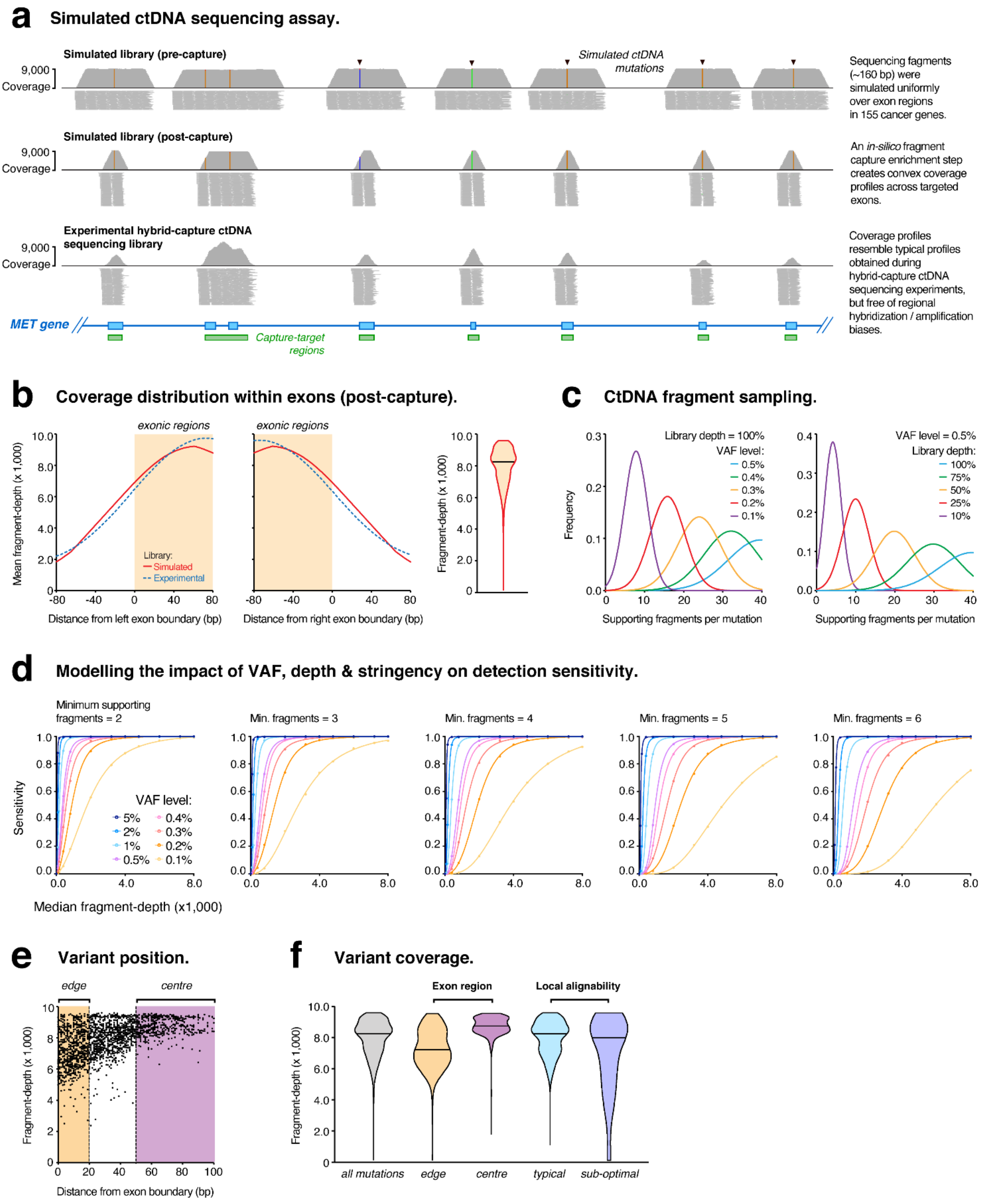
Supplementary Materials

TABLE OF CONTENTS



a Simulated ctDNA sequencing assay.

Sequencing fagments (~160 bp) were simulated uniformly over exon regions in 155 cancer genes.

An *in-silico* fragment capture enrichment step creates convex coverage profiles across targeted exons.

Coverage profiles resemble typical profiles obtained during hybrid-capture ctDNA sequencing experiments, but free of regional hybridization / amplification biases.

b Coverage distribution within exons (post-capture).

c CtDNA fragment sampling.

d Modelling the impact of VAF, depth & stringency on detection sensitivity.

e Variant position.

f Variant coverage.

Fig. S1. Simulated hybrid-capture ctDNA sequencing experiment. (**a**) Genome browser view showing coverage of simulated sequencing fragments within the *MET* oncogene before (upper) and after (middle) *in silico* hybrid-capture enrichment. Sequencing fragments from an experimental ctDNA sequencing library (lower) are shown for comparison. The experimental library is from hybrid-capture sequencing on the contrived reference samples *Lbx-low*, as analysed with the ROC ctDNA assay. (**b**) Plots show average per-base fragment-depth relative to exon boundaries, across all on-target exons, in simulated ctDNA sequencing libraries, following *in silico* capture. Equivalent profiles are also shown for an experimental ctDNA sequencing library (dashed line). (**c**) Histograms show the number of supporting read-fragments per mutation within simulated ctDNA sequencing libraries. The left plot shows distributions for mutations at different variant allele frequency (VAF) levels, at maximum simulated library depth (8,252-fold median fragment-depth). The right plot shows distributions for mutations at 0.5% VAF, at depreciating library depths (relative to maximum). The number of sequence fragments containing a given mutation follows a Poisson distribution, with a median fragment count that is proportionate to the product of VAF and global fragment-depth. (**d**) Curves modelling the relationship between simulated library depth (median fragment-depth) and detection sensitivity for simulated mutations. Within each plot, mutations are parsed into different VAF levels and separate plots show increasing levels of detection stringency (i.e., minimum number of fragments for a mutation to be called). (**e**) Scatter-plot shows fragment-depth recorded at site of each simulated mutations, relative to their distance from the nearest exon boundary (mutations >100 bp from exon boundaries not shown). (**f**) Violin plots show coverage distributions for simulated mutations, parsed according to exon region and local alignability.

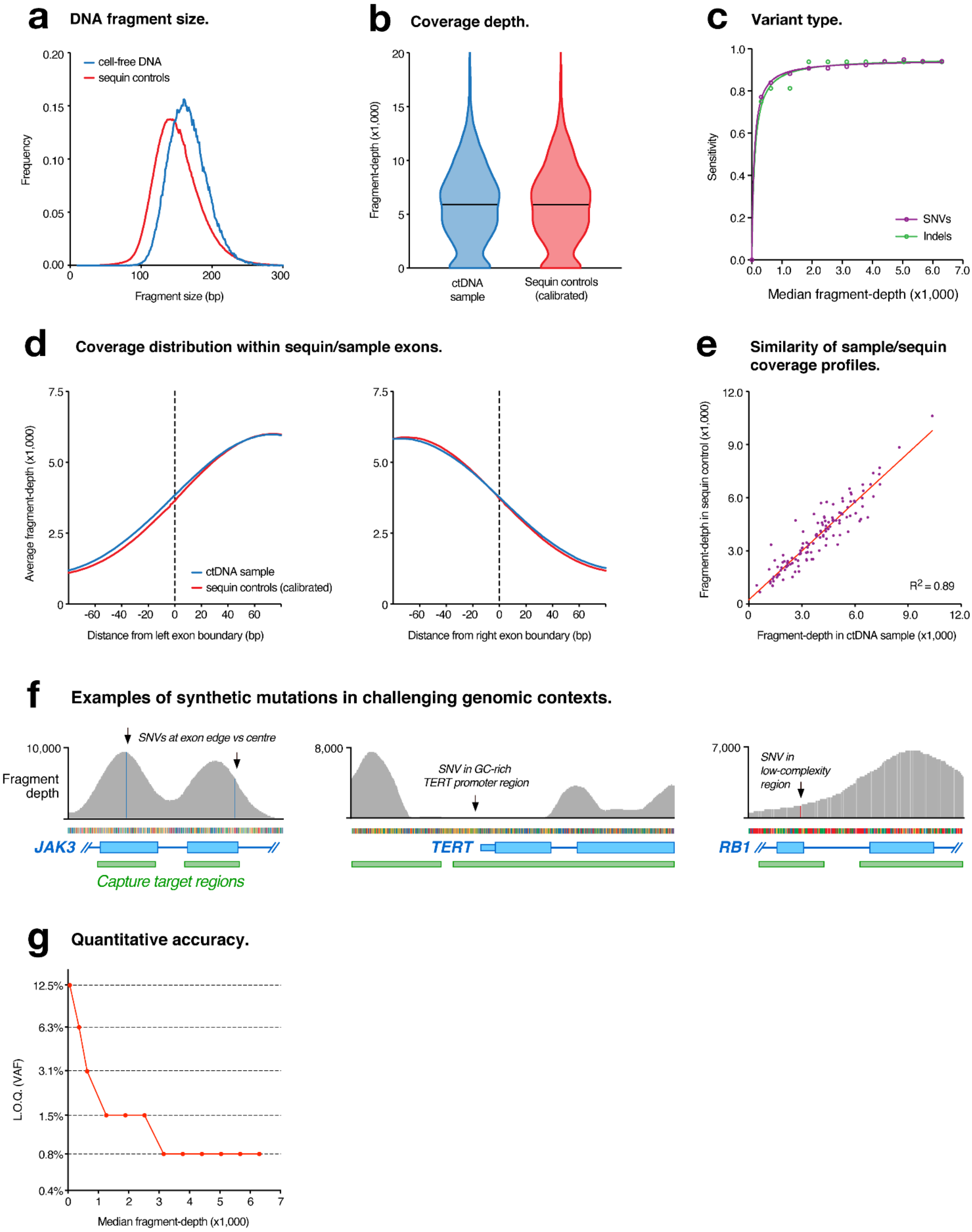
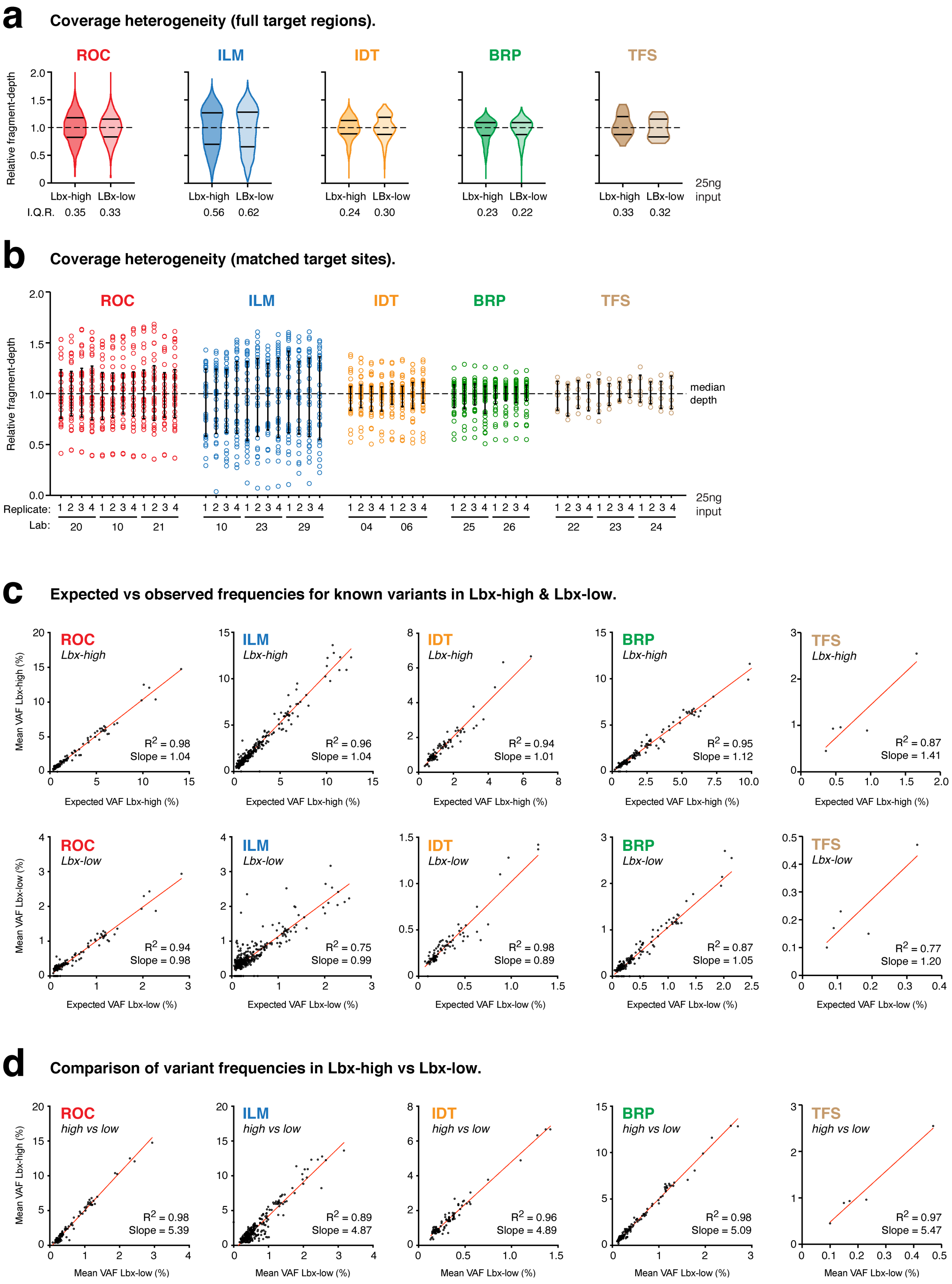

a
DNA fragment size.
0.20
0.15
0.10
0.05
0.00
Frequency
cell-free DNA
sequin controls
0
100
200
300
Fragment size (bp)
b
Coverage depth.
20
15
10
5
0
Fragment-depth (x1,000)
ctDNA sample
Sequin controls (calibrated)
c
Variant type.
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
SNVs
Indels
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
Median fragment-depth (x1,000)
d
Coverage distribution within sequin/sample exons.
7.5
5.0
2.5
0
Average fragment-depth (x1,000)
ctDNA sample
sequin controls (calibrated)
-60
-40
-20
0
20
40
60
Distance from left exon boundary (bp)
7.5
5.0
2.5
0
-60
-40
-20
0
20
40
60
Distance from right exon boundary (bp)
e
Similarity of sample/sequin coverage profiles.
12.0
9.0
6.0
3.0
0
Fragment-detph in sequin control (x1,000)
$R^2 = 0.89$
0
3.0
6.0
9.0
12.0
Fragment-depth in ctDNA sample (x1,000)
f
Examples of synthetic mutations in challenging genomic contexts.
10,000
Fragment depth
SNVs at exon edge vs centre
JAK3
Capture target regions
8,000
SNV in GC-rich TERT promoter region
TERT
7,000
SNV in low-complexity region
RB1
g
Quantitative accuracy.
12.5%
6.3%
3.1%
1.5%
0.8%
0.4%
L.O.Q. (VAF)
0
1
2
3
4
5
6
7
Median fragment-depth (x1,000)

Fig. S2. Synthetic sequin experiment. (**a**) Histograms show DNA fragment-size distributions for NGS read-fragments derived from synthetic sequin controls (red) compared to fragments from a typical patient cell-free DNA sample. (**b**) Violin plots show coverage distributions (unique fragment-depth) for on-target exon regions within sequin controls (red) compared to their accompanying human reference ctDNA sample, after sequin coverage calibration was performed. (**c**) Curves modelling the relationship between library depth (median-fragment depth) and detection sensitivity for synthetic sequin mutations, with SNVs (purple) and indels (green) shown separately. (**d**) Plots show average per-base fragment-depth relative to exon boundaries, across all on-target exons within sequin controls (red) and accompanying human reference ctDNA sample (blue). (**e**) Scatter-plot compares fragment-depth at sites of synthetic mutations observed in sequin controls to accompanying human sample. The strong correlation ($R^2 = 0.89$) indicates that coverage profiles were highly similar between sequin controls and their accompanying sample. (**f**) Genome browser views show examples of synthetic mutations occurring in challenging genome contexts. Left: comparison of SNVs at exon edge to exon centre, with the edge mutation exhibiting lower fragment-depth. Centre: Low or nil coverage was obtained within the highly GC-rich *TERT* promoter region, obscuring detection of the synthetic mutation (c.-57A>C) in this region. Right: low coverage obtained within a region of low sequence complexity in *RB1*. (**g**) Plot shows the relationship between the limit of quantitative accuracy (L.O.Q.; i.e., the lowest VAF level above which two-fold increases in VAF could be reliably detected) and fragment-depth.



a Coverage heterogeneity (full target regions).

b Coverage heterogeneity (matched target sites).

c Expected vs observed frequencies for known variants in Lbx-high & Lbx-low.

d Comparison of variant frequencies in Lbx-high vs Lbx-low.

Fig. S3. Coverage heterogeneity and observed variant frequencies for ctDNA assays. (**a**) Violin plots show coverage distributions (unique fragment-depth) after normalising to median depth in each assay for *Lbx-high* and *Lbx-low* (25 ng input) replicates. The variance of a given distribution indicates the degree of coverage heterogeneity for that assay. Bars indicate 25% and 75% quartiles, and interquartile ranges are noted below each plot. (**b**) Dot plots show normalised fragment depth measurements at 33 known variant sites that are present within the target regions of all participating hybrid-capture panels. 6 known variants within the TFS amplicon target regions are also shown for comparison. Bars indicate median +/- interquartile ranges. (**c**) Scatter plots show observed vs expected variant allele frequencies (VAFs) for on-target known variants in *Lbx-high* (upper) and *Lbx-low* (lower) at 25 ng input. (**d**) Scatter plots compare VAFs for on-target, known variants between *Lbx-high* (vertical axis) and *Lbx-low* (horizontal axis). Slope of x=5 reflects 5-fold dilution of variant alleles in *Lbx-low*, compared to *Lbx-high*.

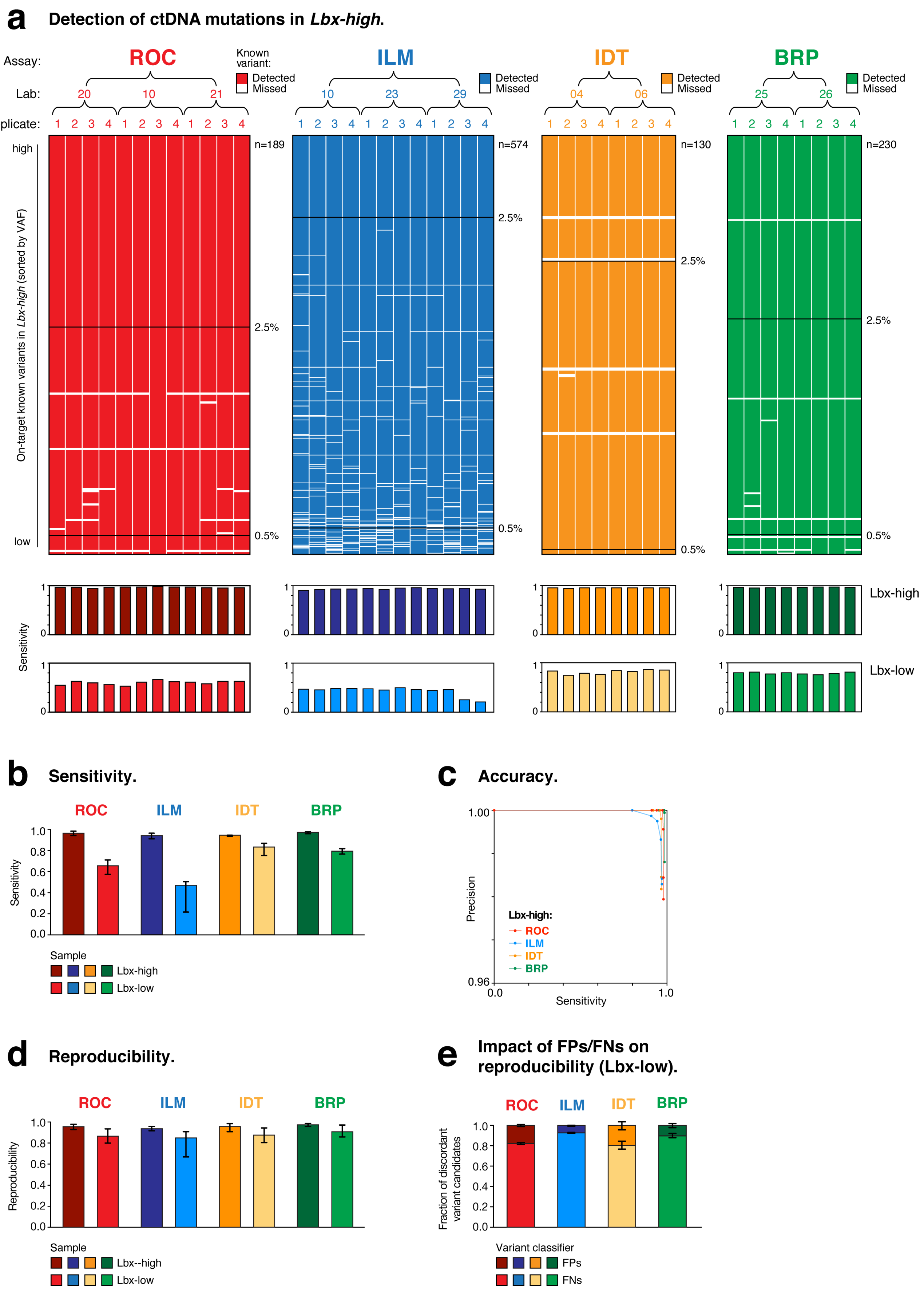

a
Detection of ctDNA mutations in *Lbx-high*.
Assay:
Lab:
Replicate:
ROC
ILM
IDT
BRP
Known variant:
Detected
Missed
20
10
21
23
29
04
06
25
26
1 2 3 4
n=189
n=574
n=130
n=230
2.5%
0.5%
high
low
On-target known variants in *Lbx-high* (sorted by VAF)
Sensitivity
1
0
Lbx-high
Lbx-low
b
Sensitivity.
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Sample
Lbx-high
Lbx-low
c
Accuracy.
Precision
1.00
0.96
0.0
1.0
Sensitivity
Lbx-high:
ROC
ILM
IDT
BRP
d
Reproducibility.
Reproducibility
Sample
Lbx--high
Lbx-low
e
Impact of FPs/FNs on reproducibility (Lbx-low).
Fraction of discordant variant candidates
Variant classifier
FPs
FNs

Fig. S4. Comparison of performance between hybrid-capture assays at 25ng input. (**a**; upper) Heatmaps show the detection of known variants (rows) in ctDNA assay replicates (columns). Known variants are sorted by expected variant allele frequency (VAF) in descending order, and replicates are arranged hierarchically by assay type, test lab and replicate number. Heatmaps show results for *Lbx-high* at 25ng input and equivalent heatmaps for *Lbx-low* are shown in **Fig. 4a.** (**a**; lower) Aligned below each heatmap column, bar charts indicate the sensitivity of variant detection in each assay. Sensitivity is reported separately for known variants (VAF 0-100%) in *Lbx-high* and *Lbx-low*. (**b**) Bar charts (*n* = 12 for ROC, ILM; *n* = 8 for IDT, BRP; median ± range) show overall sensitivity for participating assays in *Lbx-high* and *Lbx-low*. (**c**) Precision-recall curves compare diagnostic performance of participating ctDNA assays for *Lbx-high* (25ng input). Equivalent curves for *Lbx-low* are shown in **Fig. 2b.** (**d**) Bar charts (*n* = 132 for ROC, ILM; *n* = 56 for IDT, BRP; median ± range) show pairwise reproducibility scores for participating assays in *Lbx-high* and *Lbx-low*. (**e**) Plots show the fraction of discordant variants between a given pair of assay replicates, within known positions, that are false-positives (FPs) and false-negatives (FNs; mean ± 95% CI). FNs are more common than FPs, indicating that poor reproducibility is primarily caused by lack of sensitivity, rather than high FP rates.

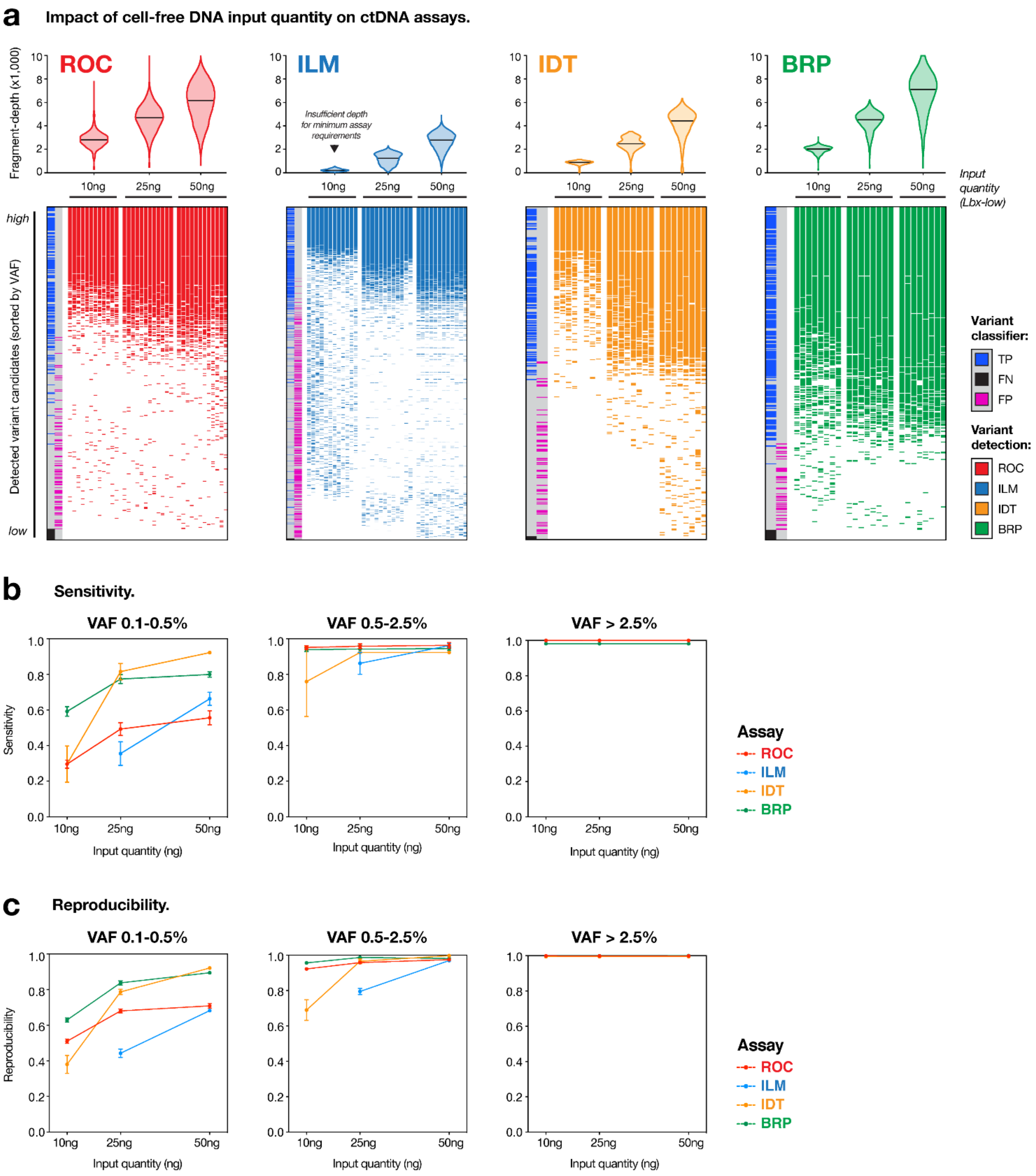


Fig. S5. Impact of cell-free DNA input quantity. (**a**) Heatmaps show the impact of increasing *Lbx-low* input quantity, and proportionate increase in fragment-depth (violin plots), on variant detection in participating ctDNA assays. All on-target variant candidates are shown (rows) and sorted by observed VAF in descending order. Assay replicates (columns) are arranged hierarchically by assay type, input amount, test lab and replicate number. Variant candidates that fall within known positions are also classified as true-positives (TPs; blue) or false-positives (FPs; pink) and known variants that were missed in every replicate are indicated (FNs; black). (**b**) Curves showing the relationship between cell-free DNA input quantity (*Lbx-low*) and sensitivity for each participating ctDNA assay. Sensitivity is reported separately for known variants at high (VAF > 2.5%), intermediate (0.5-2.5%) and low (0.1-0.5%) frequency. (**c**) Curves showing the relationship between cell-free DNA input quantity (*Lbx-low*) and reproducibility for each participating ctDNA assay. Reproducibility is reported separately for variant candidates at high (VAF > 2.5%), intermediate (0.5-2.5%) and low (0.1-0.5%) frequency. Error bars are 95% confidence intervals.

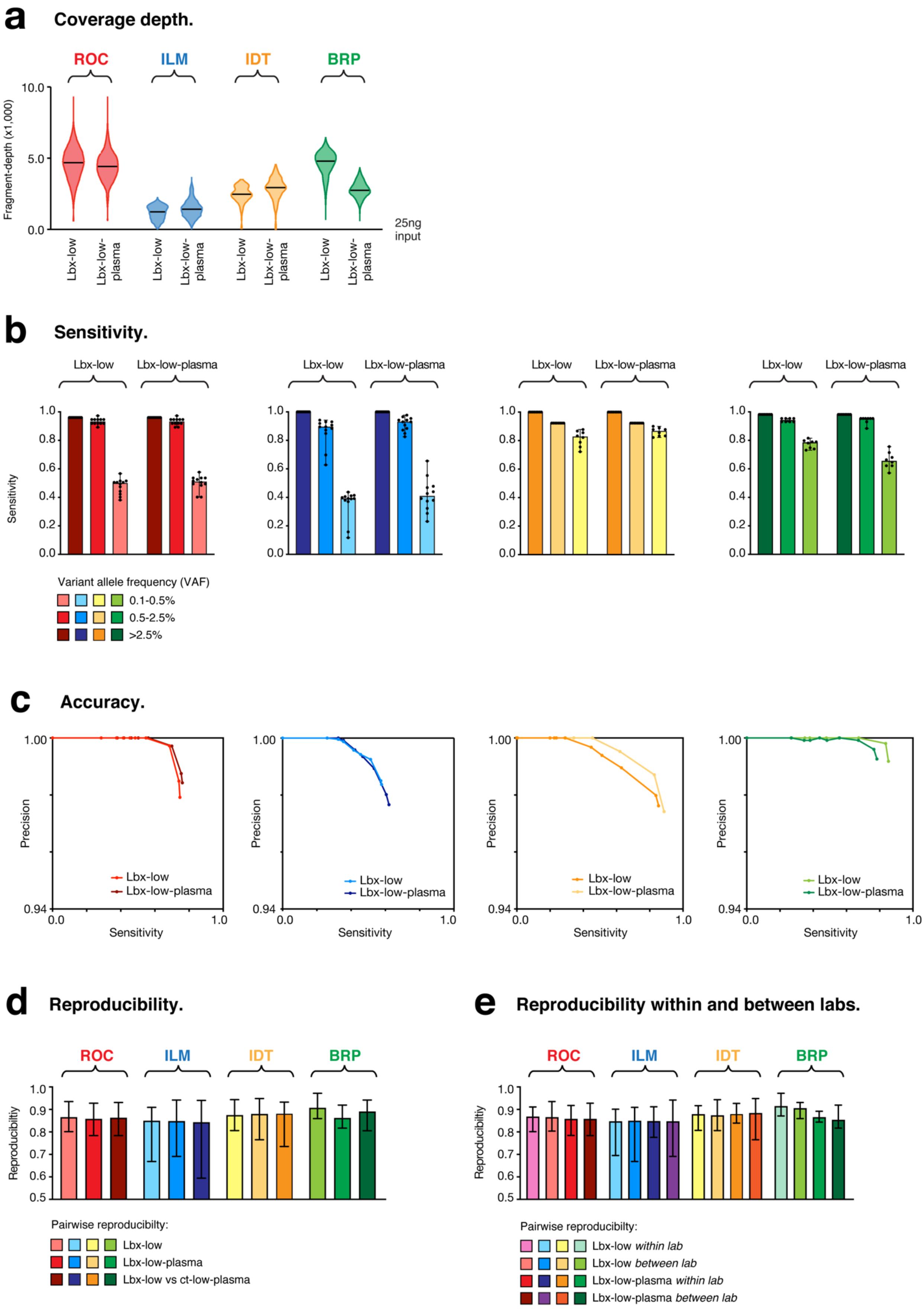

a
Coverage depth.
ROC
ILM
IDT
BRP
Fragment-depth (x1,000)
10.0
5.0
0.0
25ng input
Lbx-low
Lbx-low-plasma
b
Sensitivity.
Lbx-low
Lbx-low-plasma
Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Variant allele frequency (VAF)
0.1-0.5%
0.5-2.5%
>2.5%
c
Accuracy.
Precision
1.00
0.94
0.0
1.0
Sensitivity
Lbx-low
Lbx-low-plasma
d
Reproducibility.
ROC
ILM
IDT
BRP
Reproducibility
1.0
0.9
0.8
0.7
0.6
0.5
Pairwise reproducibilty:
Lbx-low
Lbx-low-plasma
Lbx-low vs ct-low-plasma
e
Reproducibility within and between labs.
ROC
ILM
IDT
BRP
Reproducibility
1.0
0.9
0.8
0.7
0.6
0.5
Pairwise reproducibilty:
Lbx-low within lab
Lbx-low between lab
Lbx-low-plasma within lab
Lbx-low-plasma between lab

Fig. S6. Impact of plasma-DNA extraction. (**a**) Violin plots show coverage distributions (fragment-depth) for *Lbx-low* and *Lbx-low-plasma* replicates in each participating assay. All assays used 25ng input amounts, however, we note that inaccuracy in the quantification of post-extraction DNA caused lower input amounts to be used for *Lbx-low-plasma* in BRP replicates, compared to *Lbx-low,* explaining the lower coverage for this assay. (**b**) Bar charts (*n* = 12 for ROC, ILM; *n* = 8 for IDT, BRP; median ± range) show sensitivity for participating assays in *Lbx-low* and *Lbx-low-plasma*. Sensitivity is reported separately for known variants at high (VAF > 2.5%), mid (0.5-2.5%) and low (0.1-0.5%) frequency. (**c**) Precision-recall curves compare diagnostic performance in *Lbx-low* and *Lbx-low-plasma* across participating ctDNA assays. (**d,e**) Bar charts (*n* = 132 for ROC, ILM; *n* = 56 for IDT, BRP; median ± range) show pairwise reproducibility scores for participating assays: (**d**) reproducibility is reported separately for comparisons of *Lbx-low* replicates, *Lbx-low-plasma* replicates, and for comparisons of *Lbx-low* to *Lbx-low-plasma* replicates; (**e**) reproducibility is reported separately for all within-lab and between-lab pairwise comparisons among *Lbx-low* and *Lbx-low-plasma* replicates*.* In general, no difference in sensitivity, accuracy or reproducibility was observed between *Lbx-low* and *Lbx-low-plasma* at matched input amounts (25ng).

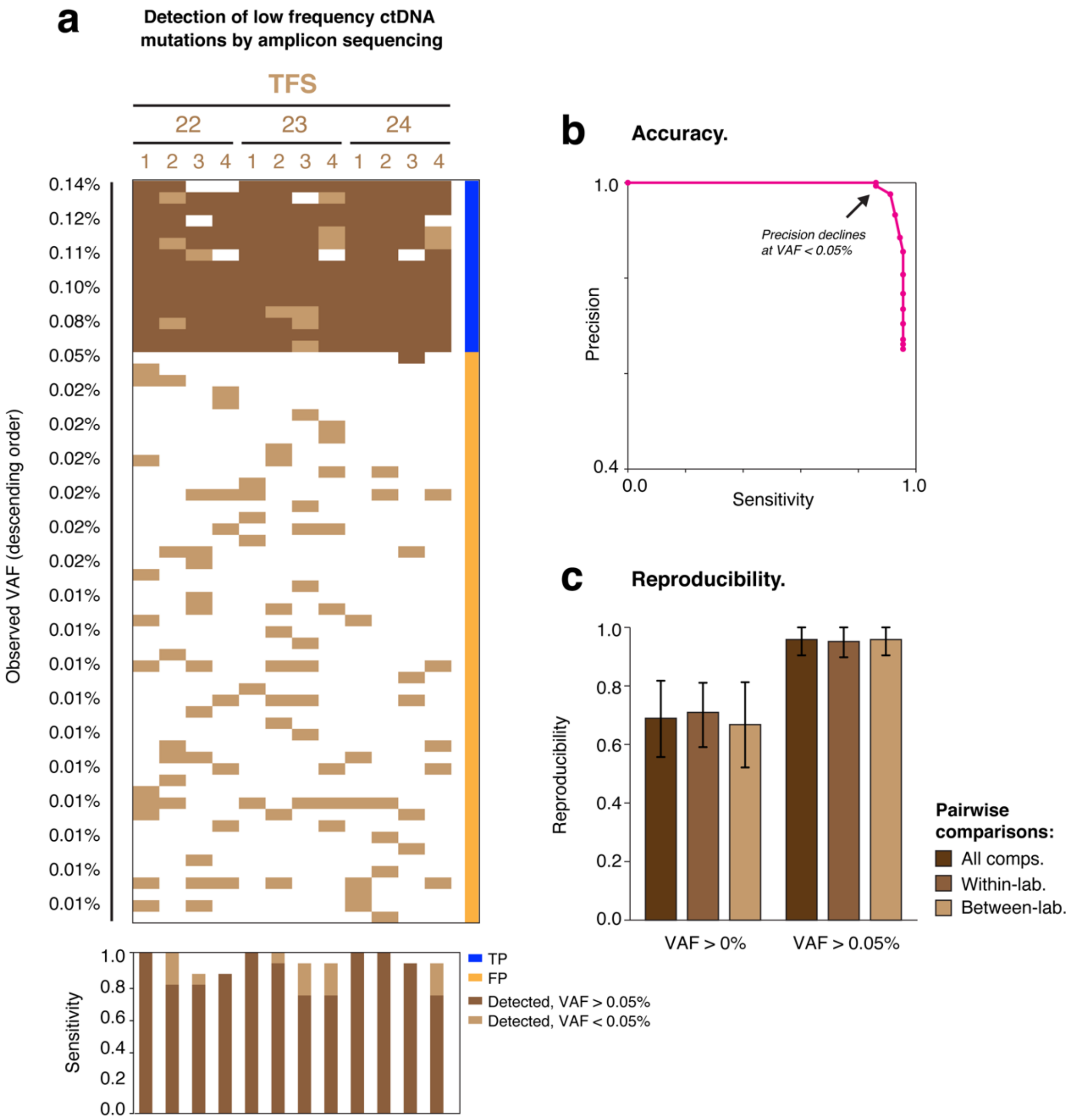


Figure S7. Evaluating detection of low-frequency mutations with TFS amplicon sequencing assay. TFS test sites analyzed AcroMetrix Oncology Hotspot Control, at 50 ng input quantity, containing 15 known cancer mutations that overlapped TFS hotspot regions, present at ~0.1% VAF. (**a**; upper) Detection heatmaps show on-target variant candidates (rows), sorted by observed VAF in descending order. Assay replicates (columns) are arranged hierarchically by test lab and replicate number. Variant candidates are classified as true-positives (TPs; blue) or false-positives (FPs; yellow) and known variants that were missed in every replicate are indicated (FNs; black). (**a**; lower) Aligned below each heatmap column, bar charts indicate the sensitivity of variant detection in each replicate. (**b**) Precision-recall curve evaluates the diagnostic accuracy of low frequency variant detection. Precision declines sharply below a detection threshold of ~0.05% VAF, with relatively small sensitivity gains beyond this cut-off. (**c**) Bar charts (median ± range) show pairwise reproducibility scores, reported separately for all within-lab and between-lab pairwise comparisons. Applying a detection threshold of VAF > 0.05% leads to a large increase in reproducibility.

Supplementary Table 1. Comparison of tumor-tissue and plasma DNA sequencing for precision oncology. This table is adapted from Aggarwal et al (2020)[6].

Characteristic	Tumor-tissue analysis	Plasma ctDNA analysis
Accessibility	May be problematic if the tissue is not immediately available or is inadequate. Additionally, other factors (e.g., comorbidities, anatomic location) may add significant risk for the acquisition of solid tissue.	Highly accessible along with extensive availability of blood collection tubes (e.g., EDTA and Streck)
Turnaround time	Usually slower, particularly if the sample must first be requested from an outside facility. If a new biopsy sample is required, an invasive procedure or small operation must be scheduled and coordinated by different medical specialties (e.g., radiology or surgery, anaesthesiology, pathology)	Fast since it is obtained by a routine blood draw. This is further facilitated by expedited shipping of the sample if a commercial lab is performing the assay.
Sensitivity	Excellent, since all tumor biopsy samples undergo a review for assessment of tumor cell count and purity. Genotyping is only performed on specimens deemed adequate for analysis.	Lower since there is no suitability review (e.g., tumor cell count and purity). Depending on the tumor type and burden of disease, ctDNA may not be detectable, particularly with tumors having minimal shed.
Specificity	Usually excellent except that germline variants may sometimes be reported as somatic. This may be resolved if a germline sample and/or specific bioinformatic approaches are included as part of the analysis.	Excellent for targetable driver mutations. False positives may result for certain genes (especially at lower allelic fractions) due to clonal hematopoiesis of indeterminate potential (CHIP). This may be resolved if a germline sample (i.e., buffy coat PBMCs) and/or specific bioinformatic approaches are included as part of the analysis.
Expense	Highly variable, depending on the type of invasive procedure needed to obtain the tissue, and the number of medical specialities involved.	Variable. If the ctDNA analysis is negative, a tumor tissue biopsy is usually required. Otherwise, the blood/plasma required for the ctDNA assay is obtained by a routine blood draw, which is a minimal expense.
Repeatability	Usually not done due to the invasiveness of tumor biopsy procedures that may result in significant risk to the patient. The tumor is usually only genotyped at initial diagnosis.	Easily done via routine blood draw. As the prices of NGS assays continue to fall, ctDNA follow-up assays may replace or augment routine imaging studies for cancer monitoring.
Scope	Possibility of *spatial bias*, depending on needle placement into a heterogeneous solid tumor. This is further exacerbated in situations involving metastatic disease.	Potential of sampling ctDNA from multiple tumor foci yielding a more comprehensive tumor analysis and therapy plan.

Supplementary Table 2. Theoretical representation of ctDNA fragments in patient plasma.

Plasma from healthy donors typically yields ~5-10 ng of cell-free DNA per mL[10]. The amount of ctDNA derived from a cancer patient may be highly variable and exhibits dependencies based on the organ of origin, burden of disease, and the particular shedding characteristics of that cancer[4].

The quantitation of DNA in a single cell is ~6 pg. Since cancer mutations are largely heterozygous due to their somatic nature, estimates below are made using a haploid genome approach. Thus, the quantitation of ctDNA in single cell is estimated at ~3 pg. A sample input of 10 ng cell free DNA is used to begin table generation: 10ng / 0.003 ng = 3000 genome equivalent copies.

Input cell-free DNA (ng/mL)	Genome Equiv. Copies	Expected ctDNA-fragment copies at VAF = X.				
		2.5%	1%	0.5%	0.25%	0.1%
10	3000	75	30	15	7.5	3
20	6000	150	60	30	15	6
30	9000	225	90	45	22.5	9
40	12000	300	120	60	30	12
50	15000	375	150	75	37.5	15
60	18000	450	180	90	45	18
70	21000	525	210	105	52.5	21
80	24000	600	240	120	60	24
90	27000	675	270	135	67.5	27
100	30000	750	300	150	75	30
200	60000	1500	600	300	150	60

Supplementary Table 3. Description of participating ctDNA assays.

Information regarding five industry-leading ctDNA assays enrolled in the proficiency study (ROC, ILM, IDT, BRP, TFS). For each ctDNA assay is shown: the total size of all reportable panel regions ('Reportable region'), the proportion of this that is within protein-coding positions ('Coding') and consensus target regions ('CTR'), the number of known negative positions ('Negatives') and ('Variants'). All region sizes are reported in kb. The reporting region of participating hybrid capture assays (ROC, ILM, IDT, BRP) ranged from 110 kb to 501 kb, which was much larger than that of the TFS amplicon panel (1.9 kb). All participating assays are for Research Use Only, not for diagnostic procedures.

Name	Vendor	ctDNA assay	Sequencing platform	Target genes	Reportable region (kb)	Coding (kb)	CTR (kb)	Negatives (× 1,000)	Variants
ROC	Roche Sequencing Solutions	AVENIO ctDNA (Expanded Kit)	Illumina NextSeq	77	161.7	140.2	103.8	47.1	189
ILM	Illumina	TruSight Tumor 170 + UMI	Illumina NovaSeq	154	501.0	390.1	338.4	133.0	574
IDT	Integrated DNA Technologies	xGen Non-small Cell Lung Cancer	Illumina NovaSeq	24	110.1	93.2	76.5	39.3	130
BRP	Burning Rock Biotech	Lung Plasma v4	Illumina NovaSeq	168	226.9	148.5	125.1	53.4	229
TFS	Thermo Fisher Scientific	Oncomine Lung cfDNA assay	Ion Torrent S5 XL	11	1.9	1.6	1.3	0.8	5

Supplementary Table 4. Assay recovery rates and coverage heterogeneity.

Interquartile range (I.Q.R.) and coefficient of variation (C.V.) indicate coverage heterogeneity relative to median fragment-depth. Estimated recovery rate calculations are based on assumption of single haploid genome = 3 pg. Hence, 10 ng = 3000 genome equivalent copies, 25 ng = 7500 genome equivalent copies, 50ng = 15,000 genome equivalent copies.

	ROC			ILM			IDT			BRP			TFS		
Input	**10ng**	**25ng**	**50ng**	**10ng**	**25ng**	**50ng**	**10ng**	**25ng**	**50ng**	**10ng**	**25ng**	**50ng**	**10ng**	**25ng**	**50ng**
Med. depth	2819	4698	6157	221	1235	2799	887	2476	4448	2027	4528	7104	1113	3280	6379
I.Q.R.	752	1523	2391	143	772	1046	179	768	1347	363	998	2259	472	1151	2197
C.V.	25%	28%	27%	43%	40%	28%	20%	22%	26%	16%	20%	26%	31%	21%	27%
Est. recovery	94%	63%	41%	7%	17%	19%	30%	33%	30%	68%	60%	47%	37%	44%	85%

Supplementary Table 5: False-positive rates.

FP-rates are reported as FP/kb for each hybrid-capture ctDNA assay across a range of minimum VAF thresholds (0-0.5%). Values reported here are for *Lbx-low* at 25 ng input.

			FP-rate (FP / kb) at specified VAF threshold		
Assay	**Known negatives (kb)**	**FPs per replicate (mean [range])**	> 0%	> 0.1%	> 0.5%
ROC	47.1	2.91 [1-6]	0.061	0.044	0.000
ILM	133	5.25 [2-10]	0.039	0.039	0.008
IDT	39.3	2.75 [0-6]	0.070	0.057	0.000
BRP	53.4	1.65 [0-5]	0.030	0.007	0.000

Supplementary Table 6. Sequencing information for hybrid-capture assays.

*8 libraries per flow cell. **This panel size was calculated from its constituent probe sequences.

Assay	Sequencing platform	Flow-cell	Libraries per lane	Read-length (bp)	Average read-pairs per library (millions)	Panel size (kb)
ROC	Illumina NextSeq 500	High-output	NA*	2 x 151	60	192
ILM	Illumina NovaSeq 6000	S4	6	2 x 151	225	533
IDT	Illumina NovaSeq 6000	S4	24	2 x 151	69	140**
BRP	Illumina NovaSeq 6000	S4	24	2 x 151	93	227

SUPPLEMENTARY METHODS

Five industry-leading NGS-based ctDNA assays were enrolled in the SEQC2 proficiency study. This included hybrid-capture gene panels from Roche Sequencing Solutions (ROC), Illumina (ILM), Integrated DNA Technologies (IDT) and Burning Rock Biotech (BRP), and an amplicon sequencing panel from Thermo Fisher Scientific (TFS). Each participating assay was performed at 2-3 independent test labs, according to the vendor's instructions, and each sequencing library then analyzed by the relevant assay vendor. Bioinformatic analysis was not standardized across the study, with each vendor employing an internal analysis pipeline, and providing a final set of variant candidates for centralized evaluation by an independent team.

Here we provide detailed information on the experimental and bioinformatic procedures employed for each participating assay.

ROC: Roche Sequencing Solutions, AVENIO ctDNA Expanded Kit

Experimental procedure

The AVENIO ctDNA Expanded Kit (For Research Use Only; not for use in diagnostic procedures) is a hybridization-based workflow requiring only DNA, allowing the detection of Single Nucleotide Variations (SNVs), Insertions and Deletions (Indels), Fusions, and Copy Number Variants (CNVs). Prior knowledge of the fusion breakpoint is not required, since the hybridization method targets whole introns of the genes of interest. In brief, the extracted cell-free DNA sample is initially ligated with adapters containing unique molecular identifiers, which allows for the deduplication of the eventual sequencing reads back to the original input molecules, significantly reducing undesired errors. After the ligation, PCR is used to universally amplify the ligated material; gene enrichment does not occur during the PCR. The sample is then incubated overnight with the gene panel, consisting of biotinylated probes designed for optimal enrichment of the genes of interest. The desired DNA-probe complexes are then captured on streptavidin beads, and after a series of washes, the samples are PCR-amplified. The final product of the workflow is enriched libraries ready for sequencing. The final sequencing libraries were sequenced using the Illumina NextSeq 500 sequencing platform. Sequencing results were analyzed by the AVENIO ctDNA Analysis Server v1.1 (for Research Use Only; not for use in diagnostic procedures).

Bioinformatics procedure

The AVENIO ctDNA analysis pipeline is commercially available in the Oncology Analysis Server v1.1 (for Research Use Only; not for use in diagnostic procedures). It reads lane-level (BCL format) data from a run of plasma samples on a NextSeq500, converts to raw sequence data by Illumina's bcl2fastq program, and then demultiplexes the data per sample using the specific sample adapter sequence ligated during its sample prep step. Each sample's genomic differences (variants) compared to the human reference genome (hg38) are identified using genomic alignments (BAM format), which are processed by the downstream variant calling algorithms. Sample-level variants that can be detected by the pipeline are SNVs, indels, fusions, and CNVs.

The workflow generates sample level variant calls in variant call format (VCF) for SNVs and indels, and bed file format for CNVs and fusions. The specifications for variant calling are 0.5% LOD for SNVs, indels and fusions and 2.3-4.5 copy number LOD for CNVs. The SNV caller detects variants in a set of 'hotspot' ("Loci of Interest") positions and across the entire panel. However, the CNV and fusion caller are restricted only to a predetermined set of genes: ERBB2, MET and EGFR for copy number amplifications, and a list of 6 fusions genes and their common partners for fusion detection. For indels, the calls are restricted mainly to a set of hotspot ("Loci of Interest") positions except for EGFR exon 19 long deletions, EGFR exon 20 long insertions and MET long insertions, which are not restricted to a predetermined set of known indels.

SNV Caller The SNV caller generates base count files from genome alignment and MID (molecular ID) deduplication of reads. MID deduplicated base count files then undergo background polishing that removes recurrent library preparation errors inferred from control samples[21,22]. The calls are made in two modes: one is Adaptive caller, where the error distribution of each of the twelve substitution types (A>C, A>G, … T>G) is modeled in the sample and sample-specific substitution-specific depth thresholds are set. The other component of SNV caller is Hotspot caller where simpler heuristic rules are used for the variants listed in the

Loci of Interest[21,22]. The Oncology Analysis Software carries out SNV post-filtering to restrict SNV calls to exons and splice sites, and filter out likely germline changes (from databases like dbSNP, ExAC), and keep changes previously observed to be somatic (presence in TCGA, COSMIC). The caller also flags potential cross-contamination across the lane. It flags variants in a lane that are low AF in one sample but high AF in another, which may indicate potential cross contamination.

Indel Caller The indel caller in the AVENIO pipeline reports the highest AF across three tools with slightly different approaches to indel calling, including one in-house proprietary algorithm. The in-house caller processes data from position deduped bam files. It calls hotspot indels ("Loci of Interest") if the highest AF from the 3 callers is >= 0.1%. The hotspot list has approximately 70 indels that are recurrently seen in solid tumors. For non-hotspot mode, long indels (> 6 bp) are called even if they are not in the Loci of Interest.

ILM: Illumina, TruSight Tumor 170 + UMI

Experimental procedure

Libraries were prepared using the TruSight Tumor 170 Reference Guide, with modifications outlined in the TruSight UMI toolkit reference guide. Briefly, DNA samples, provided as enzymatically fragmented material to mimic cfDNA were end-repaired and A-tailed in a single reaction, followed by ligation to an universal adapter containing unique molecular identifiers (UMI) to uniquely tag each molecule going into the library preparation. Post-ligation clean-up was performed using SPRI beads and then libraries were indexed using unique dual indexes by PCR. Target regions were captured using an overnight hybridization to biotinylated target-specific oligos which covers ~533 Kb of genomic targets across 154 genes, followed by capture with streptavidin magnetic beads. A second hybridization and capture reaction were performed followed by PCR amplification using the universal primers compatible with the sequencing flowcell. Libraries were quantified and manually normalized to 6nM before being pooled in equal parts per library. Libraries were then further diluted and loaded using the Xp workflow on a NovaSeq 6000 S4 flowcell, with 6 libraries per lane on the flowcell. Sequencing was performed as 2 x 151 bp with 8 bp dual-indexed reads.

Bioinformatics procedure

The libraries prepared from enzymatically fragmented DNA were processed using Illumina's standard internal pipeline with a few modifications to remove artifactual variants that were present at low levels in *Sample B*. Briefly, reads were demultiplexed, trimmed of adaptors, and converted into the FASTQ format using bcl2fastq. Reads were then aligned to the human genome version hg19 using the BWA mem. Using this initial alignment, duplicate reads were collapsed using the unique molecular identifiers (UMIs) attached to each fragment. Collapsed reads were realigned to hg19 to rescue reads that had been error-corrected during read collapsing. Indel realignment was performed and overlapping reads from short DNA fragments were stitched into a single read, with appropriate error correction. Candidate variants were then identified using the Pisces variant caller (https://github.com/Illumina/Pisces), adjusting default parameters to output candidate variants with a variant allele fraction of at least 0.01%. Candidate variant calls are then evaluated using two likelihood models. The first model empirically re-estimates the error rates associated with collapsed reads based on the strands of the reads they were derived from and the number of times each base in the fragment was sequenced by a collapsed read. The second error model attempts to capture errors associated with alignment by estimating the rate of somatic mutations observed in healthy individuals at different positions of the genome. Candidate variants with a phred-based quality score of less than 40 in either model were removed.

To remove low allele fraction artifacts introduced by the cell line in *Sample B*, a blacklist was generated by running all *Sample B* replicates from all three test sites through the pipeline described above. Any position with a variant with quality scores of greater than 20 in both likelihood models described above in any sample was added to the blacklist and further excluded. Finally, 18,588bp that consistently performed below our depth requirements were excluded from consideration (~3% of the total panel size).

IDT: Integrated DNA Technologies, xGen Non-small Cell Lung Cancer ctDNA assay

Experimental procedure

Sample Lbx-low-plasma was purified using the QIAamp® Circulating Nucleic Acid kit and quantified according to the methods described in the SEQC2 WG2 Sample Processing and Sequence Data Reporting SOP. Libraries were constructed using mock cfDNA samples in quadruplicate using the KAPA Hyper Prep Kit and IDT custom adapters. End repair and A-tailing were performed according to the manufacturer's recommendations. For adapter ligation, 3 µM, 7.5 µM, and 15 µM stocks were used for 10 ng, 25 ng, and 50 ng input samples. Libraries were purified using 0.8X AMPure and amplified using unique dual index primers with 10, 9 and 8 cycles of PCR for 10 ng, 25 ng, and 50 ng input samples. Libraries were purified using 1X AMPure and quantified using Qubit. 500 ng of each library was captured with a custom NSCLC xGen Lockdown® Probe Panel using the xGen Universal Blockers–TS Mix. After enrichment, libraries were amplified with the KAPA HiFi HotStart ReadyMix using 13 cycles for amplification. Post-capture libraries were purified with 1.5X AMPure, quantified, and pooled for sequencing on the Illumina NovaSeq S4.

Bioinformatics procedure

IDT libraries were prepared with IDT custom adapters which contain 3 bp degenerate unique molecular identifiers (UMIs). Picard v2.18.9 IlluminaBasecallsToSam was used to demultiplex BCL files and generate unmapped bam files. The in-line UMIs were extracted using fgbio v0.7.0 ExtractUmisFromBam with the read structure 3M2S146T 3M2S146T, and --molecular-index-tags=ZA ZB, --single-tag=RX parameters. Illumina adapters sequences were marked using Picard v2.18.9 MarkIlluminaAdapters and trimmed in Picard v2.18.9 SamToFastq when generating FASTQ files for alignment. FASTQ files were mapped to hg19 using bwa-mem v0.7.15, and Picard v2.18.9 MergeBamAlignment was used to generate a mapped BAM with the UMI metadata (ZA, ZB, and RX tags), using the unmapped bam files from fgbio v0.7.0 ExtractUmisFromBam. Reads originating from the same source molecule were identified using fgbio v0.7.0 GroupReadsByUmi, which assigns a unique source molecule ID to each applicable read allowing up to 1 edit distance, stores the ID in the MI tag, and outputs a BAM file that is sorted by the MI tag to combine read families. Unmapped collapsed combined reads were generated using fgbio v0.7.0 CallDuplexConsensusReads with the following parameters --error-rate-pre-umi=45, --error-rate-post-umi=30, and --min-input-base-quality=30, and collapsed combined reads were remapped to hg19 using bwa-mem v0.7.15. Read-level filtering was applied using fgbio v0.7.0 FilterConsensusRead with --min-read=2 1 1 to build combined read families, while masking bases where any single read families disagree. To evaluate target enrichment performance, Picard v2.18.9 CollectHsMetrics was used. 10ng samples with < 500x median target coverage, and other samples with < 2000x median target coverage was flagged as outliers. Overlapping reads were hard clipped using fgbio v0.7.0 ClipBam to prevent double counting evidence for downstream variant calling. Supplementary aligned reads and not primary aligned reads were removed using samtools v1.5. Variant calling was performed on the clipped collapsed combined read families BAM using AstraZeneca VarDict v1.5.8 with an allele frequency threshold of 0 and a minimum of 3 alt reads. Low frequency mutations that were called in the *Sample B* replicates were removed in all other samples, and mutations flagged as p8 in the filter column were removed.

BRP: Burning Rock Biotech, Lung Plasma v4 ctDNA assay

Experimental procedure

Sample Lbx-low-plasma was extracted using the QIAamp Circulating Nucleic Acid kit (Qiagen) according to the manufacturer's instruction. After extraction, Ep concentration was quantified using Qubit 3.0 Fluorometer and concentration adjustment was performed following the organizers' recommendation. The library prep and enrichment process were performed using Burning Rock HS UMI library preparation kit without modification. In brief, pre-fragmented SEQC2 DNA samples were end repaired, UMI adapter ligated and PCR enriched. About 1 µg of purified pre-enrichment UMI library were hybridized to LungPlasma^{TM} panel and further enriched following manufacturer instruction. The LungPlasma^{TM} panel is about 250 Kb in size, and covers 168 human lung cancer related genes. Final DNA libraries were quantified using Qubit Fluorometer with dsDNA HS assay kit (Life Technologies, Carlsbad, CA). A LabChip GX Touch System, Agilent 2100 bioanalyzer or Agilent 4200 TapeStation D1000 ScreenTape was then performed to assess the quality and size distribution of the library. The libraries were sequenced on NovaSeq 6000 sequencer (Illumina, Inc., California, US) with 2×150bp pair-end reads with unique dual index.

Bioinformatics procedure

After demultiplex and moving 6-bp UMI to the sequence header using bcl2fastq v2.20 (Illumina), sequence data in FASTQ format were filtered using the Trimmomatic 0.36 with parameters "HEADCROP:2 SLIDINGWINDOW:8:20 MINLEN:50". (HEADCROP: Cut the specified number of bases from the start of the read).

Sequencing reads were initially mapped to the human genome (hg19) using BWA aligner 0.7.10. The consensus BAM file were then created using homebrew software based on UMI sequence and read alignment position. The consensus BAM contains reads with simplex consensus and duplex consensus. The simplex consensus was derived from reads originated from one initial DNA strand. The duplex consensus was derived from reads originated from both strands of the initial DNA fragment. Reads without consensus generation capability were discarded.

VarScan v2.4.3 with parameters "--min-coverage 100 --min-var-freq 0.00001 --min-reads2 1 --output-vcf 1 --strand-filter 0 --variants 1 --p-value 0.99" was used to create initial VCFs using consensus BAM. For SNV and short InDel, variants were further filtered using the homebrew variant filter pipeline. For each valid variant, the covered depth must be greater or equal than 100 (DP>=100); and at least 1 or 2 mutation supporting count (AD>=1 or 2) for hot and other mutation, respectively; mutation allele frequency must be greater than 0.0005 and 0.001 (AF>=0.0005 or 0.001) for hot and other mutations, respectively. In order to filter out further false-positives, only variants fulfilling both the following criteria were retained: (1) at least 2 supporting simplex or duplex consensus fragments; (2a) with at least 1 duplex consensus fragment and the mutation is being read in both paired end read; or (2b) with at least 2 duplex consensus support. Each remaining variant was annotated with ANNOVAR 20160201 and SnpEff v3.6. Low frequency mutations that were presented in the fragmented *Sample B*, likely caused by enzymatic fragmentation process, were filtered in all *Lbx-high* and *Lbx-low* samples.

TFS: Thermo Fisher Scientific, Oncomine Lung Cell-Free Total Nucleic Acid Research Assay

Experimental procedure

For samples that required nucleic acid extraction, the MagMAX™ Cell-Free Total Nucleic Acid Isolation Kit (https://www.thermofisher.com/order/catalog/product/A36716) (Cat. No. A3716) was used and extraction was carried out according to manufacturer's instructions.

Sequencing libraries were constructed according to manufacturer specifications found in Oncomine™ Lung cfTNA assay User Guide (https://www.thermofisher.com/order/catalog/product/A35864)

Included in the user guide is the protocol for constructing and templating sequencing libraries using the Ion Chef™ Instrument (Cat. No. 4484177). Subsequently, each library was loaded on to an Ion 530™ chip & Ion 530™ Kit – Chef (Cat. Nos. A27757, A30010) which was then loaded on to Ion S5™ XL (Cat No. A27214) next-generation sequencing system. Each sequencing library has a sample specific TagSequencing barcode (Tag Sequencing Barcode Set 1-24, Cat. No. A31830) attached to each amplicon to enable identification of individual sample which has been pooled with other multiplexed samples loaded on an Ion 530™ chip.

Bioinformatics procedure

Signal processing and base calling were performed using Torrent Suite Software v5.8 using default parameters for the Oncomine TagSeq Liquid Biopsy. The signal processing step consists of modeling the pH dynamics on the semiconductor surface taking account of the varying local pH in each individual sensor coming from the different reagent flows across the chip and from any nucleotide incorporation that may be happening over each sensor. The base calling step consists of taking the estimated levels of nucleotide incorporation for each read and each nucleotide flow, and modeling the de-phasing process whereby some templates within each clonally-amplified population run ahead or behind in terms of their nucleotide incorporation. During the base calling process, sample-specific barcodes and 3' adapters are annotated. Once sequencing was complete, within TSS, resulting sequencing reads are mapped to the hg19 build of the human genome. Subsequently, consensus reads are built by binning read sets with common molecular tags.

After completion of primary analysis with Torrent Suite v5.8, reads were uploaded to Ion Reporter v5.6 for subsequent processing. Data were processed using a workflow specifically pre-tuned for the Oncomine TagSeq Lung v2 Assay. A consensus flowspace signal was generated for reads sharing the same molecular tags. Reads were aligned with TMAP that uses the BWA fastmap routine to map reads and applies post-processing of the alignments to optimize for technology-specific error patterns. After alignment, variant calling was performed with Torrent Variant Caller (TVC), a variant calling framework optimized for Ion Torrent data.

For the Oncomine™ Lung cfTNA, TVC takes as input the aligned reads and a list of pre-defined hotspot alleles representing variants known to be highly recurrent in cancers. All hotspot alleles are evaluated in a statistical likelihood model that compares the consensus flow signals for all of the aligned reads with the flow signals that would be expected under reference and non-reference hypotheses. The use of flow signals leads to significant improvements in variant calling compared to variant calling approaches that rely on base calls alone. The lung cfTNA assay has a default minor allele frequency threshold of 0.035% for SNV / Indel variants with a minimum molecular coverage of 2 templates harboring a putative mutant allele. No variants are called below the thresholds and variants can be called just above the threshold if the statistical evidence is sufficiently strong. Panel design also includes known systematic error positions, (also known as 'blacklist' positions), that are masked during the variant calling. Finally, a series of post-calling filters are applied to variant calls to filter out situations where the statistical model of flow signals is not a good fit for the observed data.

Assessment by AcroMetrix synthetic DNA control

To further evaluate the detection of low-frequency mutations by amplicon sequencing, TFS test sites also tested the AcroMetrix Oncology Hotspot Control at 50 ng input. The AcroMetrix synthetic DNA constructs (https://www.thermofisher.com/order/catalog/product/969056#/969056) were spiked in *Sample B* (Agilent Male Control DNA as background gDNA) at equal amount to achieve an allele frequency at 0.1% of each variants. Consistent with other test samples, the spike-in control underwent enzymatic fragmentation, size selection, and quantification, and was further aliquoted at the same concentration (5 ng/µL) for distribution to TFS test sites (see **Methods** for details). Four library replicates were generated for the control sample and sequenced at each test site. The AcroMetrix synthetic constructs contain 521 known cancer mutations in 52 genes, of which 26 overlapped TFS hotspot regions. However, pre-filtering was required to remove variants from un-callable positions within the panel design. Spike-in variants that overlap with blacklist positions were first excluded. Some spike-in variants may fall into the amplicon primer regions, which interfered with the amplification of the variant fragments and resulted in a bias amplification of only the background gDNA sequence. Therefore, the spike-in variants that fall within such amplicons were identified and excluded from the sensitivity assessment. After excluding un-callable spike-in variants, 15 synthetic AcroMetrix variants were used for further analysis.