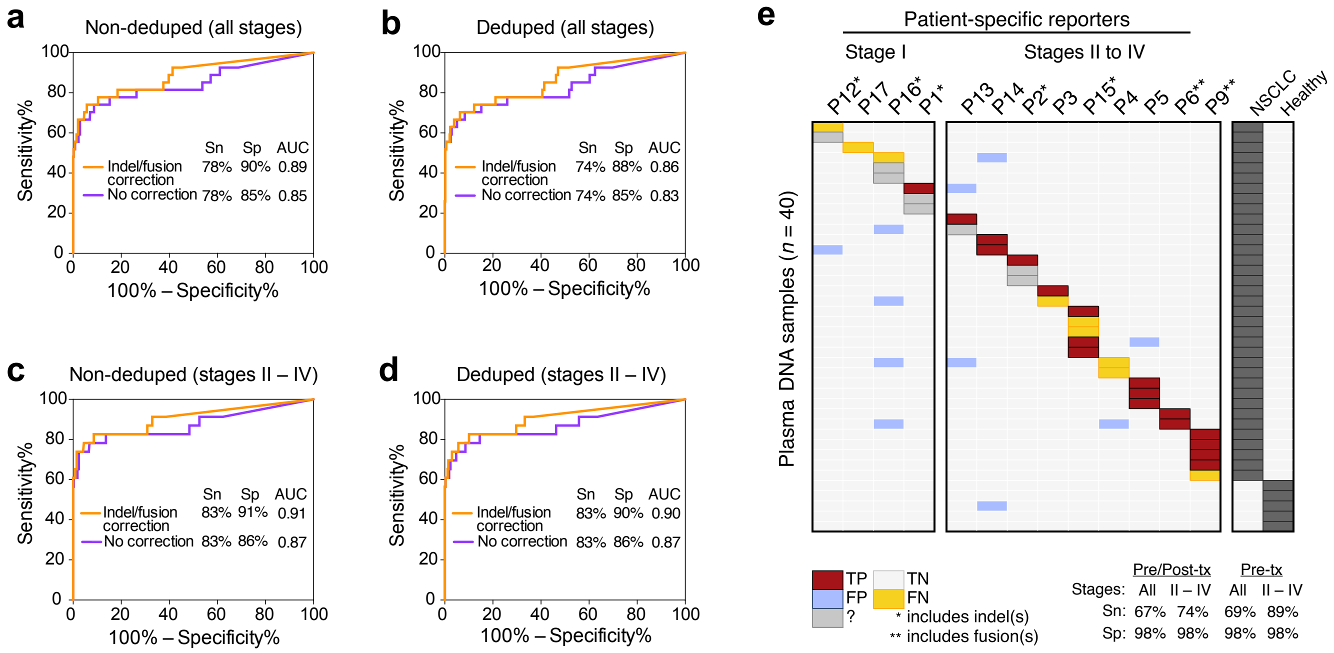
Supplementary Figure 3: Improvements in CAPP-Seq performance with optimized library preparation procedures. (**a–c**) Using 32 ng of nput circulating DNA from plasma, we compared standard versus ‘with bead’[5] library preparation methods, as well as two commercially available DNA polymerases (Phusion and KAPA HiFi). We also compared template pre-amplification by Whole Genome Amplification (WGA) using Degenerate Oligonucleotide PCR (DOP). Indices considered for these comparisons included (**a**) length of the captured circulating DNA fragments sequenced, (**b**) depth and uniformity of sequencing coverage across all genomic regions in the selector, and (**c**) sequence mapping and capture statistics, including uniqueness. Collectively, these comparisons identified KAPA HiFi polymerase and a “with bead” protocol as having most robust and uniform performance. (**d–i**) Optimizing allele recovery from low input circulating DNA during Illumina library preparation. Bars reflect the relative yield of CAPP-Seq libraries constructed from 4 ng circulating DNA, calculated by averaging quantitative PCR measurements of *n* = 4 pre-selected reporters within CAPP-Seq with pre-defined amplification efficiencies. (**d**) Sixteen hour ligation at 16 ºC increases ligation efficiency and reporter recovery. (**e**) Adapter ligation volume did not have a significant effect on ligation efficiency and reporter recovery. (**f**) Performing enzymatic reactions “with-bead” to minimize tube transfer steps increases reporter recovery. (**g**) Increasing adapter concentration during ligation increases ligation efficiency and reporter recovery. Reporter recovery is also higher when using KAPA HiFi DNA polymerase compared to Phusion DNA polymerase (**h**) and when using the KAPA Library Preparation Kit with the modifications in **d** – **g** compared to the NuGEN SP Ovation Ultralow Library System with automation on a Mondrian SP Workstation (**i**). Relative reporter abundance was determined by qPCR using the $2^{-\Delta Ct}$ method. A two-sided *t* test with equal variance was used to test the statistical significance between groups. All values are presented as means ± s.d. N.S., not significant. Based on these results, we estimate that combining the methodological modifications in **d** and **f** – **h** improves yield in NGS libraries by 3.3-fold. (**j**–**l**) CAPP-Seq performance with various amounts of input circulating DNA. (**j**) Length of the captured circulating DNA fragments sequenced. (**k**) Depth of sequencing coverage across all genomic regions in the selector (pre-duplicate removal). (**l**) Sequence mapping and capture statistics. As expected, more input circulating DNA mass correlates with more unique fragments sequenced.



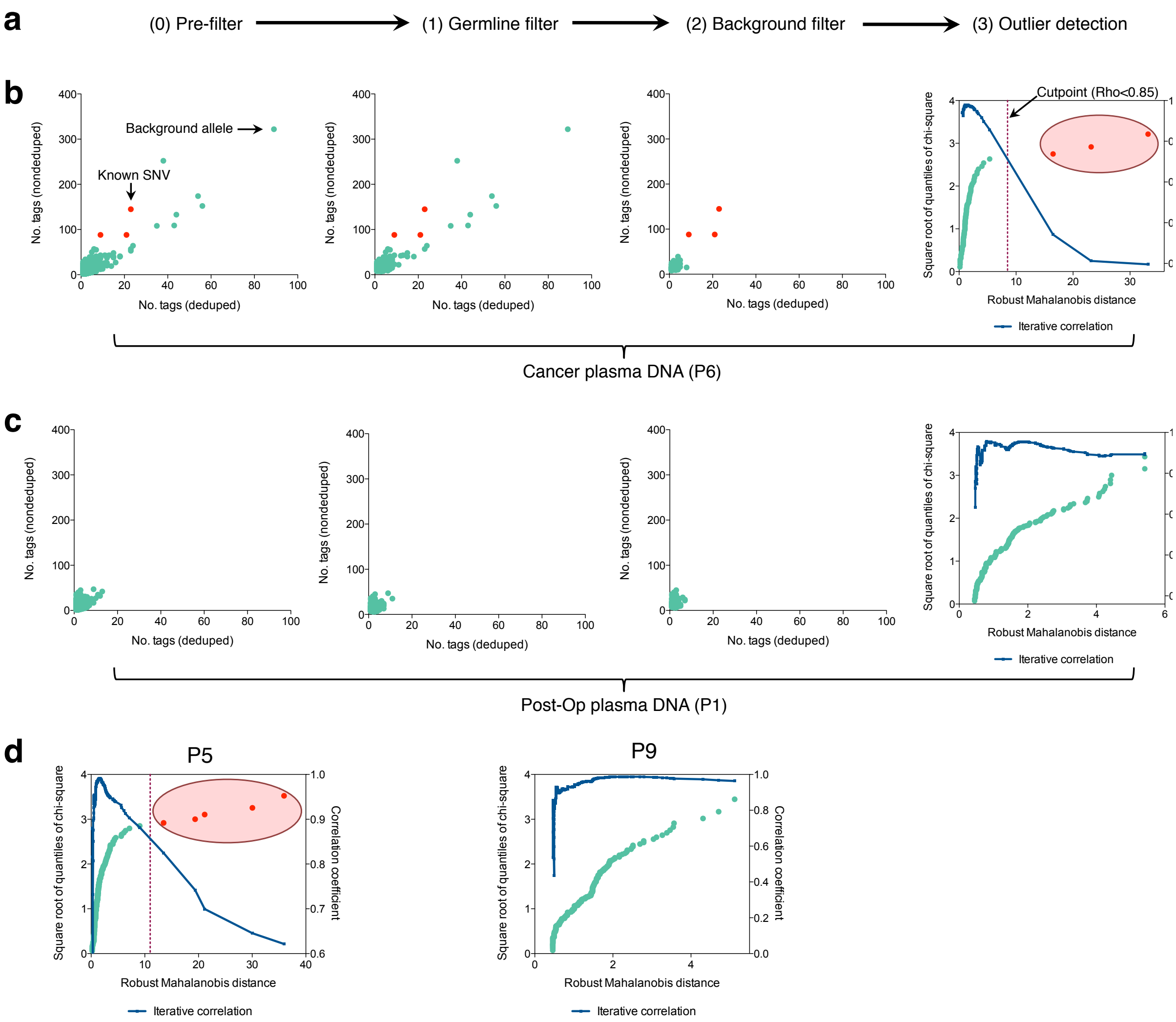
Supplementary Figure 4: Library complexity, molecule recovery, and technical assessment of CAPP-Seq. (**a**) The expected proportion of additional library complexity present in post-duplicate reads is plotted for all patient and control samples, including plasma DNA (n = 40) and paired tumor/PBL specimens (n = 17 each). Because of the highly stereotyped size of circulating DNA fragments occurring naturally in blood plasma, when compared with genomic DNA shorn by sonication, any two fragments of DNA circulating in plasma are inherently more likely by chance to have arisen from different original molecules, whether considering tumor or non-tumor cells as the source of this circulating DNA. To estimate this "missing" complexity, we reasoned that two DNA fragments (i.e., paired end reads) with identical start/end coordinates that differ by a single *a priori* defined germline variant (i.e., one maternal and one paternal allele) represent two unique and independent starting molecules rather than technical artifacts (i.e., PCR duplicates). Therefore, the number of fragments sharing identical start/end coordinates with both maternal and paternal germline alleles of heterozygous SNPs were used to estimate additional library complexity. Library complexity estimates updated to factor in these data are also provided in **Supplementary Table 2** and determined as described in **Supplementary Methods** (section B3.1). (**b**) Empirical assessment of molecule recovery in circulating plasma DNA (n = 40) by determination of the mass of DNA produced compared to the expected library yield based on mass input, number of PCR cycles, and efficiency (mean = 46%). See **Supplementary Methods** for details (section B3.3). (**a**,**b**) Values are presented as means ± 95% confidence intervals. (**c**) Analysis of library cross-contamination. Allelic fractions of patient-specific homozygous germline SNPs were assessed in circulating DNA samples multiplexed on the same lane. SNPs were called as described in **Supplementary Methods** (section B1.5). The mean "cross-contamination" rate in circulating DNA samples was 0.06%, shown by the horizontal dotted line. This level of contamination is too low to affect our estimates of tumor burden given the low fraction of tumor-derived circulating DNA in plasma of NSCLC patients (median of ~0.1%; **Fig. 5a**) (i.e., 0.06 × 0.1 = 0.006% of a given sample would on average represent contamination from ctDNA of another sample). Of note, to minimize the risk of inter-sample contamination, we use aerosol barrier tips, work in hoods, and do not multiplex tumor and plasma libraries in the same lane. (**d**) Analysis of selector-wide bias in captured sequence. Because the NSCLC selector was designed to target the hg19 reference genome, we reasoned that selector bias for SNVs, if any, should be discernable as a systematically lower ratio of non-reference to reference alleles in heterozygous germline SNPs. Therefore, we analyzed high confidence SNPs in patient PBL samples, where high confidence was defined as variants with a non-reference fraction >10% present in the common SNPs subset of dbSNP (version 137.0). As shown, we detected a very small skew toward reference (8 of 11 samples have a median non-reference allelic frequency of 49%; the remaining 3 samples are unbiased). Importantly, such bias appears too small to significantly affect our results. Boxes represent the interquartile range, and whiskers encapsulate the 10th to 90th percentiles. Germline SNPs were identified using VarScan 2, as described in **Supplementary Methods** (section B1.6).



Supplementary Figure 5: Empirical spiking analysis of CAPP-Seq using two NSCLC cell lines. (**a**) Expected and observed (by CAPP-Seq) fractions of NCI-H3122 DNA spiked into control HCC78 DNA are linear for all fractions tested (0.1%, 1%, and 10%; R^2 = 1). (**b**) Using data from **a**, analysis of the effect of the number of SNVs considered on the estimates of fractional abundance (95% confidence intervals shown in gray). (**c**) Analysis of the effect of the number of SNVs considered on the mean correlation coefficient and coefficient of variation between expected and observed cancer fractions (blue dashed line) using data from panel **a**. (**d**) Expected and observed fractions of the *EML4-ALK* fusion present in HCC78 are linear (R^2 = 0.995) over all spiking concentrations tested (see **Supplementary Fig. 2a** for breakpoints). The observed EML4-ALK fractions were normalized based on the relative abundance of the fusion in 100% H3122 DNA (see **Supplementary Methods** for details). Moreover, both a single heterozygous insertion (‘Indel’; chr7: 107416855, +T) and a 4.9 kb homozygous deletion (‘Deletion’, chr17: 29422259–29592392) in NCI-H3122 were concordant with defined concentrations. Values in **a** are presented as means ± s.e.m.



Supplementary Figure 6: Extended CAPP-Seq sensitivity and specificity analysis. (**a-d**) Receiver Operating Curve (ROC) analysis, comparing sensitivity and specificity achieved for (i) non-deduped (panels **a** and **c**) versus deduped (post PCR duplicate removal) data (panels **b** and **d**), (ii) all disease stages (panels **a** and **b**) versus intermediate to advanced disease stages (stages II-IV, panels **c** and **d**), and (iii) with or without the indel/fusion filter (described in **Supplementary Methods**). Reporter fractions for both non-deduped and deduped circulating DNA samples are provided in **Supplementary Table 4**. (**e**) CAPP-Seq sensitivity and specificity over all patient reporters and sequenced plasma DNA samples. All values shown reflect a *ctDNA detection index* of 0.03. See **Online Methods** and **Supplementary Methods** for details on detection metrics, and determination of cancer-positive, cancer-negative, and unknown categories.



Supplementary Figure 7: Biopsy-free cancer screening with CAPP-Seq. (**a**) Steps to identify candidate SNVs in plasma DNA demonstrated using a patient sample with NSCLC (P6, see **Supplementary Table 3**). Following stepwise filtration, outlier detection is applied (**Supplementary Methods**). (**b**) Screening method applied to a cancer-positive pre-treatment sample (patient P6). (**c**) Same as **b**, but using a post-treatment sample from a patient who had their tumor surgically removed. No SNVs are identified, as expected. (**d**) Two additional representative samples applying retrospective screening to patients analyzed in this study. P5 has confirmed tumor-derived SNVs, while P9 is cancer positive but lacks tumor-derived SNVs. Red points, confirmed tumor-derived SNVs; Turquoise points, background.

Cost comparison for equal detection limit in plasma

Feature	CAPP-Seq	Whole exome	Whole genome
Bases covered	0.125 Mb	50 Mb	3,000 Mb
No. mutations	4	218	15,659
Depth	10,000x	1,100x	120x
Gbs sequenced	2.1	92	360
Relative cost	**1x**	**44x**	**171x**

Supplementary Figure 8: Relative cost of CAPP-Seq compared to other methods**.** Estimates are based on achieving the same theoretical detection limit as CAPP-Seq (shown as a red curve in **Fig. 1d**). All data relate to **Fig. 1d**. Calculations are detailed in **Online Methods**.

Supplementary Methods

A. Molecular Biology Methods

A1. Cell Lines

The lung adenocarcinoma cell lines NCI-H3122 and HCC78 were obtained from ATCC and DSMZ, respectively, and grown in RPMI 1640 with L-glutamine (Gibco) supplemented with 10% fetal bovine serum (Gembio) and 1% penicillin/streptomycin cocktail. Cells were maintained in mid-log-phase growth in a 37 ºC incubator with 5% CO_2. Genomic DNA was purified from freshly harvested cells with the DNeasy Blood & Tissue Kit (Qiagen).

A2. Pleural Fluid Processing and Flow Cytometry, and Cell Sorting

Cells from pleural fluid from patients P9 and P6 were harvested by centrifugation at 300 × g for 5 min at 4 ºC and washed in FACS staining buffer (HBSS + 2% heat-inactivated calf serum [HICS]). Red blood cells were lysed with ACK Lysing Buffer (Invitrogen), and clumps were removed by passing through a 100 µm nylon filter. Filtered cells were spun down and resuspended in staining buffer. While on ice, the cell suspension was blocked for 20 min with 10 µg mL^{-1} rat IgG and then stained for 20 min with APC-conjugated mouse anti-human EpCAM (BioLegend, clone 9C4), PerCP-Cy5.5-conjugated mouse anti-human CD45 (eBioscience, clone 2D1), and PerCP-eFluor710-conjugated mouse anti-human CD31 (eBioscience, clone WM59). After staining, cells were washed and resuspended with staining buffer containing 1 µg mL^{-1} DAPI, analyzed, and sorted with a FACSAria II cell sorter (BD Biosciences). Cell doublets and DAPI-positive cells were excluded from analysis and sorting. $CD31^{-}CD45^{-}EpCAM^{+}$ cells were sorted into staining buffer, spun down, and flash frozen in liquid nitrogen. DNA was isolated with the QIAamp DNA Micro Kit (Qiagen).

A3. Optimization of NGS Library Preparation from Low Input Circulating DNA

Protocols for Illumina library construction were compared in a step-wise manner with the goal of (1) optimizing adapter ligation efficiency, (2) reducing the necessary number of PCR cycles following adapter ligation, (3) preserving the naturally occurring size distribution of circulating DNA fragments, and (4) minimizing variability in depth of sequencing coverage across all captured genomic regions. Initial optimization was done with NEBNext DNA Library Prep Reagent Set for Illumina (New England BioLabs), which includes reagents for end-repair of the circulating DNA fragments, A-tailing, adapter ligation, and amplification of ligated fragments with Phusion High-Fidelity PCR Master Mix. Input was 4 ng circulating DNA (obtained from plasma of the same healthy volunteer) for all conditions. Relative allelic abundance in the constructed libraries was assessed by qPCR of 4 genomic loci (Roche NimbleGen: NSC-0237, NSC-0247, NSC-0268, and NSC-0272) and compared by the $2^{-\Delta Ct}$ method.

Ligations were performed at 20 ºC for 15 min (as per the manufacturer's protocol), at 16 ºC for 16 h, or with temperature cycling for 16 h as previously described[2]. Ligation volumes were varied from the standard (50 µL) down to 10 µL while maintaining a constant concentration of DNA ligase, circulating DNA fragments, and Illumina adapters. Subsequent optimizations incorporated ligation at 16 ºC for 16 h in 50 µL reaction volumes.

Next, we compared standard SPRI bead processing procedures, in which new AMPure XP beads are added after each enzymatic reaction and DNA is eluted from the beads for the next reaction, to with-bead protocol modifications as previously described[3]. We compared 2 concentrations of Illumina adapters in the ligation reaction: 12 nM (10-fold molar excess to circulating DNA fragments) and 120 nM (100-fold molar excess).

Using the optimized library preparation procedures, we next compared the NEBNext DNA Library Prep Reagent Set (with Phusion DNA Polymerase) to the KAPA Library Preparation Kit (with KAPA HiFi DNA Polymerase). The KAPA Library Preparation Kit with our modifications was also compared to the NuGEN SP Ovation Ultralow Library System with automation on Mondrian SP Workstation.

A4. Evaluation of Library Preparation Modifications on CAPP-Seq Performance

We performed CAPP-Seq on 32 ng circulating DNA using standard library preparation procedures with the NEBNext kit, or with optimized procedures using either the NEBNext kit or the KAPA Library Preparation Kit. In parallel we performed CAPP-Seq on 4 ng and 128 ng circulating DNA using the KAPA kit with our optimized procedures. Indexed libraries were constructed, and hybrid selection was performed in multiplex. The post-capture multiplexed libraries were amplified with Illumina backbone primers for 14 cycles of PCR and then sequenced on a paired-end 100 bp lane of an Illumina HiSeq 2000.

We also evaluated CAPP-Seq on ultralow input following whole genome amplification (WGA). We used SeqPlex DNA Amplification Kit (Sigma-Aldrich), which employs degenerate oligonucleotide primer PCR. Briefly, 1 ng circulating DNA was amplified with real-time monitoring with SYBR Green I (Sigma-Aldrich) on a HT7900 Real Time PCR machine (Applied Biosystems). Amplification was terminated after 17 cycles yielding 2.8 µg DNA. The primer removal step yielded ~600 ng DNA, and this entire amount was used for library preparation using the NEBNext kit with optimized procedures as described above.

A5. Validation of Variants Detected by CAPP-Seq

All structural rearrangements and a subset of tumoral SNVs detected by CAPP-Seq were independently confirmed by qPCR and/or Sanger sequencing of amplified fragments. For HCC78, a 120 bp fragment containing the *SLC34A2-ROS1* breakpoint was amplified from genomic DNA using the primers: 5'-AGACGGGAGAAAATAGCACC-3' and 5'-ACCAAGGGTTGCAGAAATCC-3'. For NCI-H3122, a 143 bp fragment containing the *EML4-ALK* breakpoint was amplified using the primers: 5'-GAGATGGAGTTTCACTCTTGTTGC-3' and 5'-GAACCTTTCCATCATACTTAGAAATAC-3'. 5ng genomic DNA was used as template with 250 nM oligos and 1X Phusion PCR Master Mix (NEB) in 50 µL reactions. Products were resolved on 2.5% agarose gel and bands of the expected size were removed. The amplified DNA fragments were purified using the Qiaquick Gel Extraction Kit (Qiagen) and submitted for Sanger sequencing (Elim Biopharm). For P9, genomic DNA breakpoints were confirmed by qPCR using the primers: 5'-TCCATGGAAGCCAGAAC-3' and 5'-ATGCTAAGATGTGTCTGTCA-3' for *EML4-ALK*; 5'-CCTTAACACAGATGGCTCTTGATGC-3' and 5'-TCCTCTTTCCACCTTGGCTTTCC-3' for *ROS1-MKX*; and 5'-GGTTCAGAACTACCAATAACAAG-3' and 5'-ACCTGATGTGTGACCTGATTGATG-3' for *FYN-ROS1*. For qPCR, 10 ng of pre-amplified genomic DNA was used as template with 250 nM oligos and 1X Power SyberGreen Master Mix in 10µL reactions performed in triplicate on a HT7900 Real Time PCR machine (Applied Biosystems). Standard PCR thermal cycling parameters were used. Amplification of amplicons spanning all 3 breakpoints detected in P9 were confirmed in tumor genomic DNA as well as plasma circulating DNA, and PBL genomic DNA was used as a negative control.

CAPP-Seq confirmed somatic tumor mutations (SNVs and rearrangements) that were detected by clinical assays as a part of standard clinical care (**Supplementary Table 3**). Clinical mutation assays were performed on formalin-fixed paraffin-embedded tissues. SNVs were detected by the SNaPshot assay[4]. Rearrangements were detected by fluorescence in situ hybridization (FISH) using separation probes targeting the ALK locus (Abbott) or ROS1 locus (Cytocell).

B. Bioinformatics and Statistical Methods

B1. CAPP-Seq Detection Threshold Metrics

B1.1. Selector base-level background

We assessed the base-level background distribution of the NSCLC selector (**Fig. 2d**) using all 40 plasma DNA samples collected from NSCLC and healthy individuals analyzed in this work (**Supplementary Table 2**). For each *background* base in selector positions having ≥500x overall sequencing depth, the outlier-corrected mean across all circulating DNA samples was calculated. Although we tested dedicated outlier detection methods, such as iterative Grubbs' method and ROUT[5], our empirical analyses indicated that simple removal of the minimum and maximum values worked best. Importantly, to restrict our analysis to *background* bases, each patient sample was pre-filtered to remove germline, loss of heterozygosity (LOH), and/or somatic variant calls made by VarScan 2[6] (somatic p-value = 0.01; otherwise, default parameters).

B1.2. Significance of SNVs as reporters

To evaluate the significance of tumor-derived SNVs in plasma, we implemented a strategy that integrates circulating DNA fractions across all somatic SNVs, performs a position-specific background adjustment, and evaluates statistical significance by Monte Carlo sampling of background alleles across the selector. We note that this approach differs fundamentally from previous methods, where mutations are interrogated individually. Unlike these methods, our strategy dampens the impact of stochastic noise and biological variables (e.g., mutations near the detection limit, or tumor evolution) on tumor burden quantitation, permitting a more robust statistical assessment. In particular, this allows CAPP-Seq to quantitate low levels of ctDNA with potentially high rates of allelic drop out.

For a given plasma DNA sample θ, we begin by adjusting the allelic fraction f for each of n SNVs from patient P in order to minimize the influence of selector technical/biological background on significance estimates. Specifically, for each allele, we perform the following simple operation, $f^* = max\{0, f - (e - \mu)\}$, where f is the raw allelic fraction in plasma DNA, e is the position-specific error rate for the given allele across all circulating DNA samples (see section B1.1 above), and μ denotes the mean selector-wide background rate (=0.006% in this study, see section B1.1 and **Fig. 2d**). In effect, this adjustment nudges the mean of all n SNVs closer to the global selector mean μ, mitigating the confounding impact of technical/biological background. Using Monte Carlo simulation, we compare the adjusted mean SNV fraction F^* (=$(\sum f^*)/n$) against the null distribution of background alleles across the selector. Specifically, for each of i iterations (=10,000 in this work), n background alleles are randomly sampled from θ, after which their fractions are adjusted using the above formula and averaged. A SNV p-value for patient P is determined as the percentile of F^* with respect to the null distribution of background alleles in θ. Thus, a panel of SNVs from patient P would be assigned a detection p-value of 0.04 if F^* ranks in the 96th percentile of adjusted background alleles in θ. We note that background adjustment always improved CAPP-Seq specificity in our ROC analyses (data not shown).

B1.3. Significance of indels as reporters

We implemented an approach based on population statistics to assess the significance of indels separately from SNVs. For each indel in patient P, we use the Z-test to compare its fraction in a given plasma DNA sample θ against its fraction in every plasma DNA sample in our cohort (excluding circulating DNA samples from the same patient P). To increase statistical robustness, each read strand (positive or negative orientation) is assessed separately, yielding two Z-scores for each indel. These are combined into a single Z-score by Stouffer's method[8], an unweighted approach for integrative Z statistics. Finally, if patient P has more than 1 indel, all indel-specific Z-scores are combined by Stouffer's method into a final Z statistic, which is trivially converted to a p-value.

B1.4. Significance of fusions as reporters

Given the exceedingly low false positive rate associated with the detection of the same NSCLC fusion breakpoint in independent libraries, the recovery of a tumor-derived genomic fusion in plasma DNA by CAPP-Seq was (arbitrarily) assigned a p-value of ~0.

B1.5. Indel/fusion correction for sensitivity and specificity assessment

Related to **Figure 3**, after calculating a ctDNA detection index for every set of reporters across all plasma DNA samples using the methods described above, we applied an additional step to increase specificity. Namely, to exploit the lower technical background of indels and fusion breakpoints as compared to SNVs, we applied an "*indel/fusion correction*". Specifically, if indel/fusion reporters found in patient *X*'s tumor could be uniquely detected in patient X's plasma DNA (i.e., not detected in any other patient or control plasma DNA sample), then the ctDNA detection index corresponding to patient *X* was set to 1 (i.e., ctDNA not detectable) in every unmatched plasma DNA sample. In other words, patient X's reporters would not be called a false positive in another patient. Although we have not yet encountered two patients with the same indel/fusion reporter(s), if this was the case, the correction would not be applied from one patient to the other.

To perform this correction in a blinded manner, as shown for **Figure 3** (panels a and b), we identified germline SNPs in each plasma DNA and PBL sample, and assigned each circulating DNA sample to the tumor/normal pair with highest SNP concordance (after un-blinding, all plasma DNA samples were found to be correctly matched to their corresponding tumor/normal pairs). As shown in **Supplementary Figure 14**, this correction consistently increased CAPP-Seq specificity. Germline SNPs were identified using VarScan 2[6], with a p-value threshold of 0.01, minimum sequence coverage of 100x, a minimum average quality score of 30 (Phred), and otherwise default parameters.

B2. Sensitivity and Specificity Analysis

We tested CAPP-Seq performance in a blinded fashion by masking all patient identifying information, including disease stage, circulating DNA time point, treatment, etc. We then tested our detection metrics described above (section B1) for correctly calling tumor burden across the entire grid of de-identified plasma DNA samples (13 patient-specific sets of somatic reporters across 40 plasma samples, or 520 pairs). To calculate sensitivity and specificity, we "un-blinded" ourselves and grouped patient samples into cancer-positive (i.e. cancer was present in the patient's body), cancer-negative (i.e. patient was cured), or cancer-unknown (i.e. insufficient data to determine true classification) categories. We considered every time point of patients with radiographic evidence of recurrence and all stage IV patients as cancer-positive, regardless of clinical evaluation at the time point in question. The post-treatment time point of patient 13 (P13; stage IIB NSCLC) was considered cancer-unknown due to "No Evidence of Disease (NED)" status at last follow-up, nearly 2 years from their treatment (**Fig. 4e**). Patient 2 (P2; stage IIIB NSCLC), was classified as NED following complete surgical resections, and was also considered cancer-unknown. All post-treatment stage I NSCLC patient samples were conservatively considered "cancer unknown" rather than true negatives due to limited follow-up.

B3. Analysis of Library Complexity

B3.1. Library complexity estimation

We estimated the number of haploid genome equivalents per library using 330 genome equivalents per 1 ng of input DNA (**Supplementary Table 2**), and calculated overall 'molecule recovery' as the median depth after duplicate removal (see section B3.2) divided by the smaller of (i) the median depth before duplicate removal and (ii) the estimated number of haploid genome equivalents. Molecule recovery at a given sequencing depth was estimated to be 38% for circulating plasma DNA, 37% for tumor DNA, and 48% for PBLs (highest DNA input mass among all samples).

In contrast to genomic DNA, plasma DNA is naturally fragmented and has a highly stereotyped size distribution related to nucleosome spacing[9], with a median length of ~170bp and very low dispersion (**Fig. 2a, Supplemenary Table 2**). As such, we hypothesized that independent input molecules with identical start/end coordinates may inflate the duplication rate of circulating DNA, leading to an underestimated molecule recovery rate.

We tested this hypothesis by analyzing heterozygous germline SNPs, reasoning that DNA fragments (i.e., paired end reads) with identical start/end coordinates and differing by a single *a priori* defined germline variant are more likely to represent independent starting molecules than technical artifacts (i.e., PCR duplicates). Heterozygous SNPs were identified in all 90 samples (**Supplementary Table 2**) using VarScan 2[6] (same as described in section B1.5), and filtered for variants with an allele frequency between 40% and 60% that are present in the Common SNPs subset of dbSNP (version 137.0). For each heterozygous common SNP, A/B, we counted all fragments with unique start/end coordinates that support A, B, or AB. Among molecules with a given A/B SNP, there is a 50% chance of getting A and B together when randomly sampling two molecules (AB or BA), and there is a combined 50% chance of getting either AA or BB. Since the number of unique start/end positions for AB (denoted N) represents at least twice as many molecules ($\geq 2N$), and a combined $\geq 2N$ molecules can be assumed missing from unique start/end coordinates that support A or B, a lower bound on total missing library complexity is determined by the formula, $3N/S$, where S denotes the sum of unique start/end coordinates covering A, B, and AB. Across SNPs in each input sample, we calculated an average of 30% missing library complexity in circulating DNA samples, and 4% and 6% missing library complexity in tumor and PBL genomic DNA, respectively (**Supplementary Fig. 4a**). Molecule recovery rates adjusted for estimated loss of complexity are provided in **Supplementary Table 2**, and indicate a mean molecule recovery of at least 49% in circulating plasma DNA, 37% in tumor genomic DNA (mostly FFPE) and 51% in PBL genomic DNA.

B3.2. Duplication rate

Common deduping tools, such as SAMtools[7] rmdup and Picard tools MarkDuplicates (http://picard.sourceforge.net), identify and/or collapse reads based on sequence coordinates and quality, not sequence composition. This can result in the removal of tumor-derived reads (representing distinct molecules) that happen to share sequence coordinates with germline reads. This is particularly problematic for circulating DNA since for a large fraction of molecules there are other unique molecules with the same start and end (see above, section B3.1). To address this issue, we developed a custom Perl script that ignores bases with low quality (here, Phred Q<30), and collapses only those fragments (read pairs) with 100% sequence identity that also share genomic coordinates. The resulting post-duplicate reads are provided alongside corresponding non-deduped data in **Supplementary Tables 2 and 4**, which respectively cover sequencing statistics and plasma DNA monitoring results.

B3.3. Library complexity measured via PCR and mass input

As a separate estimation of library complexity, for each Illumina NGS library constructed from circulating DNA, we calculated the fraction of expected library yield from the actual yield and the expected (ideal) yield (**Supplementary Fig. 4b**). The actual library yield was determined from the molarity and volume of the constructed libraries (prior to hybrid selection). The expected library yield was calculated from the mass of circulating plasma DNA used for library preparation and the number of PCR cycles performed, with the assumption that ligation was 100% efficient and PCR was 95% efficient at each cycle. A PCR efficiency of 95% was observed from qPCR performed on serial dilutions of Illumina TruSeq libraries (mean R^2>0.999 from 4 independent experiments).

B4. CAPP-Seq Selector Design

Most human cancers are relatively heterogeneous for somatic mutations in individual genes. Specifically, in most human tumors, recurrent somatic alterations of single genes account for a minority of patients, and only a minority of tumor types can be defined using a small number of recurrent mutations (<5-10) at predefined positions. Therefore, the design of the selector is vital to the CAPP-Seq method because (1) it dictates which mutations can be detected with high probability for a patient with a given cancer, and (2) the selector size (in kb) directly impacts the cost and depth of sequence coverage. For example, the hybrid selection libraries available in current whole exome capture kits range from 51-71 Mb, providing ~40-60 fold maximum theoretical enrichment versus whole genome sequencing. The degree of potential enrichment is inversely proportional to the selector size such that for a ~100 kb selector, >10,000 fold enrichment should be achievable.

We employed a six-phase design strategy to identify and prioritize genomic regions for the CAPP-Seq NSCLC selector as detailed below. Three phases were used to incorporate known and suspected NSCLC driver genes, as well as genomic regions known to participate in clinically actionable fusions (phases 1, 5, 6), while another three phases employed an algorithmic approach to maximize both the number of patients covered and SNVs per patient (phases 2–4). The latter relied upon a metric that we termed “Recurrence Index” (RI), defined as the number of NSCLC patients with SNVs that occur within a given kilobase of exonic sequence (i.e., No. of patients with mutations / exon length in kb). RI thus serves to measure patient-level recurrence frequency at the exon level, while simultaneously normalizing for gene or exon size. As a source of somatic mutation data uniformly genotyped across a large cohort of patients, in phases 2–4, we analyzed non-silent SNVs identified in TCGA whole exome sequencing data from 178 patients in the Lung Squamous Cell Carcinoma dataset (SCC)[10] and from 229 patients in the Lung Adenocarcinoma (LUAD) datasets (TCGA query date was March 13, 2012). Thresholds for each metric (i.e. RI and patients per exon) were selected to statistically enrich for known/suspected drivers in SCC and LUAD data (**Supplementary Fig. 1**). RefSeq exon coordinates (hg19) were obtained via the UCSC Table Browser (query date was April 11, 2012).

The following algorithm was used to design the CAPP-Seq selector (parenthetical descriptions match design phases noted in **Fig. 1b**).

- Phase 1 (*Known drivers*)
  Initial seed genes were chosen based on their frequency of mutation in NSCLCs. Analysis of COSMIC (v57)[11] identified known driver genes that are recurrently mutated in ≥9% of NSCLC (denominator ≥500 cases). Specific exons from these genes were selected based on the pattern of SNVs previously identified in NSCLC. The seed list also included single exons from genes with recurrent mutations that occurred at low frequency but had strong evidence for being driver mutations, such as BRAF exon 15, which harbors V600E mutations in <2% of NSCLC[12-21].
- Phase 2 (*Max. coverage*)
  For each exon with SNVs covering **≥5** patients in LUAD and SCC, we selected the exon with **highest RI** that identified at least 1 new patient when compared to the prior phase. Among exons with equally high RI, we added the exon with minimum overlap among patients already captured by the selector. This was repeated until no further exons met these criteria.
- Phase 3 (*RI ≥ 30*)
  For each remaining exon with an **RI ≥ 30** and with SNVs covering **≥3** patients in LUAD and SCC, we identified the exon that would result in the largest reduction in patients with only 1 SNV. To break ties among equally best exons, the exon with highest RI was chosen. This was repeated until no additional exons satisfied these criteria.
- Phase 4 (*RI ≥ 20*)
  Same procedure as phase 3, but using **RI ≥ 20**.
- Phase 5 (*Predicted drivers*)
  We included all exons from additional genes previously predicted to harbor driver mutations in NSCLC[12,13].
- Phase 6 (*Add fusions*)
  For recurrent rearrangements in NSCLC involving the receptor tyrosine kinases *ALK*, *ROS1*, and *RET*, the introns most frequently implicated in the fusion event and the flanking exons were included.

All exons included in the selector, along with their corresponding HUGO gene symbols and genomic coordinates, as well as patient statistics for NSCLC and a variety of other cancers, are provided in **Supplementary Table 1**, organized by selector design phase.

C. CAPP-Seq Computational Pipeline

C1. Mutation Discovery: SNVs/indels

For detection of somatic SNV and insertion/deletion events, we employed VarScan 2[6] (somatic p-value = 0.01, minimum variant frequency = 5%, strand filter = true, and otherwise default parameters). Somatic variant calls (SNV or indel) present at less than 0.5% mutant allelic frequency in the paired normal sample (PBLs), but in a position with at least 1000x overall depth in PBLs and 100x depth in the tumor, and with at least 1x read depth on each strand, were retained (**Supplementary Table 3**). While the selector was designed to predominantly capture exons, in practice, it also captures limited sequence content flanking each targeted region. For instance, this phenomenon is the basis for the (thus far) uniformly successful recovery by CAPP-Seq of fusion partners (which are not included within the selector) for kinase genes such as *ALK* and *ROS1* recurrently rearranged in NSCLC. As such, we also considered variant calls detected within 500bps of defined selector coordinates. These calls were eliminated if present in non-coding *repeat* regions, since repeats may confound mapping accuracy. Repeat sequence coordinates were obtained using the RepeatMasker track in the UCSC table browser (hg19). Given a low, but measurable cross-contamination rate of ~0.06% in multiplexed circulating DNA samples, (**Supplementary Fig. 4c**) we also excluded any SNVs found as germline SNPs in samples from the same lane. Additionally, we excluded SNVs in the top 99.9th percentile of global selector background (>0.27% sample-wide background rate; see **Fig. 2d** and section B1.1 above). Finally, we excluded any SNVs not present at a depth of at least 500x in at least 1 circulating DNA sample. Variant annotation was automatically downloaded from the SeattleSeq Annotation 137 web server (http://snp.gs.washington.edu/SeattleSeqAnnotation137/). Complete details for all identified SNVs and indels are provided in **Supplementary Table 3**. Of note, all depth thresholds refer to pre-duplication removal reads (see section B3.2).

C2. Mutation Discovery: Fusions

For practical and robust de novo enumeration of genomic fusion events and breakpoints from paired-end next-generation sequencing data, we developed a novel heuristic approach, termed FACTERA (FACile Translocation Enumeration and Recovery Algorithm). FACTERA has minimal external dependencies, works directly on a preexisting .bam alignment file, and produces easily interpretable output. Additional aspects of the method will be described in a forthcoming manuscript (Newman et al., in preparation). FACTERA is coded in Perl and freely available from the authors upon request.

As input, FACTERA requires a .bam alignment file of paired-end reads produced by BWA[22], exon coordinates in .bed format (e.g., hg19 RefSeq coordinates), and a .2bit reference genome to enable fast sequence retrieval (e.g., hg19). In addition, the analysis can be optionally restricted to reads that overlap particular genomic regions (.bed file), such as the CAPP-Seq selector used in this work.

FACTERA processes the input in three sequential phases: identification of discordant reads, detection of breakpoints at base pair-resolution, and *in silico* validation of candidate fusions. Each phase is described in detail below.

C2.1. Identification of discordant reads

To iteratively reduce the sequence space for gene fusion identification, FACTERA, like other algorithms (e.g. BreakDancer[23]), identifies and classifies discordant read pairs. Such reads indicate a nearby fusion event since they either map to different chromosomes or are separated by an unexpectedly large insert size (i.e. total fragment length), as determined by the BWA mapping algorithm. The bitwise flag accompanying each aligned read encodes a variety of mapping characteristics (e.g., improperly paired, unmapped, wrong orientation, etc.) and is leveraged to rapidly filter the input for discordant pairs. The closest exon of each discordant read is subsequently identified, and used to cluster discordant pairs into distinct gene-gene groups, yielding a list of genomic regions R adjacent to candidate fusion sites. For each member gene of a discordant gene pair, the genomic region R_i is defined by taking the minimum of all 3' exon/read coordinates in the cluster, and the maximum of all 5' exon/read coordinates in the cluster. These regions are used to prioritize the search for breakpoints in the next phase.

C2.2 Detection of breakpoints at base pair-resolution

Discordant read pairs may be introduced by NGS library preparation and/or sequencing artifacts (e.g., jumping PCR). However, they are also likely to flank the breakpoints of *bona fide* fusion events. As such, all discordant gene pairs identified in the preceding phase are ranked in decreasing order of discordant read depth (duplicate fragments are eliminated to correct for possible PCR bias), and genomic regions with a depth of at least 2x (by default) are further evaluated for potential breakpoints. Within each region, FACTERA analyzes all properly paired reads in which one of the two reads is "soft-clipped," or truncated. Soft-clipped reads allow for precise breakpoint determination, and are easily identified by parsing the CIGAR string associated with each mapped read, which compactly specifies the alignment operation used on each base (e.g. My = y contiguous bases were mapped, Sx = x bases were skipped). To simplify this step, only soft-clipped reads with the following two patterns are considered, SxMy and MySx, and the number of skipped bases x is required to be at least 16 (≤1 in 4.3B by random chance) to reduce the impact of non-specific sequence alignments.

To validate potential genomic breakpoints, defined as the edges of soft-clipped reads, FACTERA executes the following routine. For each discordant gene pair, all candidate breakpoints are tabulated, and the support (i.e. read frequency) for each is determined. Breakpoints supported by less than 2 reads (by default) are excluded from further analysis. Starting with the two breakpoints with highest support, FACTERA selects a representative soft-clipped read for each breakpoint, such that the length of the clipped sequence is closest to half of the read length. If the mapped region of one read matches the soft-clipped region of the other, FACTERA records a putative fusion event. To assess inter-read concordance, FACTERA employs the following algorithm. The mapped region of read 1 is parsed into all possible subsequences of length k (i.e., k-mers) using a sliding window (k = 10, by default). Each k-mer, along with its lowest sequence index in read 1, is stored in a hash table data structure, allowing k-mer membership to be assessed in constant time. Subsequently, the soft clipped sequence of read 2 is parsed into non-overlapping subsequences of length k, and the hash table is interrogated for matching k-mers. If a minimum matching threshold is achieved (=0.5 × the minimum length of the two compared subsequences), then the two reads are considered concordant. FACTERA will process at most 1000 (by default) putative breakpoint pairs for each discordant gene pair. Moreover, for each gene pair, FACTERA will only compare reads whose orientations are compatible with valid fusions. Such reads have soft-clipped sequences facing opposite directions. When this condition is not satisfied, FACTERA uses the reverse complement of read 1 for k-mer analysis.

In some instances, genomic subsequences flanking the true breakpoint may be nearly or completely identical, causing the aligned portions of soft-clipped reads to overlap. Unfortunately, this prevents an unambiguous determination of the breakpoint. As such, FACTERA incorporates a simple algorithm to arbitrarily adjust the breakpoint in one read (i.e., read 2) to match the other (i.e., read 1). For each read, FACTERA calculates the distance between the breakpoint and the read coordinate corresponding to the first k-mer match between reads. For example, let x be defined as the distance between the breakpoint coordinate of read 1 and the index of the first matching k-mer, j, and y be defined as the corresponding distance for read 2. Then, the offset is estimated as the difference in distances (x, y) between the two reads.

C2.3. In silico validation of candidate fusions

To confirm each candidate breakpoint *in silico*, FACTERA performs a local realignment of reads against a template fusion sequence (± 500bp around the putative breakpoint) extracted from the .2bit reference genome. BLAST is currently employed for this purpose, although BLAT or other fast aligners could be substituted. A BLAST database is constructed by collecting all reads that map to each candidate fusion sequence, including discordant reads and soft-clipped reads, as well as all unmapped reads in the original input .bam file. All reads that map to a given fusion candidate with at least 95% identity and a minimum length of 90% of the input read length (by default) are retained, and reads that span or flank the breakpoint are counted. As a final step, output redundancies are minimized by removing fusion sequences within a 20 bp interval of any fusion sequence with greater read support and with the same sequence orientation (to avoid removing reciprocal fusions).

FACTERA produces a simple output text file, which includes for each fusion sequence, the gene pair, the chromosomal sequence coordinates of the breakpoint, the fusion orientation (e.g., forward-forward or forward-reverse), the genomic sequences within 50 bp of the breakpoint, and depth statistics for reads spanning and flanking the breakpoint. Fusions identified in patients analyzed in this work are provided in **Supplementary Table 3**.

C2.4. Experimental validation of FACTERA

To experimentally evaluate the performance of FACTERA, we generated NGS data from two NSCLC cell lines, HCC78 (21.5M × 100 bp paired-end reads) and NCI-H3122 (19.4M × 100 bp paired-end reads), each of which has a known rearrangement (*ROS1* and *ALK*, respectively)[24,25] with a breakpoint that has, to the best of our knowledge, not been previously published. FACTERA readily revealed evidence for a reciprocal *SLC34A2-ROS1* translocation in the former and an *EML4-ALK* fusion in the latter. Precise breakpoints predicted by FACTERA were experimentally validated by PCR amplification and Sanger sequencing (**Supplementary Fig. 2a**; see also *Validation of Variants Detected by CAPP-Seq*). Importantly, FACTERA completed each run in practical time (~90 sec), using only a single thread on a hexa-core 3.4 GHz Intel Xeon E5690 chip. These initial results illustrate the utility of FACTERA as part of the CAPP-Seq analysis pipeline.

C2.5. Templated fusion discovery

We implemented a user-directed option to “hunt” for fusions within expected candidate genes. A fusion could be missed by FACTERA if the fusion detection criteria employed by FACTERA are incompletely satisfied—such as if discordant reads, but not soft-clipped reads, are identified—and will most likely occur when fusion allele frequency in the tumor is extremely low. As input, the method is supplied with candidate fusion gene sequences as "baits". All unmapped and soft-clipped reads in the input .bam file are subsequently aligned to these templates (using blastn) to identify reads that have sufficient similarity to both (for each read, 95% identity, e-value < 1.0e-5, and at least 30% of the read length must map to the template, by default). Such reads are output as a list to the user for manual analysis.

We tested this simple approach on a low purity tumor sample found to harbor an ALK fusion by FISH, but not FACTERA (i.e., case P9). Using templates for *ALK* and its common fusion partner, *ELM4*, we identified 4 reads that mapped to both, in a region with an overall depth of ~1900x. The estimated allele frequency of 0.21% was strikingly similar to the 0.22% tumor purity measured by FACS (data not shown), confirming the utility of the templated fusion discovery method. We subsequently FACS-depleted CD45+ immune populations and re-sequenced this patient's tumor. In the enriched tumor sample, FACTERA identified the *EML4-ALK* fusion, along with two novel ROS1 fusions (**Fig. 4b, Supplementary Table 3**).

C3. Mutation Recovery:SNVs/indels

Using a custom Perl script, previously identified reporter alleles were intersected with a SAMtools mpileup file generated for each plasma DNA sample, and the number and frequency of supporting reads was calculated for each reporter allele. Only reporters in properly paired reads at positions with at least 500x overall depth (pre-duplication removal; see section B2.1) were considered (**Supplementary Table 4**).

C4. Mutation Recovery: Fusions

For enumeration of fusion frequency in sequenced plasma DNA, FACTERA executes the last step of the discovery phase (i.e., *in silico validation of candidate fusions*, above) using the set of previously identified fusion templates. The fusion allele frequency is calculated as α / β, where α is the number of breakpoint-spanning reads, and β is the mean overall depth within a genomic region ± 5bps around the breakpoint. Regarding the NSCLC selector described in this work, the latter calculation was always performed on the single gene contained in the NSCLC selector library. If both fusion genes are targeted within a selector library, overall depth is estimated by taking the mean depth calculated for both genes.

Notably, in some cases we observed lower fusion allele frequencies than would be expected for heterozygous alleles (e.g., see cell line fusions in **Supplementary Table 3**). This was seen in cell lines, in an empirical spiking experiment, and in one patient's tumor and plasma samples (i.e., P6), and could potentially result from inefficient "pull-down" of fusions whose partners are not represented in the selector. Regardless, fusions are useful reporters—they possess virtually no background signal and show linear behavior over defined concentrations in a spiking experiment (**Supplementary Fig. 5d**). Moreover, allelic frequencies in plasma are easily adjusted for such inefficiencies by dividing the measured frequency in plasma by the corresponding frequency in the tumor. In cases where sequenced tumor tissue is impure, tumor content can be estimated using the frequencies of SNVs (or indels) as a reference frame, allowing the fusion fraction to be normalized accordingly (**Supplementary Table 4**).

C5. Screening Plasma DNA without Knowledge of Tumor DNA

We devised the following statistical algorithm as an initial step toward non-invasive tumor genotyping and cancer screening with CAPP-Seq. The method identifies candidate tumor-derived SNVs using sequential models of (i) background noise in paired germline DNA (in this work, PBLs), (ii) base-pair resolution background frequencies in plasma DNA across the selector, and (iii) sequencing error in plasma DNA. Anecdotal examples are provided in **Supplementary Fig. 7**. The algorithm works in four main steps, detailed below.

As input, the algorithm takes allele frequencies from a single plasma DNA sample and analyzes high quality *background alleles*, defined in a first step for each genomic position as the non-dominant base with highest fractional abundance, excluding germline SNPs. Only alleles with depth of at least 500x and strand bias <90% (conservative, by default) are analyzed. For consistency with variant calling, we allowed the screening approach to interrogate selector regions within 500 bp of defined coordinates, expanding the effective sequence space from ~125 kb to ~600 kb.

Second, the binomial distribution is used to test whether a given non-reference allele in plasma DNA is significantly different from the corresponding paired germline allele (**Supplementary Fig. 7a,b**). Here the probability of success is taken to be the frequency of the background allele in PBLs, and the number of trials is the allele's corresponding depth in plasma DNA. To avoid contributions from alleles in rare circulating tumor cells that might contaminate PBLs, input alleles with a fractional abundance greater than 0.5% in paired PBLs (by default) or a Bonferroni-adjusted binomial probability greater than 2.08×10^{-8} are not further considered (alpha of 0.05 / [~600 kb × 4 alleles per position]).

Third, a database of plasma DNA background allele frequencies is assembled. Here, we used samples analyzed in the present study (i.e., pre-treatment NSCLC samples and 1 sample from a healthy volunteer), except the input sample is left out to avoid bias. Based on the assumption that all background allele fractions follow a normal distribution, a Z-test is employed to test whether a given input allele differs significantly from typical circulating DNA background at the same position (**Supplementary Fig. 7a,b**). All alleles within the selector are evaluated, and those with an average background frequency of 5% or greater (by default) or a Bonferroni-adjusted single-tailed Z-score <5.6 are not further considered (alpha of 0.05, adjusted as above).

Finally, candidate alleles are tested for remaining potential sequencing errors. This step leverages the observation that non-tumor variants (i.e., "errors") in plasma DNA tend to have a higher duplication rate than *bona fide* variants detectable in the patient's tumor (data not shown). As such, the number of supporting reads is compared for each input allele between nondeduped (all fragments meeting QC critiera; see **Online Methods**) and deduped data (only unique fragments meeting QC criteria). An outlier analysis is then used to distinguish candidate tumor-derived SNVs from remaining background noise (**Supplementary Fig. 7a–c**). Specifically, to reveal outlier tendency in the data, the square root of the robust distance *Rd* (Mahalanobis distance) is compared against the square root of the quantiles of a chi-squared distribution *Cs*. By applying this transformation, we observed a natural separation between true SNVs and false positives in cancer patients (**Supplementary Fig. 7a,c**), and notably, an absence of outlier structure in patient samples lacking tumor-derived SNVs (**Supplementary Fig. 7b,c**). To automatically call SNVs without prior knowledge, the screening approach iterates through data points by decreasing *Rb* and recalculating the Pearson's correlation coefficient *Rho* between *Rd* and *Cs* for points 1 to *i*, where Rd_i is the current maximum *Rd*. The algorithm iteratively reports outliers (i.e., candidate tumor-derived SNVs) until it terminates when *Rho* ≥ 0.85.

Supplementary References

1. Fan, H.C. & Quake, S.R. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. *PLoS One* **5**, e10439 (2010).
2. Lund, A.H., Duch, M. & Pedersen, F.S. Increased cloning efficiency by temperature-cycle ligation. *Nucleic acids research* **24**, 800-801 (1996).
3. Fisher, S., *et al.* A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome biology* **12**, R1 (2011).
4. Dias-Santagata, D., *et al.* Rapid targeted mutational analysis of human tumours: a clinical platform to guide personalized cancer medicine. *EMBO molecular medicine* **2**, 146-158 (2010).
5. Motulsky, H.J. & Brown, R.E. Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate. *BMC bioinformatics* **7**, 123 (2006).
6. Koboldt, D.C., *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576 (2012).
7. Li, H., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
8. Stouffer, S., DeVinney, L. & Suchmen, E. *The American Soldier: Adjustment During Army Life*, (Princeton University Press, Princeton, NJ, 1949).

9. Fan, H.C., Blumenfeld, Y.J., Chitkara, U., Hudgins, L. & Quake, S.R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci U S A* **105**, 16266-16271 (2008).
10. Hammerman, P.S.*, et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).
11. Forbes, S.A.*, et al.* COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic acids research* **38**, D652-657 (2010).
12. Ding, L.*, et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075 (2008).
13. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27**, 175-181 (2011).
14. Okuda, K., Sasaki, H., Yukiue, H., Yano, M. & Fujii, Y. Met gene copy number predicts the prognosis for completely resected non-small cell lung cancer. *Cancer science* **99**, 2280-2285 (2008).
15. Su, Z.*, et al.* A platform for rapid detection of multiple oncogenic mutations with relevance to targeted therapy in non-small-cell lung cancer. *J Mol Diagn* **13**, 74-84 (2011).
16. Tsao, M.S.*, et al.* Prognostic and predictive importance of p53 and RAS for adjuvant chemotherapy in non small-cell lung cancer. *J Clin Oncol* **25**, 5240-5247 (2007).
17. Chaft, J.E.*, et al.* Coexistence of PIK3CA and other oncogene mutations in lung adenocarcinoma-rationale for comprehensive mutation profiling. *Molecular cancer therapeutics* **11**, 485-491 (2012).
18. Paik, P.K.*, et al.* Clinical characteristics of patients with lung adenocarcinomas harboring BRAF mutations. *J Clin Oncol* **29**, 2046-2051 (2011).
19. Stephens, P.*, et al.* Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* **431**, 525-526 (2004).
20. Jin, G.*, et al.* PTEN mutations and relationship to EGFR, ERBB2, KRAS, and TP53 mutations in non-small cell lung cancers. *Lung Cancer* **69**, 279-283 (2010).
21. Malanga, D.*, et al.* Activating E17K mutation in the gene encoding the protein kinase AKT1 in a subset of squamous cell carcinoma of the lung. *Cell Cycle* **7**, 665-669 (2008).
22. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
23. Chen, K.*, et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681 (2009).
24. Bergethon, K.*, et al.* ROS1 rearrangements define a unique molecular class of lung cancers. *J Clin Oncol* **30**, 863-870 (2012).
25. McDermott, U.*, et al.* Genomic alterations of anaplastic lymphoma kinase may sensitize tumors to anaplastic lymphoma kinase inhibitors. *Cancer research* **68**, 3389-3395 (2008).