**TECHNICAL REPORT**

**High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants**

Pedram Razavi1,2,9, Bob T. Li1,9, David N. Brown3,9, Byoungsok Jung4, Earl Hubbell4, Ronglai Shen5, Wassim Abida1, Juluru Krishna6, Ino De Bruijn7, Chenlu Hou4, Oliver Venn4, Raymond Lim3, Anand Aseen1, Tara Maddala4, Sante Gnerre4, Ravi Vijaya Satya4, Qinwen Liu4, Ling Shen4, Nicholas Eattock4, Jeanne Yue4, Alexander W. Blocker4†, Mark Lee4§, Amy Sehnert4¥, Hui Xu4, Megan P. Hall4, Angie Santiago-Zayas1, William F. Novotny4\*, James M. Isbell8, Valerie W. Rusch8, George Plitas8, Alexandra S. Heerdt8, Marc Ladanyi3, David M. Hyman1, David R. Jones8, Monica Morrow8, Gregory J. Riely1, Howard I. Scher1, Charles M. Rudin1, Mark E. Robson1, Luis A. Diaz, Jr.1, David B. Solit1,2,7, Alexander M. Aravanis4, Jorge S. Reis-Filho2,3

**Affiliations**

1Memorial Sloan Kettering Cancer Center, Department of Medicine, New York, NY.

2Memorial Sloan Kettering Cancer Center, Human Oncology and Pathogenesis Program, New York, NY.

3Memorial Sloan Kettering Cancer Center, Department of Pathology, New York, NY.

4GRAIL, Inc. Menlo Park, CA.

5Memorial Sloan Kettering Cancer Center, Department of Epidemiology and Biostatistics, New York, NY.

7Memorial Sloan Kettering Cancer Center, Marie-Josée and Henry R. Kravis Center for Molecular Oncology, New York, NY.

8Memorial Sloan Kettering Cancer Center, Department of Surgery, New York, NY.

9Equal Contribution

†Current affiliation: Foresite Capital Management, San Francisco, CA.

§Current affiliation: Genentech, Inc., South San Francisco, CA.

¥Current affiliation: MyoKardia, Inc., South San Francisco, CA.

\*Current affiliation: BeiGene, Ltd., San Mateo, CA.

To whom correspondence should be addressed: [razavip@mskcc.org](mailto:razavip@mskcc.org), [reisfilj@mskcc.org](mailto:reisfilj@mskcc.org)

**ABSTRACT**

Accurate identification of tumor-derived somatic variants in plasma circulating cell-free DNA (cfDNA) requires understanding the various biologic compartments contributing to the cfDNA pool. We sought to define the technical feasibility of a high-intensity sequencing assay of cfDNA and matched white-blood cell (WBC) DNA covering a large genomic region (508 genes, 2Mb, >60,000x raw-depth) in a prospective study of 124 metastatic cancer patients, with contemporaneous matched metastatic tumor tissue biopsies, and 47 non-cancer controls. The assay displayed a high sensitivity and specificity, allowing for *de novo* detection of tumor-derived mutations and inference of tumor mutational burden, microsatellite instability, mutational signatures and sources of somatic mutations identified in cfDNA. The vast majority of cfDNA mutations (81.6% in controls and 53.2% in cancer patients) had features consistent with clonal hematopoiesis (CH). This cfDNA sequencing approach robustly detected somatic mutations in plasma and revealed that CH constitutes a pervasive biological phenomenon emphasizing the importance of matched cfDNA-WBC sequencing.

**INTRODUCTION**

The presence of circulating cell-free DNA (cfDNA) in peripheral blood was initially detected over 70 years ago1. Subsequent early studies demonstrating higher levels of cfDNA in the plasma of cancer patients drew attention to cfDNA as a potential source of tumor-derived DNA2,3. Massively parallel sequencing analysis of cfDNA samples from cancer patients revealed that tumor-derived cfDNA (ctDNA) accounts for only a fraction of the total cfDNA, and this fraction varies according to disease burden, site, and tumor biologic features including histology, tumor vascularization, and proliferation and apoptosis rates4,5. ctDNA fraction is extremely low in many early-stage and some metastatic cancers6,7, therefore, methods for detecting plasma ctDNA must detect mutations at extremely low allele fractions8. Even in metastatic cancers with high disease burden and high ctDNA fractions, it is often necessary to identify subclonal mutations (those altered in only a small fraction of cancer cells within a patient) present at very low allele fractions. Most previous studies focused on analysis of patients with advanced disease using a panel of hotspot mutations or limited genomic regions of key cancer genes sequenced at high depths9-11, a large number of genes at sequencing depths comparable to those employed for analysis of tumor biopsies12,13, or a combination of methods to define ctDNA fraction using shallow whole-genome sequencing or PCR-based methods followed by whole-exome analysis of samples with a high ctDNA fraction7,14.

Even when accurate cfDNA assays are utilized, cfDNA sequencing results may still be confounded by biological signals arising from somatic mosaicism: somatic non-tumor-derived variants that non-malignant cells acquire through aging, cell divisions and/or as a result of encountering intrinsic or extrinsic mutagens15. One common form of somatic mosaicism is clonal hematopoiesis (CH), which results from the accumulation of somatic mutations in hematopoietic stem cells (HSCs) that are clonally propagated to the subsequent progeny from these stem cells16. These somatic mutations are reported to provide a fitness advantage to some HSCs and/or their descendant cells, resulting in their disproportionate expansion9,17-19; however, reports have also suggested that CH may arise through neutral drift20. CH increases with age and occurs in up to 31% of older individuals11,19,21-25. Importantly, CH can also be detected in cfDNA sequencing analysis26 and, in this context, confound the interpretation of cfDNA sequencing, particularly because a large proportion of the cfDNA fragments originates from hematopoietic cells27.

Previous studies comparing somatic genetic alterations detected in cfDNA samples and their respective tumor biopsies have revealed relatively good concordance between cfDNA and tumor biopsy sequencing, particularly among patients with advanced disease7,10,28-31. Additional somatic variants not present in tumor biopsies have also been documented5. Further work is required to define the nature and source of these additional somatic variants detected in cfDNA (tumor-derived vs. other sources) and to allow for accurate cfDNA sequencing as near as possible to the molecular limits of detection.

Here we report on the development of a high-intensity sequencing assay of matched cfDNA and white blood cells (WBCs) for *de novo* characterization of the repertoire of somatic mutations in cfDNA, without *a priori* knowledge of variants present in a matched tumor biopsy. This approach, combined with sequencing of DNA samples extracted from matched tumor tissue biopsies using an FDA-cleared sequencing assay (MSK-IMPACT), allowed for categorization and quantification of cfDNA variant sources.

**RESULTS**

## Study design and demographic information

This prospective observational study examined the technical feasibility of a high-intensity circulating cfDNA-based platform in patients with advanced untreated or progressivemetastatic breast cancer (MBC), non-small cell lung cancer (NSCLC), or castration-resistant prostate cancer (CRPC), as well as non-cancer control participants (**Methods**). Briefly, plasma cfDNA and matched WBC genomic DNA (gDNA) from patients with MBC, NSCLC, CRPC, or non-cancer controls were subjected to a targeted capture sequencing assay comprising the entire coding region of 508 genes and intronic and/or regulatory regions of selected genes (**Fig. 1a, Supplementary Table 1**). In cancer patients, tumor biopsies and matched normal WBC samples were collected within 6 weeks of plasma cfDNA samples with no intervening therapy change, and were sequenced using the Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT), a Food and Drug Administration-authorized capture-based sequencing assay targeting the coding regions of 410 genes and intronic and/or regulatory regions of selected genes (**Fig. 1a, Supplementary Table 1)**32,33. For the purpose of comparison to tumor biopsies, only variants mapping to the intersection of the 410 genes present in the two gene panels were considered.

Of 161 eligible cancer patients (53 MBC, 53 NSCLC and 55 CRPC) enrolled between September 24, 2015-August 01, 2016, 124 (39 MBC, 41 NSCLC and 44 CRPC) were included in the concordance subset (evaluable for both tumor tissue and cfDNA analysis, **Supplementary Fig. 1**). Of the 50 non-cancer control samples, three failed quality control due to incorrect plasma pooling during cfDNA extraction (n=2) and cross-contamination detected by the quality control pipeline (n=1), resulting in 47 evaluable samples. The baseline demographic characteristics of the cohort are presented in **Supplementary Table 2**. Among evaluable patients in the MBC cohort, the median age was 60 (range 30-79), 26 (67%) were hormone receptor-positive and HER2-negative, and 32 (82%) had invasive ductal carcinoma. In the evaluable metastatic NSCLC cohort, the median age was 67 (range 33-83), 28 (68%) were female, 38 (93%) were adenocarcinomas, and 28 (68.3%) had M1b (extrathoracic metastases) disease. The median age of evaluable CRPC patients was 67 (range 46-87), and 39 (89%) were adenocarcinoma. Overall, the majority of patients enrolled in the study received prior treatment in the (neo)adjuvant and/or metastatic settings (85% in MBC, 41% in NSCLC, 100% in CRPC): 38% of the MBC patients and 10% of the NSCLC patients had received at least three lines of therapy in the metastatic setting, and all CRPC patients had disease progression after initial castration therapy with or without androgen receptor antagonists, with 32 (78%) also receiving additional systemic therapy prior to sample collection.

## De novo detection of tumor-derived cfDNA mutations

To identify the source of somatic variants found in cfDNA, sequencing was performed independently on cfDNA, WBC gDNA, and each cancer patient’s matched tumor biopsy and WBC gDNA samples (**Methods, Fig. 1a**). The high-intensity cfDNA sequencing approach simultaneously analyzed plasma cfDNA and WBC gDNA using a targeted DNA assay spanning approximately 2 Mb and utilizing unique molecular identifier (UMI) sequences to suppress technical assay errors at a minimum average raw coverage depth of 60,000X (**Supplementary Fig. 2**). A joint-variant-calling of plasma cfDNA and WBC gDNA variants was performed utilizing a machine learning-based noise model (**Supplementary Methods**). Together, this resulted in an assay with performance the**1b-c**low false positive (; **Supplementary Fig. 3**)high**d-e, Supplementary Fig. 4**) and a detection performance comparable to that of digital droplet PCR (ddPCR; **Fig. 1f**). Our high-intensity sequencing assay was found to have a favorable per base error rate (**Supplementary Methods,** **Supplementary Figs. 2-3, Supplementary Tables 3-5**) ranging from 1x10-5 to 3x10-5 compared to other high-fidelity cfDNA sequencing assays, such as 2x10-5 for Integrated Digital Error Suppression (iDES) (PMID: 27018799), 0.9x10-5 for Safe-Sequencing System (Safe-SeqS) (PMID: 21586637), or <1.5 variants detected per age-matched control sample in approximately 1x106 bases sequenced per experiment compared to <1 error per 1.56x106 bp for Guardant G360 (PMID: 26474073) and <1 false positive per 3x106 bp sequenced for targeted error correction sequencing (TEC-Seq) (PMID: 28814544).

We first sought to define the performance of the cfDNA sequencing assay for the detection of tumor-derived biopsy-matched somatic mutations without prior knowledge of the somatic alterations in the tumor cfDNA sequencing analysis (henceforth referred to as *de novo* detection). *De novo* detection of at least one tumor-derived mutation, as defined by MSK-IMPACT sequencing of the tumor biopsy, was observed in 104 of the 124 evaluable patients (84%, 95% confidence interval [CI]: 76%-90%; **Fig. 2a**). The detection rate in MBCs (95%, 95% CI: 83%-99%) was significantly higher than in NSCLCs (76%, 95% CI: 60%-80%; p = 0.0258), and comparable to that of CRPCs (82%, 95% CI: 67%-92%). This assay’s large genomic footprint further allowed *de novo* cfDNA detection of 530 of 740 mutations detected by MSK-IMPACT in tumor biopsies (72%, 95% CI: 66%-75%; **Fig. 2a**), with similar percentages of tumor-derived mutations detected in MBCs (73%, 95% CI: 67%-79%), NSCLCs (71%, 95% CI: 65%-76%), and CRPCs (71%, 95% CI: 63%-78%).

We next sought to define the genes targeted by *de novo-*detected somatic mutations in cfDNA. Our analysis of genes recurrently mutated in cfDNA as defined by the *de novo*-detected somatic mutations revealed that they consisted mostly of the same genes found to be mutated in the respective tumor samples (**Fig. 2b** and **Supplementary Fig. 5**). Most importantly, this led to *de novo* detection of somatic mutations in cfDNA that were present in tumor biopsies but below the MSK-IMPACT assay limit of detection (subthreshold for previously established clinical variant calling cut-offs)32,34, or were neither detected in the tumor biopsy nor WBCs (variants of unknown source; VUSo).

Given the low false positive rates (**Fig. 1d-e**) and the assay’s accuracy for measuring variant allele fractions (VAFs; **Fig. 1d-f**), we quantified the VAFs of somatic mutations not present in the WBCs in controls and cancer patients. All but two of the 67 mutations (97%) detected in controls were detected at VAFs of <1% (**Fig. 2c**), whereas 51.1%, 56.6%, and 54.5% of the variants detected in MBCs, NSCLCs, and CRPCs, respectively, were detected at VAFs of <1%. In the vast majority of patients (88%), somatic mutations with the highest VAF (mean highest VAF 15.10%; median highest VAF 9.18%) were tumor-matched (biopsy-matched or biopsy-subthreshold; **Fig. 2d**).

We next investigated whether the sensitivity of the assay would vary according to the prevalence of a given mutation within the tumor biopsy. The detection rate of mutations in cfDNA was significantly correlated with their cancer cell fractions in the tumor biopsies (the percentage of cancer cells within a biopsy inferred to harbor a given mutation in a particular patient; **Online Methods**), with mutations defined as clonal being significantly more frequently detected than subclonal ones (p<1e-4; **Fig. 2e**). Additionally, the ctDNA fraction (the fraction of tumor-derived cfDNA) in metastatic cancer patients varied significantly according to tumor type (p=4.6e-3; **Fig. 2f**). To assess the association between disease burden and disease fraction, volumetric assessment of the disease burden was performed (**Methods**). We found a significant association between the estimated disease volume in ml and the ctDNA fraction in both MBC (n=34, p=1.03e-4) and NSCLC (n=29, p=0.042; **Fig. 2g**) patients. An assessment of the association between the automated bone scan index (aBSI; **Methods**) and the ctDNA fraction also revealed a significant association (n=39, p = 0.023; **Fig. 2g**).

Taken together, these analyses demonstrate that this cfDNA sequencing assay has high sensitivity for detection of tumor-derived somatic mutations and identifies mutations subclonally present in tumor tissues. As further described below, VUSo detected in neither tumor biopsy nor WBC may derive from multiple origins but comprise a set of alterations from which a subset may reflect aspects of ongoing tumor evolution and heterogeneity not captured in a small and anatomically constrained biopsy.

## Tumor mutation burden and mutational signatures

Besides the identification of mutations, sequencing-derived data have been used to assess mutational burden, mutational signatures or genomics patterns consistent with microsatellite instability (MSI) and have been shown to provide information that can guide therapeutic decisions35,36. Therefore, we performed an exploratory analysis to determine whether tumor mutation burden (TMB), mutational signatures37,38, and MSI score39 could be defined solely using cfDNA data.

Analysis of TMB by MSK-IMPACT sequencing of tumor biopsies revealed six samples with a high TMB (based on the predefined >13.8 mutations/Mb cut-off)32, including two MBCs, three NSCLCs and one CRPC. Four of these cases were also classified as having high TMB by cfDNA (>22.7 mutations/Mb; **Methods**). The remaining two samples displayed relatively low ctDNA fractions (0.2 and 8.6%) and borderline cfDNA TMB (18.2 and 20.0 mutations/Mb, respectively). Importantly, cfDNA analysis identified six additional cases with a high TMB (**Fig. 3a**, a total of ten cases henceforth referred to as hypermutated) not detected as hypermutators by MSK-IMPACT analysis of the tumor biopsy. Potential explanations for this observation include spatial tumor heterogeneity between metastatic sites with only some sites demonstrating a hypermutator phenotype. The ten hypermutated cfDNA samples accounted for 75% of the cfDNA biopsy subthreshold mutations and VUSo across the entire cohort (**Fig. 3b**) and displayed mutational signatures consistent with the modalities of genetic instability documented in MBCs, NSCLCs, and CRPC. All hypermutated MBCs (n=5) and one of the three hypermutated CRPCs displayed the mutational signature related to activity of the APOBEC cytidine deaminases37,38, which generate a hypermutator phenotype and is commonly acquired late in the evolution of MBCs and CRPCs40. Consistent with the results of previous analysis of NSCLCs41, the mutational signatures of the two hypermutated NSCLCs comprised the smoking-related signature and a combination of other mutational signatures, including APOBEC, homologous recombination DNA repair-deficiency (HRD), and loss-of-function of mismatch repair (MMR; **Fig. 3c** and **Supplementary Fig. 6**).

High microsatellite instability (MSI-H), stemming from loss of MMR, has been validated as a biomarker of response to immune-checkpoint inhibitors42. We therefore assessed the MSI status of the cohort utilizing MSIsensor39, adjusted for the ultra-high sequencing depth of cfDNA, tumor biopsy. and matched normal WBCs (**Methods, Supplementary Fig. 7**). Our cfDNA analysis revealed one CRPC with genomics features of MSI-H (**Fig. 3d**), which was also found to display a dominant MMR mutational signature (**Fig. 3c**). Consistent with the notion that MSI is a predictor of response to immune-checkpoint inhibitors, this CRPC patient32 who had previously been treated with enzalutamide received an anti-PD-L1 inhibitor and displayed rapid and sustained tumor regression, as defined by the response evaluation criteria in solid tumors (RECIST v1.1) and prostate-specific antigen (PSA) serological levels (**Fig. 3e**). Taken together, these results suggest that this cfDNA sequencing assay may accurately detect tumor-derived mutations across a large portion of the genome, potentially allowing for the characterization of tumor mutation burden, MSI status and mutational signatures.

## Characterization of the biological sources of cfDNA variants

Considering the cfDNA results alone, despite the specificity of the assay for somatic mutation detection, tumor-matched alterations (biopsy-matched and biopsy-subthreshold) accounted for only 24.4% (739 of 2983 mutations) of all somatic mutations detected in the cfDNA of cancer patients (**Fig. 4a**). In fact, a median of 7.27 (range 0.91-20.91) mutations per Mb were detected in the cfDNA samples of non-cancer controls (**Fig. 4a**). Although previous studies have suggested that these alterations likely constitute technical artifacts of ultra-high-depth sequencing analysis43,44, based on the specificity of this assay we posited that these variants instead stemmed from somatic mosaicism, in particular CH, and tumor-derived events resulting from spatial genetic heterogeneity (as seen in the hypermutated cancer cases).

To address this hypothesis, we first investigated the presence of somatic mutations in the WBC sequencing results (WBC-matched mutations) for the mutations defined as somatic by cfDNA analysis but were not biopsy-matched. This analysis revealed that in non-cancer controls, the vast majority (81.6%, 297 of 364) of somatic mutations detected were also identified in WBCs, suggesting that these somatic genetic alterations were likely not technical artifacts but rather a result of CH (**Fig. 4a**). Likewise, the majority (53.2%, 918 of 1727) of the mutations identified in cfDNA samples of non-hypermutated cancer patients were also WBC-matched (**Fig. 4a**). Importantly, the number of WBC-matched cfDNA variants in cancer patients did not correlate with the number of tumor-matched mutations (biopsy-matched or biopsy-subthreshold), making them less likely to be of tumor origin (**Fig. 4b, Supplementary Fig. 8**). As CH is related to age22, we examined the association of age with the number of somatic DNA variants in the cfDNA in samples from individual participants. As expected, the number of WBC-matched variants, but not of the biopsy-matched or biopsy-subthreshold variants, significantly correlated with age (smoking-adjusted p = 7.4e-41; **Fig. 4c**). Based on this interpretation, the cfDNA and WBC sequencing analysis as performed here suggests that 89.5% of cancer patients and 83% of non-cancer controls have evidence of CH in their cfDNA (**Fig. 4b**). Consistent with recent observations26 and with the notion that these mutations constitute CH events, the vast majority of the WBC-matched somatic mutations detected in cfDNA involved canonical CH genes, such as *DNMT3A, TET2, PPM1D* and *TP53* (**Fig. 4d**, **Supplementary Fig. 9**)11, while some of these also included pathogenic cancer alterations (**Supplementary Table 6**). Additionally, the VAFs of WBC-matched cfDNA variants were significantly correlated with their VAFs in the WBCs (**Fig. 4e, Supplementary Fig. 10**), hence, they were unlikely to be a result of systematic sequencing errors or background noise.

Previous studies of hypermutated tumors demonstrated significant spatial heterogeneity resulting in numerous subclonal mutations private to each tumor site45. Consistently, the overall proportion of tumor-matched mutations (biopsy-matched and biopsy-subthreshold) was significantly lower in the 10 hypermutated patients (17.2% [216 of 1210] non-WBC-matched mutations) compared to the 114 non-hypermutated patients (30.3% [523 of 1727]; p = 1.2e-16), whereas a higher proportion of tumor-derived variants were subclonal biopsy-subthreshold variants in hypermutated cases (41.2% [89 of 216] vs. 15.3% [80 of 523] respectively; p = 1.7e-13). These findings indicated that a single tumor biopsy may not capture the full landscape of tumor mutational profile in patients whose tumors harbor a hypermutator phenotype.

We next sought to define the biological source of nonsynonymous VUSo in cfDNA (**Methods**). After removing variants with known source-of-origin (WBC-matched in all samples; biopsy-matched and biopsy-subthreshold variants in cancer patient samples), approximately 31.9% of non-cancer controls had no additional variants identified in cfDNA, with the remaining 68.1% harboring at least one VUSo (**Fig. 4b**).

In cancer patients, 77.7% (994 of 1280) of the VUSo were detected in the 10 hypermutated cancer samples. In fact, VUSo accounted for 82.1% (994 of 1210) of the total non-WBC-matched somatic cfDNA mutations in hypermutated samples as compared with 35.4% (286 of 809) in non-hypermutated tumors (p = 9.3e-103). Additionally, VUSo rarely constituted the mutation at the highest VAF in cancer patients (17.6%, 23.1% and 19.5% of MBCs, NSCLCs and CRPCs, respectively). These findings indicate that a large proportion of the VUSo likely originated from the tumor and may not have been detected in the biopsy sample taken, due to spatial tumor heterogeneity and sampling bias. To investigate the potential origins of the VUSo further, we evaluated the genes harboring variants classified as such in cancer patients (**Fig. 2b, Supplementary Figs. 5 and 11**). A subset of VUSo affected specific genes known to harbor somatic mutations occurring late in the evolution of the respective cancer type and commonly found altered at subclonal levels in metastatic cancers, including mutations in *ESR1, RB1,* and *NF1* in MBC, the *EGFR* T790M mutation in NSCLC, and *AR* mutations in CRPC (**Fig. 2b**, **Supplementary Figs. 5 and 11**)34. The VAF distribution for these mutations mostly mirrored that of biopsy-matched variants (**Fig. 4e, Supplementary Fig. 10**), consistent with the hypothesis that a large proportion of VUSo are tumor-derived. In hypermutated cases, however, a significant correlation between the size of the sequenced coding region of a gene harboring VUSo and the number of VUSo affecting this given gene was observed (p = 4.4e-16; **Supplementary Fig. 12**). We posit that these mutations are for the most part tumor-derived and stem from increased mutational rates found in cancer cells from patients with tumors displaying a hypermutator phenotype. It should be noted that, in controls, the genes most frequently harboring VUSo included canonical CH genes (**Supplementary Fig. 11**). Consistent with the notion that at least a subset of VUSo arose from CH or other sources of somatic mosaicism not present in matched WBC samples, VUSo were weakly but significantly associated with age at sample collection (p = 0.0026; **Fig. 4c**), affected canonical CH genes in both cancer patients and controls (**Supplementary Fig. 11**), with some having similar allele frequencies as WBC-matched variants (**Fig. 4e, Supplementary Fig. 10**).

Taken together, this high-intensity cfDNA sequencing assay, combining joint mutation detection based on cfDNA and high-depth WBC sequencing, identified CH mutations as the most probable origin of non-tumor derived mutations detected in cfDNA, and provided evidence that subclonal tumor-derived mutations absent in the tumor biopsy can be detected in cfDNA. Our results also indicate that the CH frequency in cancer patients and healthy individuals appears to be substantially more prevalent than previously reported with lower-depth WBC sequencing approaches11,19,21-23,25.

***Characterization of WBC variants***

High-depth sequencing analysis of WBCs currently constitutes the main approach for the detection of somatic alterations originating from CH. Here, the cfDNA assay detected 57.3% of the somatic variants with supporting reads in WBCs which were also sequenced utilizing the same high-intensity assay (**Fig. 5a**; **Methods**). Importantly, CH was detected in 99.1% of the WBCs of the cancer patients analyzed, and in 93.6% of the non-cancer controls. In 41.6% of patients with metastatic cancer, the mutation found at the highest VAF affected one of the 15 canonical CH-related genes, with *DNMT3A* and *TET2* being the genes whose mutations were most frequently detected at the highest VAFs in both non-cancer controls and metastatic cancer patients (**Fig. 5b**). In fact, if a patient harbored a mutation affecting a canonical CH gene, there was a high likelihood of other CH mutations being detected in the same patient, and in those with CH events, the number of mutations was also significantly correlated with age (p = 6.09e-64, **Fig. 5c**).

Consistent with previous studies suggesting that therapeutic interventions may result in the acquisition of specific types of CH events19,46, our results indicated that somatic mutations affecting *PPM1D* were significantly more frequently detected in cancer patients than in controls (age-adjusted p = 0.0115, **Fig. 5c**). In addition, within the cohort of cancer patients, mutations affecting *PPM1D*, in particular in the form of truncating variants preferentially affecting the C-terminal domain (**Fig. 5d**, **Supplementary Fig. 13**),were significantly more common in those who received chemotherapy and/or radiation therapy than in patients who had no prior history of such treatments (age- and smoking-adjusted p = 0.0008, **Fig. 5c**).

**Gene copy number variation (CNV) detection**

As an exploratory, hypothesis generating analysis, we sought to define whether the high-intensity cfDNA assay would *de novo* detect CNVs. We observed a relatively good concordance between CNVs detected in tumor biopsies and cfDNA only in cases where the ctDNA fractions were ≥10% (**Supplementary Figs. 14 and 15**). Despite this limitation on ctDNA fraction, we sought to determine whether actionable CNVs reported in the tumor biopsies could still be detected by the high-intensity cfDNA assay. This analysis revealed that in five patients (n = 4 *ERBB2* (*HER2*)amplified MBCs and n = 1 *MET* amplified NSCL), three of the *ERBB2* amplifications could be detected *de novo*. In the two cases were these actionable alterations were present in the MSK-IMPACT tumor biopsy but not in cfDNA, the ctDNA fractions were 1.3% and 1.9% **(Supplementary Fig. 16)**. None of the remaining samples tested harbored amplifications of these two genes, demonstrating the specificity but that sensitivity of this assay for the detection of gene amplifications in cfDNA is highly dependent on the ctDNA fraction.

# DISCUSSION

Most cfDNA assays currently used are based solely on the analysis of a small panel of genes or hotspot mutations in key cancer genes, and do not incorporate matched WBC sequencing. Previous attempts at broadening the genomic area probed by cfDNA sequencing assays resulted in the identification of not only mutations known to be present in tumors but also a large number of variants absent from the respective tumor tissues and inferred to be somatic. Despite the use of multiple strategies to mitigate sequencing artifacts, it has been postulated that high-depth sequencing assays covering a large genomic region would inevitably result in the identification of a high number of false positive sequencing variants43,44. Here, we devised a high-intensity cfDNA sequencing assay covering a large genomic region based on a joint analysis of cfDNA and WBC gDNA, utilizing UMIs to suppress technical assay errors and hierarchical Bayesian error correction models to mitigate mutation detection artifacts stemming from ultra-high sequencing depths. Our findings highlight the importance of having methods to mitigate sequencing errors (e.g. UMIs and robust error correction methods) coupled with matched WBC sequencing performed at similar depths to those employed for the cfDNA analysis. Our high-intensity method demonstrated that cfDNA sequencing without considering the results of WBC sequencing, as currently performed, might be misleading, as some CH mutations affecting cancer genes may be interpreted as tumor-derived mutations (e.g. *TP53* mutations, which can be present in tumor-derived cfDNA and also be part of CH).

This cfDNA sequencing approach allowed for robust *de novo* detection of somatic mutations with a sensitivity similar to that of ddPCR (**Fig. 1f**). In fact, this assay was comparable to previous high-depth targeted sequencing efforts47-49 for the detection of ctDNA in plasma samples from patients with MBC, NSCLC and CRPC, and identified 77.4% of the repertoire of somatic mutations reported in the matched tumor biopsy samples from patients with advanced cancers. Given the large genomic footprint and the limited number of false positive variants (**Figs. 1d-e**), an exploratory analysis revealed the potential use of the high-intensity cfDNA sequencing assay for *de novo* characterization of tumor mutational burden and mutational signatures, including MSI (**Fig. 3a-e**), broadening the potential applications of cfDNA sequencing analysis in the context of patients with advanced cancers.

Our analyses revealed that the majority of non-tumor-matched nonsynonymous somatic mutations identified in cfDNA had supporting reads present in the respective WBC DNA samples. These WBC-matched mutations were present in the vast majority of non-cancer controls and cancer patients. These mutations preferentially affected genes previously implicated in CH11,19,21-23,25, and their presence was strongly associated with age at collection of the blood sample, consistent with the hypothesis that these alterations likely constitute CH. Importantly, the number of these probable CH variants per patient was on average higher than the number of tumor-matched variants in metastatic patients. The higher prevalence of CH found in WBCs in this study (93.6% of non-cancer controls and 99.1% in patients with advanced cancers) relative to that reported in prior studies11,19,21-23,25 likely resulted from the high sensitivity of the assay employed to detect variants in the WBC samples50,51 and was consistent with a recent observation utilizing a similar approach as reported in Liu *et al.*26. In the present study, however, both cfDNA and WBC samples were ultra-deep sequenced at comparable raw depths, allowing for the detection of CH at a higher sensitivity in WBCs and, consequently, the distinction between CH and tumor-derived mutations, which was not investigated by Liu *et al.*26. Although the genes recurrently affected by these somatic genetic alterations were genes previously implicated in CH, the majority of these WBC-matched variants were private to individual patients, suggesting that accounting for them in cfDNA-based clinical assays requires the sequencing of cfDNA and matched WBC DNA in a patient-specific manner. Indeed, recent studies24,27,52 have demonstrated in a limited number of patients that a large proportion of somatic variants in WBCs were also identified in cfDNA, resulting in the detection of ‘false-positive’ tumor-derived mutations in cfDNA. These findings provide a plausible explanation for the relatively low positive predictive value of prior cfDNA-based assays for the detection of tumor-derived mutations. Further, recent reports of inconsistent results between cfDNA and tumor tissue assays may be due to a subset of non-tumor origin (e.g. CH) cfDNA variants being interpreted as tumor-derived24,27,52. These findings emphasize the importance of joint analysis of cfDNA and matched WBC, given that mutations related to CH may result in inaccurate tumor mutation burden and mutational signature quantification8,27.

We also demonstrated that VUSo (cfDNA mutations not detected in WBC or tumor tissue) could have multiple origins, including tumor heterogeneity, CH occurring at extremely low levels, other sources of somatic mosaicism, or a small amount of residual technical noise. Our data suggest that a majority of the observed VUSo were tumor-derived and arose from minor tumor subclones. In support of this hypothesis, 77.7% of all VUSo in cancer patients were identified in 10 patients whose tumors harbored hypermutator mutational processes, such as APOBEC, known to amplify tumor heterogeneity and subclonal diversity. Additionally, many of these variants were also observed in genes previously identified to be altered in tumor subclones, such as variants associated with mechanisms of resistance to therapies, including *AR* and *ESR1* variants in CRPC and hormone receptor-positive MBC patients, respectively. Our results suggest that high-intensity cfDNA assays may offer a more comprehensive landscape of tumor mutational profile than tumor tissue sequencing alone.

This study has several limitations. Colorectal carcinomas, another common form of cancer, were not included in this study; hence, further studies are warranted to define the sources of cfDNA mutations in patients with this disease. The tumor assessment was limited to the analysis of a single tumor biopsy due to limitations in obtaining multiregional biopsies in the clinical setting. As such, the full scope of tumor heterogeneity may not have been entirely captured10. This caveat, however, would remain regardless of the number of sites biopsied. Healthy controls were from a different source and were processed in different batches from the tumor samples, potentially affecting results. Although the median collapsed target coverage (unique molecule counts) of cancer patient samples in this study was ~4,400X, increased mean collapsed target coverage could have revealed additional lower-prevalence variants in cfDNA. Given that the number of samples in each tumor subgroup was relatively small, the analysis performed here may not have captured the full spectrum of MBCs, NSCLCs, and CRPCs, and their respective subtypes. Additionally, the samples used to train the hierarchical Bayesian model was composed of <50 baseline samples from healthy controls; it is likely that the site-specific error rate estimates will be more accurate with repeat samples from hitherto healthy volunteers who would have been followed longitudinally. Our findings also emphasize the importance of high-depth WBC sequencing, and even when this approach is employed, a subset of VUSo might still originate from CH not detected in the matched WBC sample, other sources of somatic mosaicism, benign neoplasms and/or other forms of occult cancers not detected in the extensive clinical work up performed in the patients included in this study. Finally, the high-cost of the high-intensity cfDNA sequencing assay may preclude its broader adoption in the clinical context at present.

Despite these limitations, the cfDNA sequencing assay described here constitutes an advance in the development of approaches for *de novo* detection of the repertoire of somatic genetic alterations in cancer patients and provides further evidence that CH likely constitutes a biological phenomenon and a technical pitfall more prevalent than previously anticipated. Our findings also emphasize the importance of matched cfDNA-WBC sequencing at comparable depths to avoid the potential misclassification of CH affecting cancer genes as tumor-derived mutations.

# METHODS

## Study design

This was a prospective observational study of patients with metastatic breast (MBC), non-small cell lung (NSCLC), and castration resistant prostate (CRPC) cancer designed to characterize the detection of variants in plasma cfDNA using a targeted DNA assay (GRAIL, Inc.; Menlo Park, CA), and to evaluate the concordance of variant detection between tissue and plasma as evidence of ctDNA detection. The primary objectives were to assess the tumor cfDNA detection rate based on observing at least one variant (single-nucleotide variants [SNVs], indels); and to assess the concordance of the MSK-IMPACT variants32,33 detected in tumor biopsy samples versus cfDNA. Secondary objectives included assessing the ctDNA detection rate based on observing at least one MSK-IMPACT variant, characterizing the ctDNA detection rate as a function of the type of variant (SNV, indels) and the number of variants detected, and characterizing the proportion of patients with variants detected.

## Patient enrollment

All patients provided written informed consent for tumor, cfDNA, and WBC sequencing and review of patient medical records for detailed demographic, pathologic, and treatment information under an IRB-approved biospecimen umbrella protocol (MSKCC protocol 12-245, clinicaltrials.gov ID: NCT01775072). At least 50 patients of each type of cancer were enrolled to obtain evaluable patients with both the targeted DNA assay and MSK-IMPACT analysis. Clinical data (baseline demographics, cancer history, and prior lines of therapy) were collected from medical records.

Patients with MBC, NSCLC, or CRPC with disease progression as assessed by the investigator were eligible. Disease progression was based on objective radiographic and/or physical exam and/or biomarker results. Patients diagnosed with *de novo* or recurrent stage IV NSCLC or MBC were allowed to be included if enrolled prior to initiation of the first line of treatment for metastatic disease. No new therapies were permitted to be initiated between tissue biopsy and blood draw. Patients with progressive disease on stable doses of treatment (e.g. hormone therapy) were eligible. Blood was drawn within 6 weeks of tissue biopsy for MSK-IMPACT analysis either prior to or after tissue biopsy. Whole blood samples received outside of the stability timeframe for Streck DNA BCT (5 days) were excluded.

Fifty de-identified whole blood samples from self-reported healthy individuals (no diagnosis of cancer) were obtained from the San Diego Blood Bank (San Diego, CA). Limited clinical data were provided with the samples. Healthy participants were required to be at least 20 years of age, meet all eligibility for blood donation per standardized assessment and criteria, to lack a diagnosis of cancer, and to have no prior history of cancer. Participants were excluded if they had a prior history of cigarette smoking for at least one year, a current history of cigarette smoking, were pregnant, had a personal history of cancer, or had prior medical or surgical treatment of any type of cancer. Results were not returned to any patients, health care providers, or the San Diego Blood Bank.

## Tumor sample accessioning, processing, and analysis

For all the 161 patients, tumor DNA was extracted from FFPE biopsy samples and matched normal DNA was extracted from mononuclear cells from peripheral blood. All specimens underwent next-generation sequencing in the MSKCC CLIA-certified laboratory using MSK-IMPACT, a hybridization capture-based next-generation sequencing assay, which analyzes all protein-coding exons of 410 cancer-associated genes (**Supplementary Table 1**), as previously described32,33. Average sequencing coverage across all tumors was greater than 900X. Somatic mutations, DNA copy number alterations, and structural rearrangements were identified as previously described33. After excluding samples with insufficient tumor tissue, with insufficient data quality due to low total DNA quantity and purity, or that failed library preparation, a total of 124 had complete MSK-IMPACT results (**Supplementary Fig. 1**).

In addition to the gene-level amplification and deletion calls generated by the clinical laboratory pipeline, genome-wide total and allele-specific DNA copy numbers were determined using the FACETS algorithm53 for prospectively sequenced patients. Purity, average ploidy, and allele-specific integer-copy number for each segment were then determined by maximum likelihood. To determine the clonality of each mutation, we used allele-specific copy number inference from FACETS to calculate the fraction of mutated cancer cells (cancer cell fraction, CCF) as previously described54. Clonal mutations were those with a CCF (assuming the number of mutant copies was equal to the number of copies of the more frequent allele) greater than 0.8 or the upper bound of the CCF confidence interval was >0.85. Mutations with CCFs not meeting these conditions were defined as subclonal.

## Whole blood sample collection, accessioning, and preparation

Peripheral blood from patients with metastatic cancer was collected into two 10 mL Cell-Free DNA BCT (Streck; La Vista, NE) at Memorial Sloan Kettering Cancer Center (New York, NY) and shipped to GRAIL, Inc. (Menlo Park, CA) at room temperature. Whole blood from healthy individuals drawn into Streck BCTs were purchased from San Diego Blood Bank (San Diego, CA) and shipped to GRAIL, Inc. at room temperature. Received whole blood Streck BCTs were separated into plasma and buffy coat and stored at -80°C unless processed the same day.

cfDNA was extracted from two tubes of plasma (up to a combined volume of 8 ml) per subject using a modified QIAamp Circulating Nucleic Acid kit (Qiagen; Germantown, MD). Extracted cfDNA was quantified using the Fragment Analyzer High Sensitivity NGS kit (Advanced Analytical Technologies; Ankeny, IA). Genomic DNA (gDNA) from matching buffy coat (paired plasma and buffy coat from the same blood tube) was extracted using the Qiagen DNEasy Blood and Tissue kit. Extracted gDNA was quantified using NanoDrop (Thermo Scientific; Waltham, MA) and fragmented to a mean size of 180 base pairs using the Covaris E220 ultrasonicator (Woburn, MA). Sheared gDNA was subsequently size-selected using Agencourt AMPure XP magnetic beads (Beckman Coulter; Beverly, MA), then quantified using the Fragment Analyzer Standard Sensitivity NGS kit (Advanced Analytical Technologies; Ankeny, IA).

***Library preparation, target enrichment, and sequencing***

Buffy coat gDNA (50ng) and plasma cfDNA (≤75ng) were used for NGS library construction with a modified Illumina TruSeq DNA Nano protocol. Details are available in the **Supplementary Methods**.

## Analysis pipeline

A modular analysis pipeline was implemented to enable detection of mutations at very low allele fraction by suppressing noise caused by assay and alignment processes. The details of this pipeline are provided in the **Supplementary Methods**. In brief, this methodology consisted of: (1) preprocessing and a first-pass alignment, (2) collapsing and read-pair stitching, (3) candidate variant generation by *de novo* assembly, (4) edge effect scoring, (5) candidate variant analysis with recalibrated quality scores based on a hierarchical Bayesian model, and (6) joint variant analysis using the machine learning error model (**Supplementary Methods**), which was critical in accounting for clonal hematopoiesis of indeterminate potential and other artifacts.

## Source of origin of plasma variants

Variants reported by *de novo* assembly from control and cancer samples were stacked, and their source-of-origin were labeled through a hierarchical schema. First, variants with low read coverage (<200), high frequency of recurrence in WBCs, failed edge-variant filter, or below the noise model threshold were labeled as noise. Second, variants with allele fraction >20% matched in WBC were labeled as potentially germline. Third, synonymous variants were labeled as an independent category. Fourth, variants present in WBCs identified by joint-calling or leaking through joint-calling but failing additional thresholds were labeled as `WBC-matched’. The additional threshold filtered variants on smoothed cfDNA allele ratio and matching WBC alternative allele depth variation. Variants unable to be joint-called as separable from WBC were labeled ambiguous (no positive evidence for variant alleles in WBC, but insufficient depth of sequencing to prove allele frequency was statistically different in cfDNA and WBC results). The remaining variants were labeled as somatic. Somatic variants also present in the MSK-IMPACT sequencing of the tumor biopsy were labeled as biopsy-matched if they had been reported or biopsy-subthreshold if they were below the limit of clinical detection. Variants not matched were labeled as `variants of unknown source’ (VUSo).

## Tumor concordance

Overall agreement between variants in plasma and tumor tissue was measured using positive percent agreement (PPA) with tumor tissue as the reference; this can be expressed as the percent of tissue variants also detected in plasma. The top mutated cancer genes were generated by merging the top 15 genes reported by MSK-IMPACT analysis from each cancer cohort. Somatic variants (VUSo, biopsy-matched, and biopsy-subthreshold) from the top mutated cancer genes were selected from plasma variants for plotting and comparison.

***Disease burden and ctDNA fraction***

The ctDNA fraction for each plasma sample was estimated from clonal biopsy-matched mutations. Briefly, we first obtained the CCF estimate for somatic mutations detected in the matched tumor biopsy sample using the FACETS algorithm as previously described (PMID: 27270079), and then derived the ctDNA based on the VAF in cfDNA of the biopsy-matched clonal mutations.

Seventy-seven of the 80 patients in the NSCLC and MBC cohorts had computerized tomography (CT) scans available from which volumetric tumor measurements could be obtained. Of these, 34 of the exams were CTs of the chest, abdomen, and pelvis without IV contrast, obtained as part of a positron emission tomography (PET)/CT exam; 32 exams were CTs of the chest, abdomen, and pelvis with IV contrast; 5 exams were CTs of the chest only with IV contrast; 4 exams were CTs of the chest only without IV contrast; and 2 exams were CTs of the chest and abdomen with IV contrast. Exams were acquired on several different scanners at slice thicknesses ranging from 3.75 - 5 mm.

All exams were reviewed by a board-certified radiologist specializing in imaging of the chest, abdomen, and pelvis (KJ). All metastatic lesions >1 cm in diameter were identified. Volumes were measured on all lesions except bone lesions. Bone lesions often have poorly defined borders and overlap the findings in active metastasis vs treated disease. Volumes were measured using the Aquarius iNtuition advanced visualization software, version 4.4.13.P3 (TeraRecon, Inc, Foster City, CA). Of the 77 patients with available volumetric assessment, 34 MBC and 29 NSCLC patients had evaluable ctDNA fraction and included in this analysis.

Given that the majority of CRPC patients included in this study had extensive bone disease and had undergone bone scans prior to enrollment in the study, the approach employed for the volumetric assessment of disease burden was different from that used for MBCs and NSCLCs. We obtained the automated bone scan index (aBSI, platform version 3.3, EXINI Diagnostics AB, Lund, Sweden), a fully quantitative assessment of a patient’s bony disease on a bone scan that reports the number of lesions, area and the fraction of the total skeleton weight that is involved by tumor, as a proxy for bone disease burden. The methodology of the automated platform has been described in previous studies (PMID: 22306323). In brief, a neural network automatically segments the different anatomical regions of the skeleton followed by detection and classification of the abnormal hotspots. The weight fraction of the skeleton for each metastatic hotspot was calculated and the aBSI was calculated as the sum of all such fractions. The aBSI method utilized in this study has been shown to be an objective measure of the quantitative change in disease burden bone scans and a prognostic biomarker in patients with CRPC (PMID: 29799999).

## Mutation burden and association with age at diagnosis

Mutation burden was calculated as the number of nonsynonymous mutations per megabase pair of genome sequenced. The relationship of mutation burden with age and cancer status was examined by fitting a zero-inflated Poisson regression with the cancer status as covariate. To assess the age relationship with variant source, the analysis above was stratified by variant source of origin.

## Mutational signatures from hypermutated patients

The threshold of mutation burden used to define hypermutated patients was defined as 13.8 mutations/Mb 32 for the tumor biopsy whilst the corresponding value for cfDNA was evaluated *de novo* from the samples of cancer patients as median (cfDNA mutation burden) + 2 × IQR (cfDNA mutation burden), where IQR is the interquartile range. The contributions of different mutation signatures were identified for each sample according to distribution of the six substitution classes (C>A, C>G, C>T, T>A, T>C, T>G) and the bases immediately 5′ and 3′ of the mutated base, producing 96 possible mutation subtypes using deconstructSigs55. For analyses in the manuscript, we focused on six signatures: (1) aging (signature 1 and 5), (2) APOBEC (signatures 2 and 13), (3) homologous recombination repair deficiency (HRD, signature 3), (4) MMR (signatures 6, 15, 20 and 26), (5) smoking (signature 4), and (6) POLE (signature 10).

## Prevalence of clonal hematopoiesis in cfDNA

WBC-matched variant occurrence was measured at the gene level using the ratio between the number variants in a gene versus the total number of variants. The cumulative frequency was generated by first ranking the ratio by descending order and then recursively adding ratio together. Top mutated genes carrying WBC-matched variants were visualized by a heatmap. The top 20 genes were selected from each cohort and merged to form the final list of top genes. The number of patients carrying WBC-matched variants in each gene was used to measure the gene occurrence.

## Prevalence of clonal hematopoiesis in WBC

Candidate variants in WBC were generated by *de novo* assembly of error corrected and stitched read pairs and post-filtered as follows: (1) following quality score recalibration, variants with low quality (<60) or low depth (<500X) as well as *bona fide* somatic variants found in the corresponding tumor biopsy were excluded from downstream analyses; (2) variants recurring at >5% were filtered out unless (a) they had previously been reported as somatic in any of COSMIC (v86), Kandoth *et al.*56 or Chang *et al.*57, (b) they were frameshifting indels or truncating SNVs, or (c) they occurred in one of 15 canonical genes known to be associated with CH; (3) variants with VAF >30% were labelled germline and filtered out unless (a) they were frameshifting indels or truncating SNVs, or (b) they occurred in one of the 15 canonical CH genes; (4) variants occurring at any allele frequency in ExAC or gnomAD58 were labelled germline and filtered out; (5) variants mapping to the HLA-A locus were excluded; and (6) only nonsynonymous exonic variants passing the above filters were considered further. The 15 canonical genes known to be associated with were *DNMT3A,* *TET2, ASXL1, PPM1D, TP53, JAK2, RUNX1, SF3B1, SRSF2, IDH1, IDH2, U2AF1, CBL, ATM, CHEK2*.

## Statistical analyses

The difference in detection rate between the three cohorts was assessed using two-by-two Fisher exact tests, and the resulting p-values were adjusted using the Bonferroni correction for multiple testing. The association between the tumor cancer cell fraction and the cfDNA detection rate was assessed using stratified by cancer type. The exact confidence intervals were calculated for the detection rates by cancer type, ctDNA fraction or tumor burden. The difference in ctDNA fraction estimate and cancer types was assessed using exact Fisher test. The tests of trends were conducted using Jonckheere-Terpstra test or Kendall correlation as appropriate.

For WBC-matched variants, a two-sided Mann-Whitney U test was used to test whether the cancer cohort had a different mutation burden than the control cohort. For VUSo, a one-sided Mann-Whitney U test was used to test whether the cancer cohort had a greater mutation burden than the control cohort. We used zero-inflated Poisson regression to assess the association between the number of mutations in each category (biopsy-matched, biopsy-subthreshold, WBC-matched and VUSo) with age.

The association of clonal hematopoiesis measured in WBC in each of the 15 canonical CH genes and cancer status and prior history of radio- or chemotherapy was assessed using exact Fisher test or Chi-Square test as appropriate. All statistical hypothesis tests were two-sided with α = 0.05 and carried out in R/Bioconductor.

## Summary of variants and variant allele fractions in cfDNA

The mean and median number of each type of identified variant in the samples, as well as the mean and median VAF in the samples, are described in **Supplementary Table 7**. In cfDNA samples, more WBC-matched variants than biopsy-matched variants or VUSo were identified. Median VAF in cfDNA was higher for biopsy-matched variants than for WBC-matched variants or VUSo.

## Sensitivity and specificity of the targeted DNA assay

Prior to analysis of patient samples with the targeted DNA assay, analytical characterization was performed using titrations of DNA from cell lines. Genomic DNA extracted from EBV-immortalised lymphoblastoid cell line (NA12878) was purchased from Coriell Institute (Camden, NJ). The HD753 Structural Multiplex Reference Standard gDNA, which contains known SNVs, indels, fusions, and deletions, was purchased from Horizon Discovery (Cambridge, MA) (**Supplementary Table 8**). Fifteen DNA titrations using the HD753 standard and the NA12878 gDNA were prepared in triplicate to have nominal expected VAFs of 0, 0.1, 0.25, 0.5, and 1% for a majority of variants. The gDNA titrations were verified using a droplet digital PCR (ddPCR; Bio-Rad; Hercules, CA) to ensure dilution accuracy (**Supplementary Table 9**). Following ddPCR verification, DNA mixtures were sheared and size-selected according to the targeted DNA assay protocol. 30 ng of sheared, size-selected gDNA was used for library construction, resulting in a mean collapsed target coverage of 2,430X.

**Fig. 1b** shows the estimated sensitivity of the targeted DNA assay at various VAFs, using a probit regression model of variant calling status of 14 known small variants in all HD753 gDNA titrations. Each half fastq of one replicate was also combined into the other two fastqs of replicates in the same titration to create three additional FASTQ (i.e. if the triplicates are labelled A, B and C, the three simulated samples are AB = 0.5A+0.5B, AC = 0.5A+0.5C and BC = 0.5B+0.5C), simulating higher input sample cases. Therefore, the mean target collapsed depth of these three simulated samples was theoretically twice rather than three times that of a single replicate at any given titration. In fact, the mean collapsed target coverage of simulated samples (n=10) at the fastq level was 4,577X, which is similar to the median of mean collapsed target coverages for all cancer patient samples reported here (4,408X). The estimated 95% limit of detection was 0.36% for 30 ng of input DNA (mean collapsed target coverage of 2,430X), and 0.16% for simulated cases (mean collapsed target coverage of 4,577X).

**Fig. 1c** summarizes the specificity of the targeted DNA assay using non-cancer control samples (n=47). After *de novo* variant calling and WBC variant filtering, the mean number of called variants was 120.8, corresponding to a specificity of 99.9891%. After the machine learning-based joint variant calling and filtering, the mean number of called variants was reduced to 2.3, corresponding to a specificity of 99.9998%. While this drastically improved the specificity, the decrease of variant calling sensitivity was marginal. Using the same variant calling settings, the estimated sensitivity using the HD753 titrations were comparable between *de novo* variant calling and joint variant calling.

## Reproducibility of the targeted DNA assay

The high-intensity sequencing assay was validated using two distinct approaches, namely (1) repeated sequencing of the same sample using two versions of the assay (V1 and V2), and (2) ddPCR analysis of biopsy-matched mutations and VUSo. For details, please see the **Supplementary Methods, Figs. 1d-f, Supplementary Figs. 4, RR8, RR14 and RR15, Supplementary Table 6**).

***Microsatellite instability detection in high depth-of-read cfDNA assays***

An adjusted version of MSIsensor39 described in the **Supplementary Methods**, was employed. Using the distributions obtained from MSIsensor and applying updated parameters and filters, more robust results were obtained in both tumor-normal utilizing MSK-IMPACT and the higher depth-of-read cfDNA-WBC samples (**Supplementary Fig. 7**). These results suggest that the high depth-of-read cfDNA data generated in this study are suitable for detecting MSI in cancer, and that MSI detection can be further improved for shallow sequencing biopsies.

**Data Availability**

The assembled prospective somatic mutational data from cfDNA, WBC, and tumors for the entire cohort are provided as supplementary tables (**Supplementary Tables 11-13**) and deposited in the European Genome-phenome Archive (EGA) under accession number EGAS00001003755.

**REFERENCES**

1. Mandel, P. & Metais, P. [Not Available]. *C R Seances Soc Biol Fil* **142**, 241-243 (1948).

2. Stroun, M., Anker, P., Lyautey, J., Lederrey, C. & Maurice, P.A. Isolation and characterization of DNA from the plasma of cancer patients. *Eur J Cancer Clin Oncol* **23**, 707-712 (1987).

3. Leon, S.A., Shapiro, B., Sklaroff, D.M. & Yaros, M.J. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res* **37**, 646-650 (1977).

4. Jr, L.A.D. & Bardelli, A. Liquid Biopsies: Genotyping Circulating Tumor DNA. *Journal of Clinical Oncology* **32**, 579-586 (2014).

5. Wan, J.C.M.*, et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer* **17**, 223-238 (2017).

6. Lanman, R.B.*, et al.* Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA. *PLoS One* **10**, e0140712 (2015).

7. Adalsteinsson, V.A.*, et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017).

8. Aravanis, A.M., Lee, M. & Klausner, R.D. Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell* **168**, 571-574 (2017).

9. Acuna-Hidalgo, R.*, et al.* Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am J Hum Genet* **101**, 50-64 (2017).

10. Jamal-Hanjani, M.*, et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2109-2121 (2017).

11. Jaiswal, S.*, et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498 (2014).

12. Choi, M.*, et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096-19101 (2009).

13. Murtaza, M.*, et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108-112 (2013).

14. Przybyl, J.*, et al.* Combination Approach for Detecting Different Types of Alterations in Circulating Tumor DNA in Leiomyosarcoma. *Clin Cancer Res* **24**, 2688-2699 (2018).

15. Risques, R.A. & Kennedy, S.R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet* **14**, e1007108 (2018).

16. Steensma, D.P.*, et al.* Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood* **126**, 9-16 (2015).

17. Bowman, R.L., Busque, L. & Levine, R.L. Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. *Cell Stem Cell* **22**, 157-170 (2018).

18. Busque, L., Buscarlet, M., Mollica, L. & Levine, R.L. Concise Review: Age-Related Clonal Hematopoiesis: Stem Cells Tempting the Devil. *Stem Cells* **36**, 1287-1294 (2018).

19. Coombs, C.C.*, et al.* Therapy-Related Clonal Hematopoiesis in Patients with Non-hematologic Cancers Is Common and Associated with Adverse Clinical Outcomes. *Cell Stem Cell* **21**, 374-382 e374 (2017).

20. Zink, F.*, et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742-752 (2017).

21. Jaiswal, S.*, et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N Engl J Med* **377**, 111-121 (2017).

22. Xie, M.*, et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20**, 1472-1478 (2014).

23. Genovese, G.*, et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487 (2014).

24. Phallen, J.*, et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* **9**(2017).

25. Gillis, N.K.*, et al.* Clonal haemopoiesis and therapy-related myeloid malignancies in elderly patients: a proof-of-concept, case-control study. *Lancet Oncol* **18**, 112-121 (2017).

26. Liu, J.*, et al.* Biological background of the genomic variations of cf-DNA in healthy individuals. *Ann Oncol* (2018).

27. Hu, Y.*, et al.* False-Positive Plasma Genotyping Due to Clonal Hematopoiesis. *Clin Cancer Res* **24**, 4437-4443 (2018).

28. Janku, F.*, et al.* Development and Validation of an Ultradeep Next-Generation Sequencing Assay for Testing of Plasma Cell-Free DNA from Patients with Advanced Cancer. *Clin Cancer Res* **23**, 5648-5656 (2017).

29. Thompson, J.C.*, et al.* Detection of Therapeutically Targetable Driver and Resistance Mutations in Lung Cancer Patients by Next-Generation Sequencing of Cell-Free Circulating Tumor DNA. *Clin Cancer Res* **22**, 5772-5782 (2016).

30. Guibert, N.*, et al.* Amplicon-based next-generation sequencing of plasma cell-free DNA for detection of driver and resistance mutations in advanced non-small cell lung cancer. *Ann Oncol* **29**, 1049-1055 (2018).

31. Sacher, A.G.*, et al.* Prospective Validation of Rapid Plasma Genotyping for the Detection of EGFR and KRAS Mutations in Advanced Lung Cancer. *JAMA Oncol* **2**, 1014-1022 (2016).

32. Zehir, A.*, et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* **23**, 703-713 (2017).

33. Cheng, D.T.*, et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* **17**, 251-264 (2015).

34. Razavi, P.*, et al.* The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. *Cancer Cell* **34**, 427-438 e426 (2018).

35. Khagi, Y.*, et al.* Hypermutated Circulating Tumor DNA: Correlation with Response to Checkpoint Inhibitor-Based Immunotherapy. *Clin Cancer Res* **23**, 5729-5736 (2017).

36. Clark, T.A.*, et al.* Analytical Validation of a Hybrid Capture-Based Next-Generation Sequencing Clinical Assay for Genomic Profiling of Cell-Free Circulating Tumor DNA. *J Mol Diagn* **20**, 686-702 (2018).

37. Alexandrov, L.B.*, et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).

38. Nik-Zainal, S.*, et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).

39. Niu, B.*, et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015-1016 (2014).

40. Polak, P.*, et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat Genet* **49**, 1476-1486 (2017).

41. de Bruin, E.C.*, et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251-256 (2014).

42. Le, D.T.*, et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med* **372**, 2509-2520 (2015).

43. Merker, J.D.*, et al.* Circulating Tumor DNA Analysis in Patients With Cancer: American Society of Clinical Oncology and College of American Pathologists Joint Review. *Arch Pathol Lab Med* **142**, 1242-1253 (2018).

44. Cohen, J.D.*, et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926-930 (2018).

45. Schultheis, A.M.*, et al.* Massively Parallel Sequencing-Based Clonality Analysis of Synchronous Endometrioid Endometrial and Ovarian Carcinomas. *J Natl Cancer Inst* **108**, djv427 (2016).

46. Hsu, J.I.*, et al.* PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. *Cell Stem Cell* **23**, 700-713 e706 (2018).

47. Bettegowda, C.*, et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **6**, 224ra224 (2014).

48. Dawson, S.J.*, et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med* **368**, 1199-1209 (2013).

49. Chabon, J.J.*, et al.* Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. *Nat Commun* **7**, 11815 (2016).

50. Young, A.L., Challen, G.A., Birmann, B.M. & Druley, T.E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat Commun* **7**, 12484 (2016).

51. Swanton, C.*, et al.* Prevalence of clonal hematopoiesis of indeterminate potential (CHIP) measured by an ultra-sensitive sequencing assay: Exploratory analysis of the Circulating Cancer Genome Atlas (CCGA) study. *J Clin Oncol* **36**(2018).

52. Mansukhani, S.*, et al.* Ultra-Sensitive Mutation Detection and Genome-Wide DNA Copy Number Reconstruction by Error-Corrected Circulating Tumor DNA Sequencing. *Clin Chem* (2018).

53. Shen, R. & Seshan, V.E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* **44**, e131 (2016).

54. Carter, S.L.*, et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* **30**, 413-421 (2012).

55. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).

56. Kandoth, C.*, et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013).

57. Chang, M.T.*, et al.* Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov* **8**, 174-183 (2018).

58. Lek, M.*, et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).

# Acknowledgements

# We thank the following GRAIL, Inc. and Memorial Sloan Kettering Cancer Center associates for their helpful discussions and contributions to this body of work: M. Berger, N. Schultz, C. Bain, M. Chung, M. Eriksen, T. Liu, R. Mauntz, A. Mich, J. Nguyen, Y. Park, S. Ramani, E. Scott, K. Shashidhar, C. Tom, and S. Wen. This work was supported by GRAIL, Inc. and National Institutes of Health awards P30 CA008748, R01 CA190642, Breast Cancer Alliance Young Investigator Award (PR), and the Breast Cancer Research Foundation (JSR-F), Congressionally Directed Medical Research Programs W81XWH‑15‑1‑0547 and GC229671 (JSR-F).

# Author Contributions

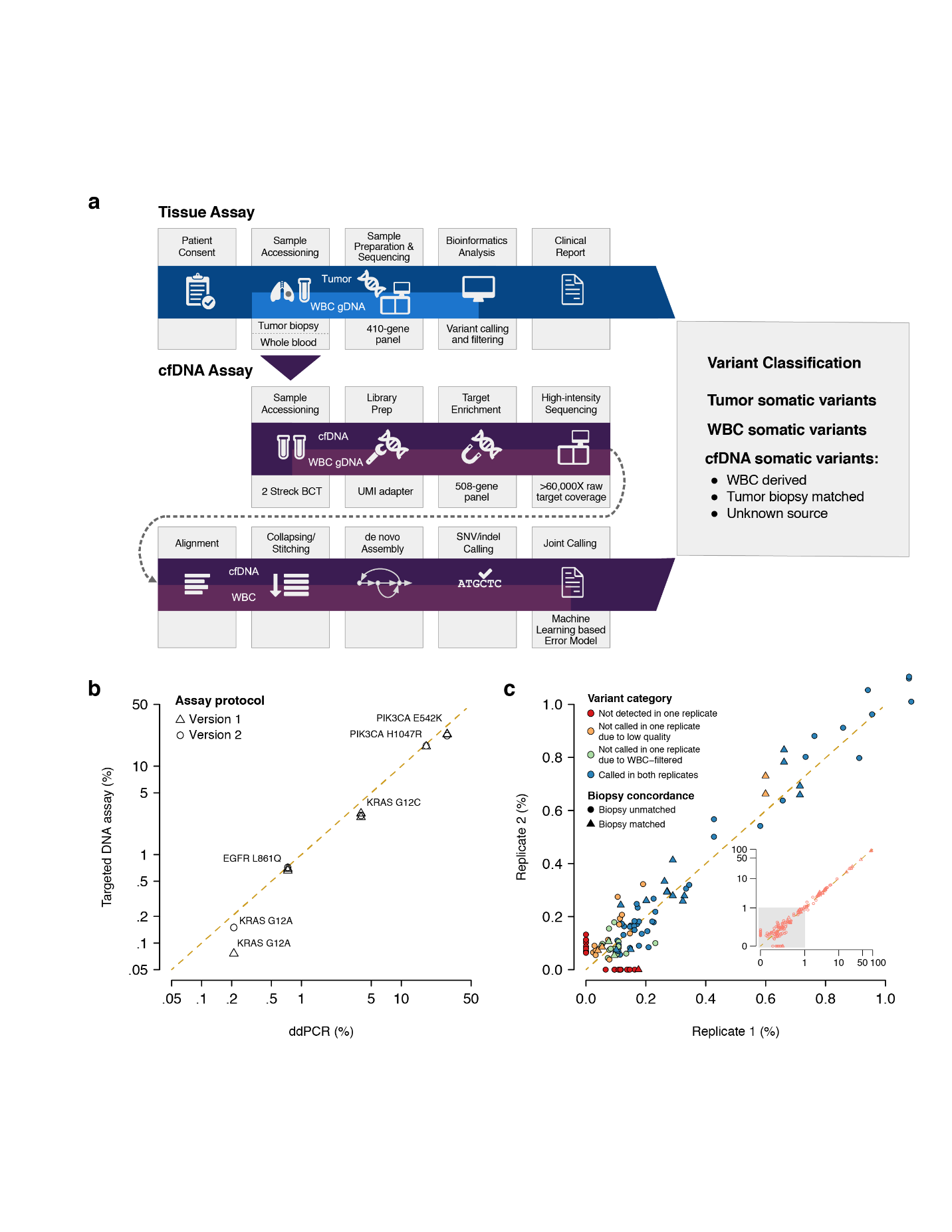
Conceived the study: PR, BTL, DBS, AMA, JSR-F; Data acquisition: PR, BTL, BJ, WA, CH, RVS, KL, LS, NE, JY, HZ, MPH, AS-Z, WEN, JMI, BWR, GP, ASH, ML, DMH, DRJ, MM, GJR, HIS, CMR, MER, LAD, DBS, AA; Data analysis and interpretation: PR, DNB, EH, RS, IDB, OV, RL, TM, AWB, AMA, JSR-F. Bioinformatics and genomic analysis: PR, DNB, EH, RS, IDB, OB, SG, AWB, AA, JSR-F. Manuscript first draft: PR, DNB, EH, MPH, AMA, JSR-F wrote the manuscript with input from all authors. Manuscript review and approval: all authors.

**Competing Interests**

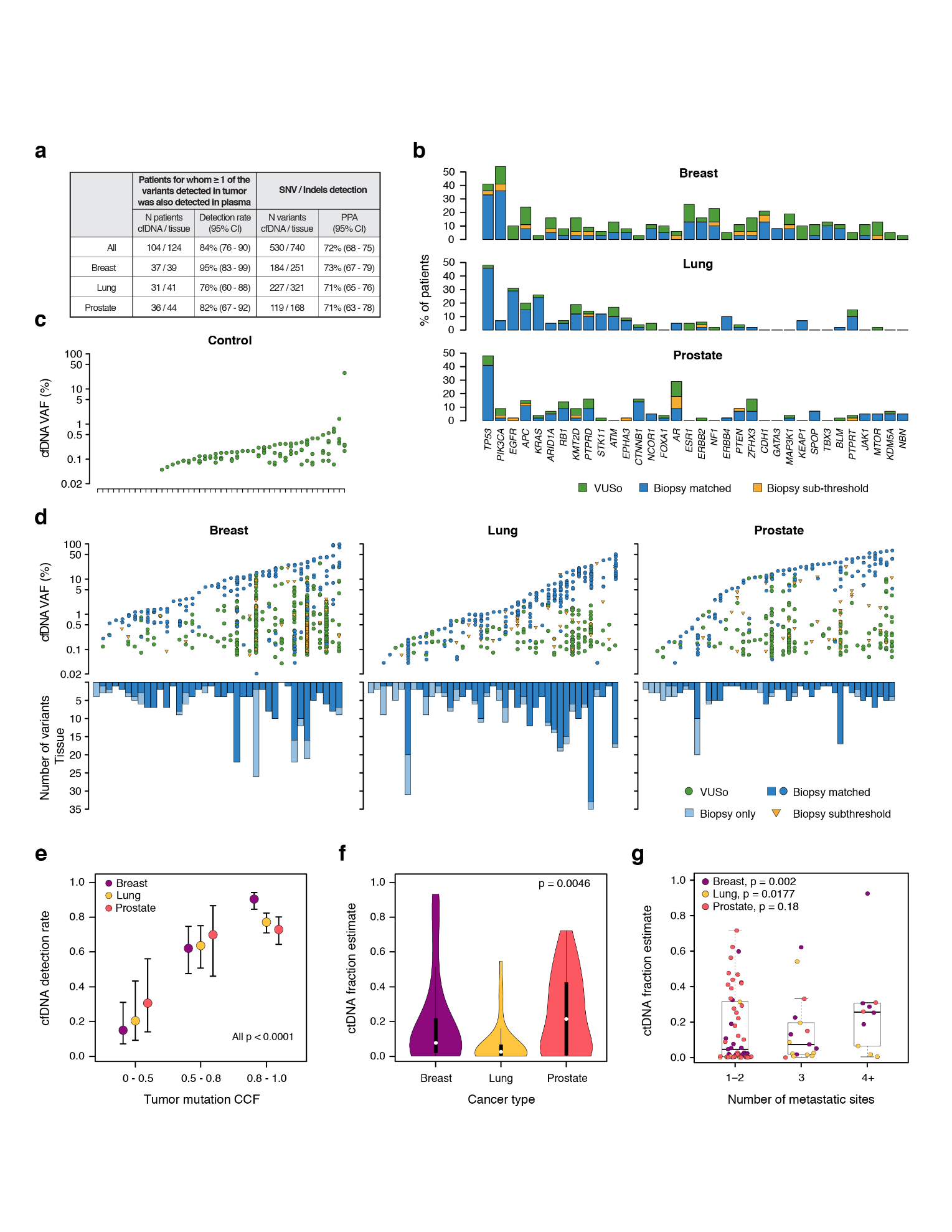
PR reports consulting/advisory board for Novartis and institutional research support from Illumina and GRAIL, Inc. BTL reports consulting/advisory board for Genentech, ThermoFisher Scientific, Guardant Health, Hengrui Therapeutics, Mersana Therapeutics, Biosceptre Australia and institutional research support from Illumina, GRAIL, Inc., Genentech, AstraZeneca. WA reports consulting or advisory role from Clovis Oncology, Janssen and MORE Health, and received Honoraria from CARET and received institutional research support from AstraZeneca, Zenith Epigenetics, Clovis Oncology and GlaxoSmithKline and also received travel/accommodations/expenses from GlaxoSmithKline and Clovis Oncology. GP is on the Scientific Advisory Board Member for Tizona Therapeutics and has consulted for Merck, BMS, Kyowa Hakko Kirin Pharma. DMH has received personal fees from Atara Biotherapeutics, personal fees fromChugai Pharma, personal fees from Boehringer Ingelheim, personal fees from AstraZeneca, personal fees from Pfizer, personal fees from Bayer, personal fees from

Debiopharm Group, personal fees from Genetech, grants from AstraZeneca, grants from Puma Biotechnology, grants from Loxo Oncology, outside the submitted work. GJR received consulting for Genentech/Roche in 2016 and received institutional research support for clinical research from Pfizer, Roche/Genentech, and Takeda. CMR has consulted on oncology drug development with Abbvie, Amgen, Ascentage, AstraZeneca, Bicycle, Celgene, Chugai, Daiichi Sankyo, Genentech/Roche, GI Therapeutics, Loxo, Novartis, Pharmamar, and Seattle Genetics; he is on the Scientific Advisory Boards of Harpoon Therapeutics and Elucida. LAD is a member of the board of directors of Personal Genome Diagnostics (PGDx) and Jounce Therapeutics.  LAD holds equity in PapGene, Personal Genome Diagnostics (PGDx) and Phoremost.  He is a paid consultant for Merck, PGDx and Phoremost.  LAD is an inventor of licensed intellectual property related to technology for circulating tumor DNA analyses and mismatch repair deficiency for diagnosis and therapy (WO2016077553A1) from Johns Hopkins University. These licenses and relationships are associated with equity or royalty payments to LAD.  The terms of all these arrangements are being managed by Johns Hopkins and Memorial Sloan Kettering in accordance with their conflict of interest policies. In addition, in the past 5 years, LAD has participated as a paid consultant for one-time engagements with Caris, Lyndra, Genocea Biosciences, Illumina and Cell Design Labs. DBS. received honoraria/consulted for Pfizer, Loxo Oncology, Illumina, Intezyne and Vivideon Therapuetics. JSR-F reports personal/consultancy fees from VolitionRx, Page.AI, Goldman Sachs, Grail, Ventana Medical Systems, Roche, Genentech and Invicro, outside the scope of the submitted work. BJ, EH, CH, OV, TM, SG, RVS, QL, LS, NE, JY, AWB, ML AS, HX, MPH, WFN, AMA are or were GRAIL, Inc. employees and hold stock and/or other ownership interests in GRAIL, Inc. The other coauthors report no competing interests.

# Figures

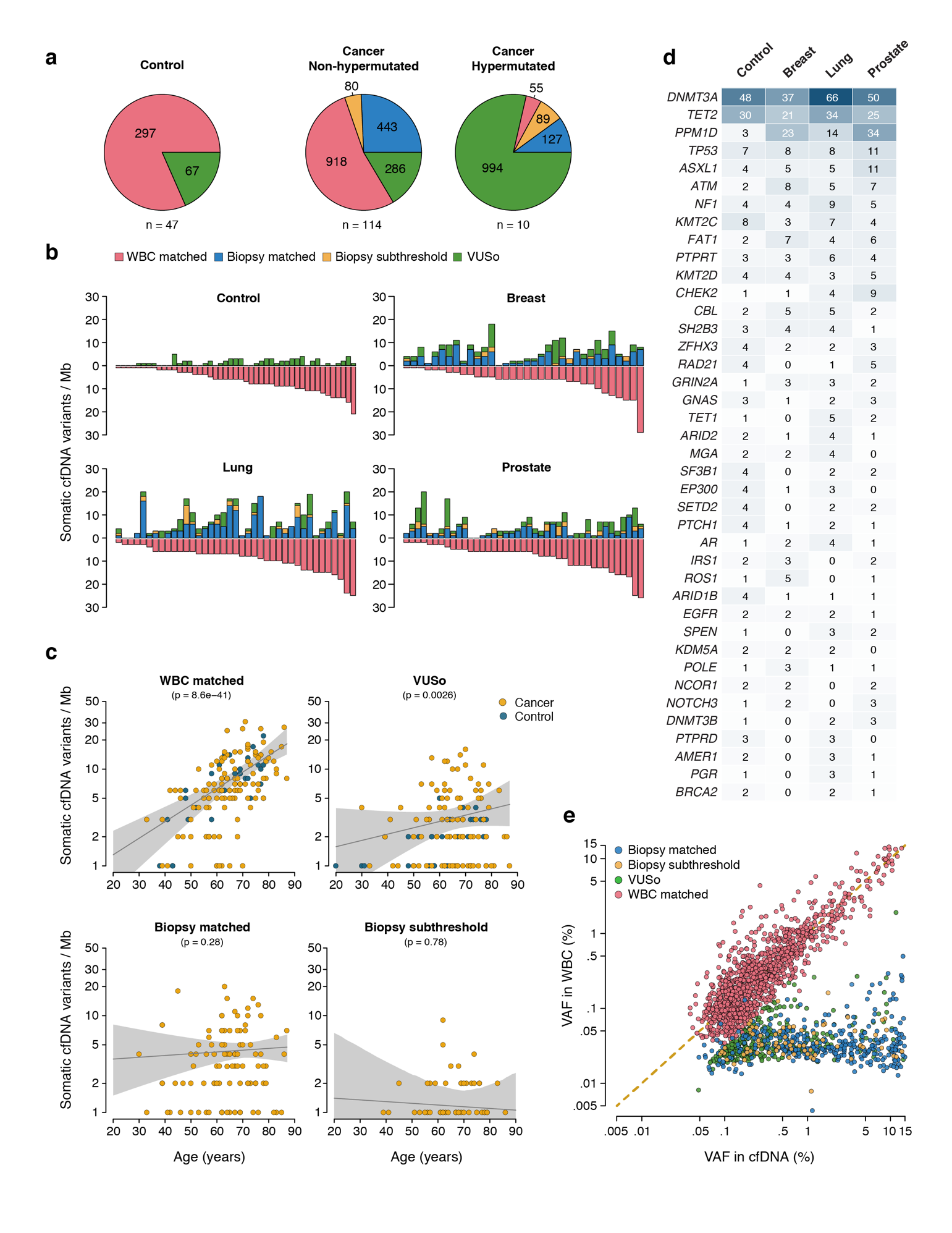


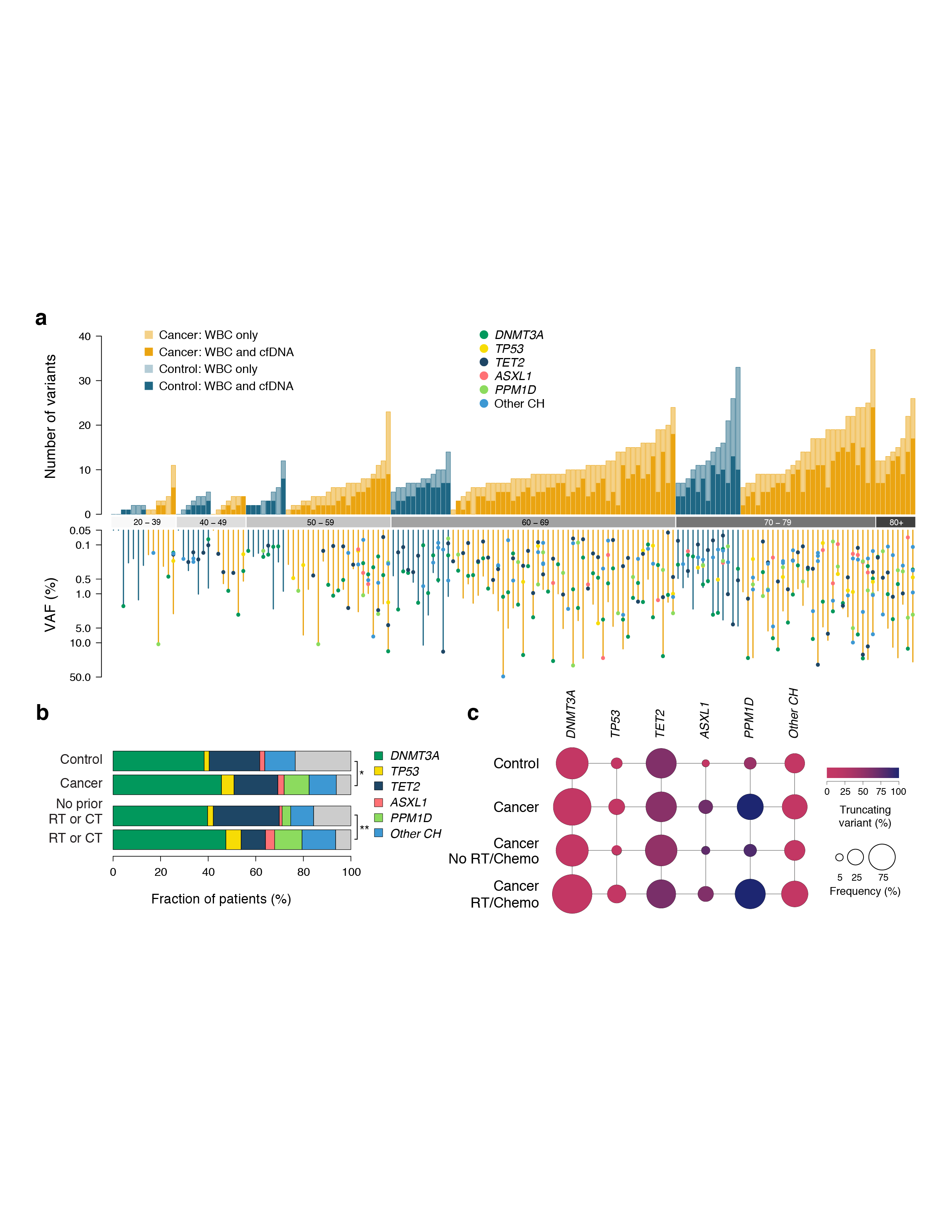
**Fig. 1. Assay workflow and reproducibility.** (**a**) Tumor and cfDNA samples were collected from patients with metastatic breast (MBC), lung (NSCLC), and prostate (CRPC) cancers. Tumor and matched normal samples were sequenced using the MSK-IMPACT assay, while plasma and buffy coat samples from cancer patients and non-cancer controls from the San Diego Blood Bank underwent sequencing followed by *de novo* assembly and mutation detection using the high-intensity targeted cfDNA assay by GRAIL, Inc (Menlo Park, CA) based on a bespoke joint-variant-calling pipeline. Tumor and cfDNA somatic variant detection results were unblinded for concordance analyses. (**b**) Comparison of variant allele fraction (VAF) measured using the targeted DNA assay (y-axis) and ddPCR (x-axis). cfDNA extracted from five cancer patients with canonical hotspot mutations were subjected to ddPCR. An aliquot of the same cfDNA isolate was used for targeted DNA assay using two versions of the protocol (V1 and V2). One sample lacking canonical hotspot mutation in the ddPCR measurements was excluded. (**c**) Comparison of allele fraction of variants detected using either of the two targeted DNA assay protocols in five patients. One MBC hypermutated patient was excluded from this analysis to avoid biased regression. Concordant mutation detection between the two replicates (triangles indicate biopsy-matched, circles indicate biopsy-unmatched variants) is enriched in allele fraction above limit of detection. The colors of the circles and triangles indicate whether the variants were detected in both replicates (blue), called in only one replicate (red), not called in one replicate due to low sample quality (yellow), or not called in one replicate due to filtering against WBC (green).

****

**Fig. 2. Concordance of cfDNA variants with tumor biopsy.** (**a**) Summary statistics of concordance between cfDNA and tumor biopsy assays for 124 patients with MBC (N = 39), NSCLC (N = 41), and CRPC (N = 44) cancer. (**b**) Frequency of genomic alterations in cfDNA of the same patients with MBC (top), NSCLC (middle), and CRPC (bottom) cancers. The genes were sorted by their frequency of alterations in the tumor. The colors indicate whether the alterations were biopsy-matched, detected in the tumor but below the threshold of the MSK-IMPACT assay (biopsy-subthreshold), or were specific to cfDNA i.e. variants of unknown source (VUSo). (**c**) Plasma variant allele fractions (VAF) of somatic variants sorted by the maximum VAF in control individuals. (**d**) Upper panels depict plasma VAFs of somatic variants in MBC, NSCLC, and CRPC. The lower panels show the number of variants identified in each individual by MSK-IMPACT. Colors indicate whether alterations were biopsy-matched, biopsy-subthreshold, detected in tumor only and not in cfDNA, or were VUSo. (**e**) Increasing detection rate of tumor variants in cfDNA with clonality of mutations in the tumor biopsy. The box plot shows the proportion of tumor mutations from the MSK-IMPACT assay that were also detected in cfDNA of MBC, NSCLC, and CRPC patients, stratified by the cancer cell fraction (CCF) in the tumor. The error bars indicate the 95% binomial confidence intervals. The CCF was strongly associated with detection rate in cfDNA (all p values by cancer type < 0.0001). (**f**) Distribution of tumor derived cfDNA fraction estimates in MBC, NSCLC, and CRPC patients (p = 0.0046). (**g**) Distribution of tumor derived cfDNA fraction estimates as a function of number of metastatic sites.



**Fig. 3. Tumor mutational burden and mutational signatures derived from cfDNA targeted assay.** (**a**) Distribution of the somatic tumor mutation burden (TMB), defined as the number of nonsynonymous mutations per megabase (Mb), in tumor (x-axis) and cfDNA (y-axis). The vertical dashed line indicates the threshold for samples with a high TMB based on tumor biopsy (13.8 mutations/Mb) and the horizontal dashed line indicates the threshold for samples with a high TMB in cfDNA (22.7 mutations/Mb). (**b**) Venn diagrams showing the total number of mutations detected in cfDNA (red) and tumor (blue) and their overlap. The upper panel shows the distribution of mutations in the 10 hypermutated cases (MBC N = 5, NSCLC N = 2, and CRPC N = 3), while the lower panel shows the same in the remaining 114 patients (MBC N = 34, NSCLC N = 39, CRPC N = 41). The 10 hypermutated cases account for 60% of total cfDNA variants and 75% of cfDNA-only variants (VUSo). (**c**) Bar charts displaying the fraction of mutational signatures in the hypermutated cases. The upper panel shows the Pearson correlation between the observed and expected 96 base substitutions profile. All the MBC cases and one of the CRPC cases demonstrated a dominant APOBEC signature. (**d**) Microsatellite instability (MSI) scores obtained using a modified MSIsensor algorithm39 from the tumor (x-axis) and cfDNA (y-axis). (**e**) A 55-year-old patient with castration- and enzalutamide-resistant prostate cancer displaying an MMR signature and high MSI score based on both cfDNA and tumor targeted sequencing data. Upon initiation of treatment on an anti-PD-L1 immunotherapy regimen, rapid tumor regression was observed. Line charts show relative tumor size based on Response Evaluation Criteria in Solid Tumors (RECIST v1.1) criteria and serum prostate-specific antigen (PSA) levels. CT images show the decreasing tumor size at indicated time points.**Fig. 4. Characterization of biological sources and composition of cfDNA variants.** (**a**) Pie charts representing the distribution of cfDNA somatic mutations. The colors indicate WBC-matched, tumor biopsy-matched, biopsy-subthreshold variants, and VUSo. (**b**) Bar plots showing the number of somatic variants detected in plasma cfDNA per megabase (Mb, y-axis) for each sample (x-axis) stratified by cancer status and biological sources and ordered by increasing number of somatic WBC-matched variants. The panels show control samples (top left) and patients with MBC (top right), NSCLC (bottom left) and CRPC (bottom right) cancers. The colors are indicated in (a). (**c**) Association between age and number of cfDNA variants categorized as WBC-matched, VUSo, tumor biopsy-matched and biopsy-subthreshold. In all panels, blue indicates control samples, and yellow indicates cancer samples. The x-axis indicates age, and the y-axis describes the number of somatic cfDNA variants per Mb. (**d**) Top mutated genes carrying WBC-matched variants for each cohort. The number in the cells indicate the overall number of variants for each gene in the corresponding cohort. (**e**) Posterior distribution of variant allele fractions (VAF). The scatter plot shows the correlation in VAFs of somatic mutations detected in cfDNA and WBC using the targeted DNA assay and color coded according to source of origin. The diagonal represents the line *y* = *x*.



**Fig. 5. Characterization of WBC variants.** (**a**) Direct analysis of somatic variants in WBC. The upper bar plot shows the number of somatic variants detected across 1.1 Mb of genome grouped by age category and ordered by increasing mutational burden. Cancer patients are displayed in yellow while control individuals are represented in blue. The lighter shade indicates WBC-only variants while the darker shade indicates somatic variants also found in the matched plasma cfDNA. The bottom panel shows the variant allele fractions (VAFs) of all somatic variants in 15 canonical genes involved in clonal hematopoiesis (CH) together with the variant occurring at maximal VAF in WBC. (**b**) Stacked bar plot showing the percentage of cancer patients and control individuals harboring a mutation with maximal VAF in a given CH gene. \* indicates p = 0.015 and \*\* indicates p = 0.0072. (**c**) Frequency of mutations in CH genes (percentages for each category) colored according to the percentage of truncating mutations including: frame-shifting indel, nonsense and nonstop mutations, relative to single nucleotide mutations. Note that for some of these patients, ≥1 variants affecting some of the canonical CH genes (e.g. *DNMT3A*, *TET2*, *PPM1D*, and *ASXL1*). The sum of the size of the circles can, therefore, exceed 100%.

**Supplementary Materials**

**Supplementary Figures**

**Supplementary Fig. 1.** Study overview

**Supplementary Fig. 2.** Analytical performance of the targeted DNA assay

**Supplementary Fig. 3.** Targeted assay reproducibility for one hypermutated MBC patient

**Supplementary Fig. 4.** Top mutated genes from cfDNA somatic variants excluding the hypermutated cases

**Supplementary Fig. 5.** 96 base substitution profiles for the 10 hypermutated cases

**Supplementary Fig. 6.** Characterization of the biological sources and composition of cfDNA variants

**Supplementary Fig. 7.** Top mutated genes carrying WBC-matched variants

**Supplementary Fig. 8.** Top mutated genes carrying VUSo

**Supplementary Fig. 9.** Association between the number of VUSo and the size of sequenced region for each gene in hypermutated cases

**Supplementary Fig. 10.** Bayesian hierarchical model for calibrated analysis of somatic cfDNA variants

**Supplementary Fig. 11.** MSIsensor analysis

**Supplementary Tables**

**Supplementary Table 1.** List of genes assayed in the cfDNA targeted panel

**Supplementary Table 2.** Baseline demographic of the different cohorts

**Supplementary Table 3**. Baseline demographic of the different cohorts

**Supplementary Table 4.** Known small variants in HD753 Structural Multiplex Reference Standard gDNA

**Supplementary Table 5.** ddPCR verification of cell-line titrations

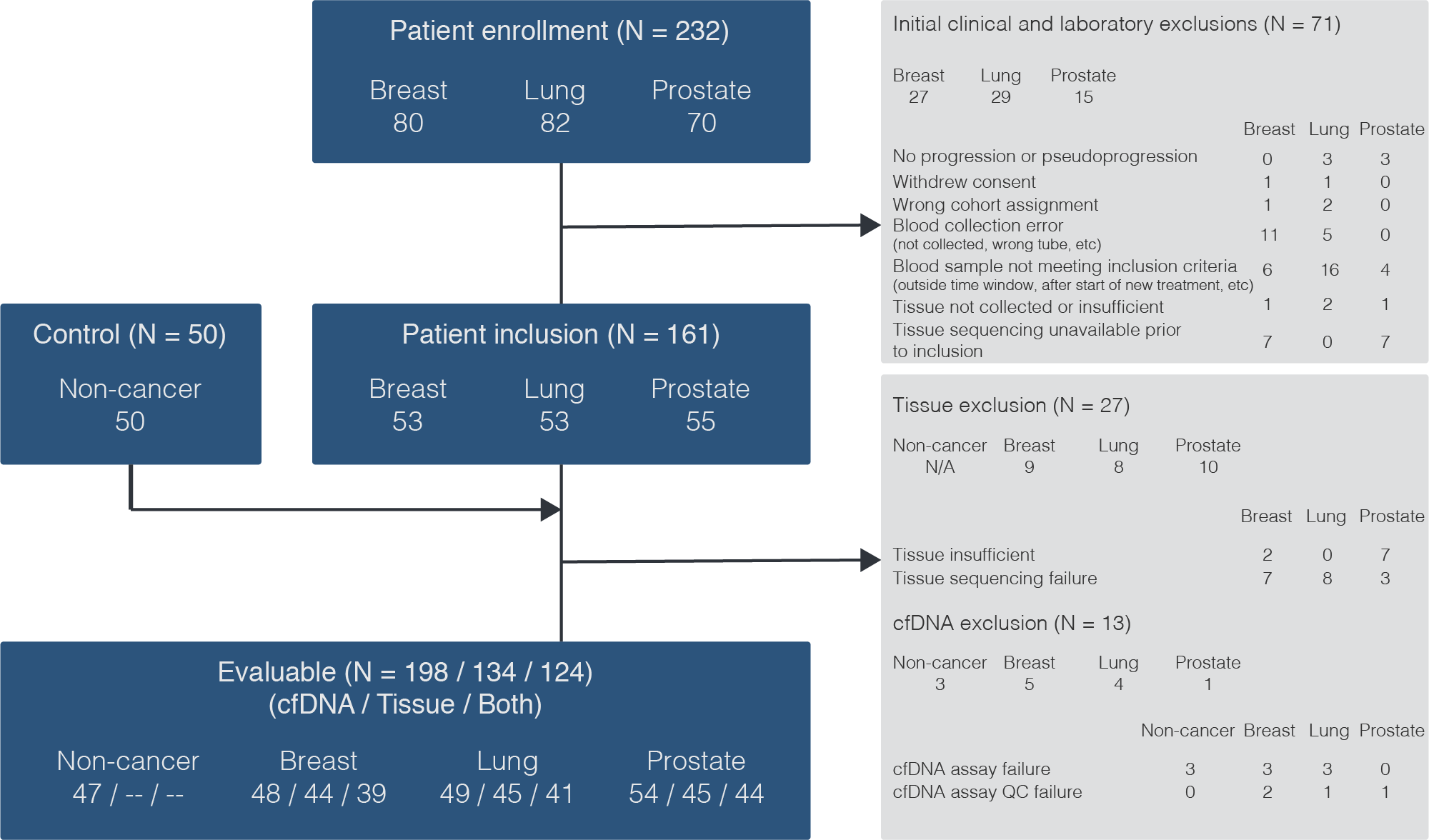
**Supplementary Table 6.** Assay reproducibility in patient samples

**Supplementary Table 7.** Somatic cfDNA mutational data for the prospective cohort

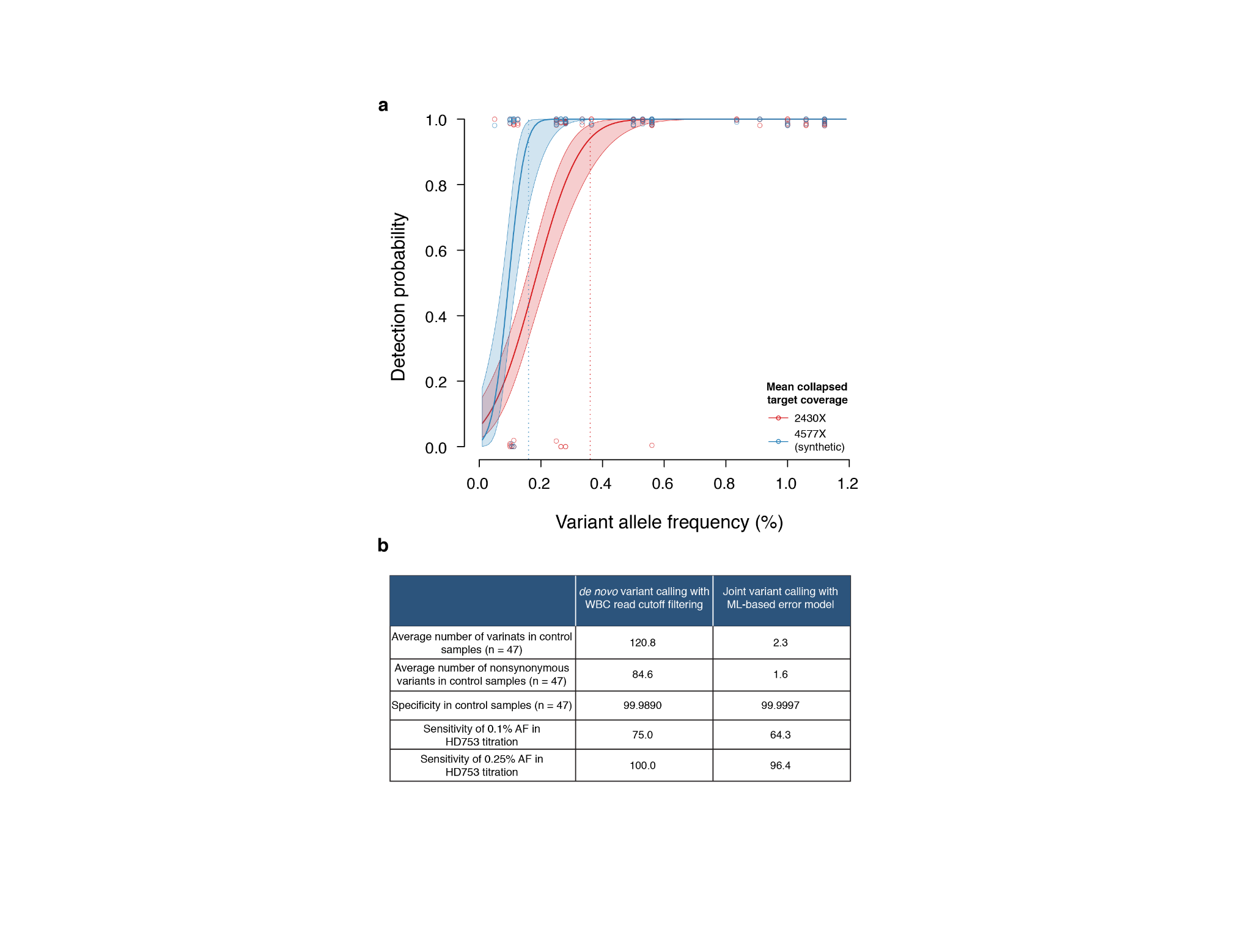
**Supplementary Table 8.** Somatic WBC mutational data for the prospective cohort

**Supplementary Table 9.** Somatic tumor tissue mutational data for the prospective cohort

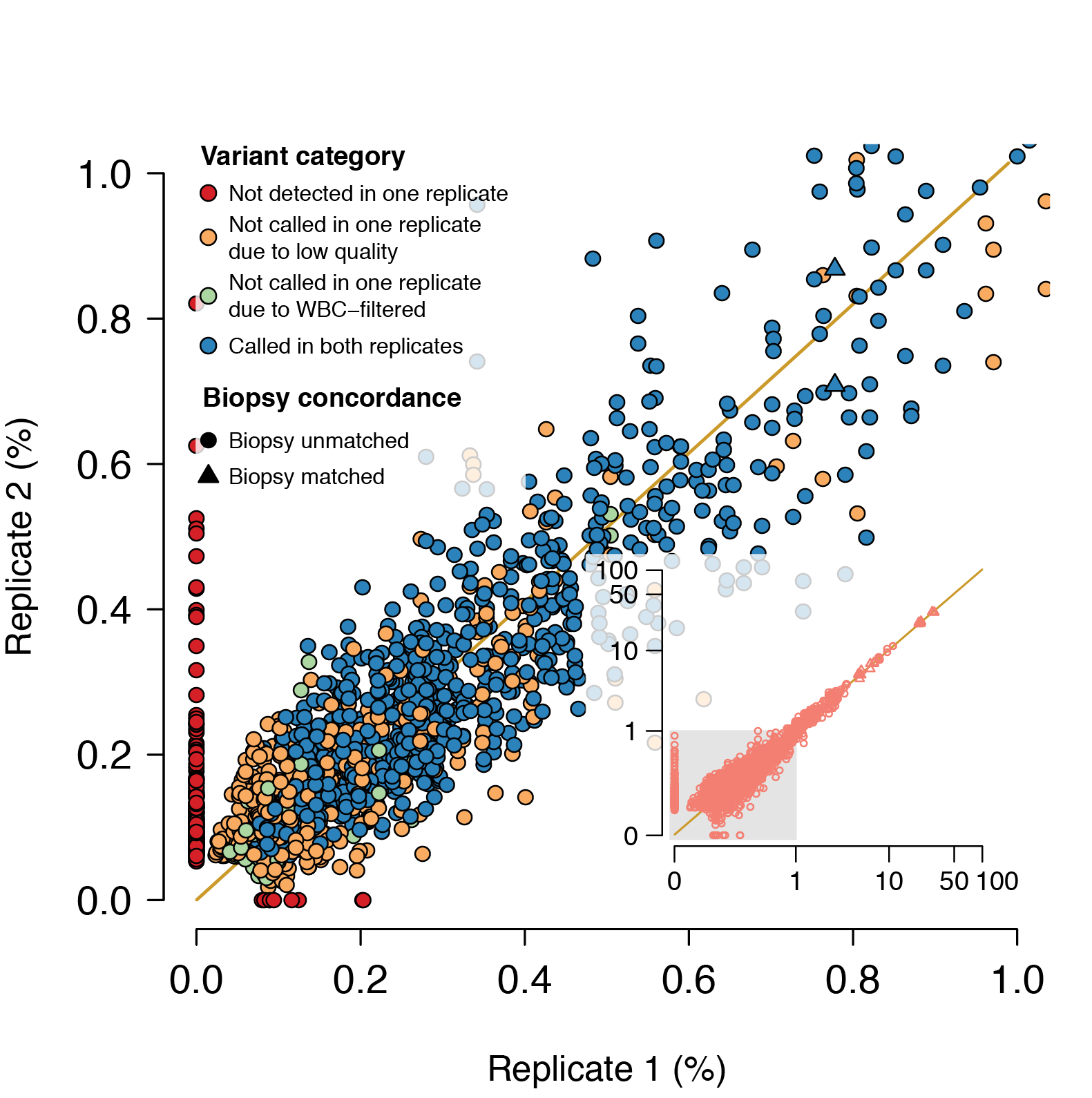
**Supplementary Figures**



**Supplementary Fig. 1.** **Study overview.** Patient enrollment, inclusion, and evaluable group are defined in the blue boxes. Clinical, tissue and cfDNA exclusions are shown in the gray boxes.



**Supplementary Fig. 2. Analytical performance of the targeted DNA assay.** (**a**) Probability of detection with increasing variant allele fraction in HD753 cell line DNA titrations. The curves show the mean target coverage of 2430X from 30 ng cell line DNA input (red) and the mean target coverage of 4577X from simulated fastqs (blue). (**b**) Estimated variant calling specificity using non-cancer control samples and corresponding variant calling sensitivity using methods as described in the methods: joint variant analysis using the machine learning error model. Non-cancer controls were not used to train the model here.

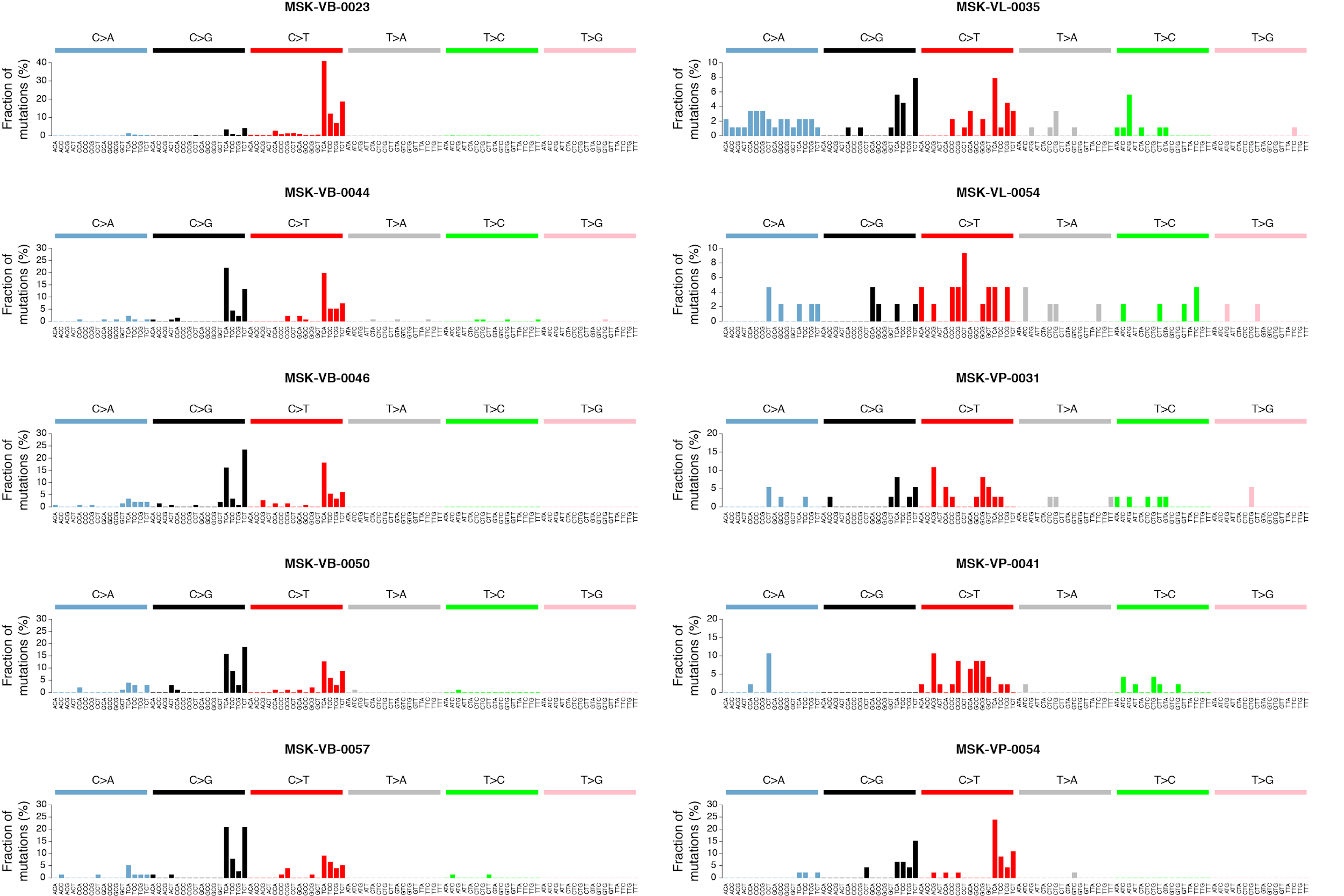


**Supplementary Fig. 3. Targeted assay reproducibility for one hypermutated MBC patient.** Comparison of the allele fraction of variants detected using either of the two targeted DNA assay protocols. Concordant variants detected in the two replicates (triangles indicate biopsy-matched, circles indicate biopsy-unmatched variants) are enriched in allele fractions above the limit of detection. The colors of the circles and triangles indicate whether the variants were detected in both replicates (blue), called in only one replicate (red), not called in one replicate due to low sample quality (yellow), or not called in one replicate due to filtering against WBC (green).

A screenshot of a cell phone

Description automatically generated

**Supplementary Fig. 4. Top mutated genes from cfDNA somatic variants excluding the hypermutated cases.** The bar plots show the frequency of genomic alterations in cfDNA of patients with MBC (top), NSCLC (middle), and CRPC (bottom) excluding the 10 hypermutated cases. The genes were sorted by their frequency of alterations. The colors indicate whether the alterations were tumor biopsy-matched, detected in the tumor but were below the threshold of MSK-IMPACT assay (biopsy-subthreshold) or were variants of unknown source (VUSo) specific to cfDNA.



**Supplementary Fig. 5. 96 base substitution profiles for the 10 hypermutated cases.** For each patient, the number of C>A, C>G, C>T, T>A, T>C, and T>G substitutions together with the sequence context immediately 3’ and 5’ are expressed as a percentage of the total number of substitutions.

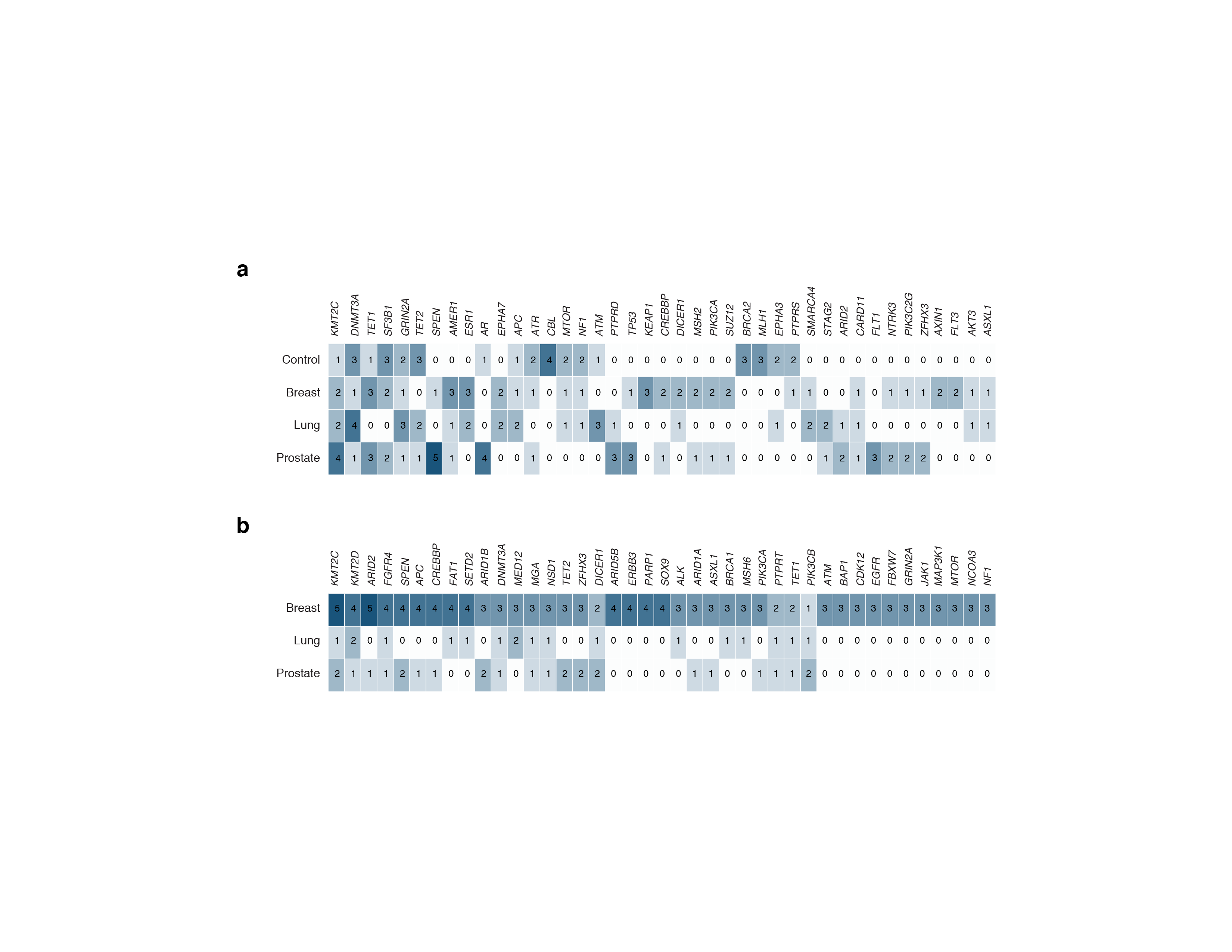
A close up of a map

Description automatically generated

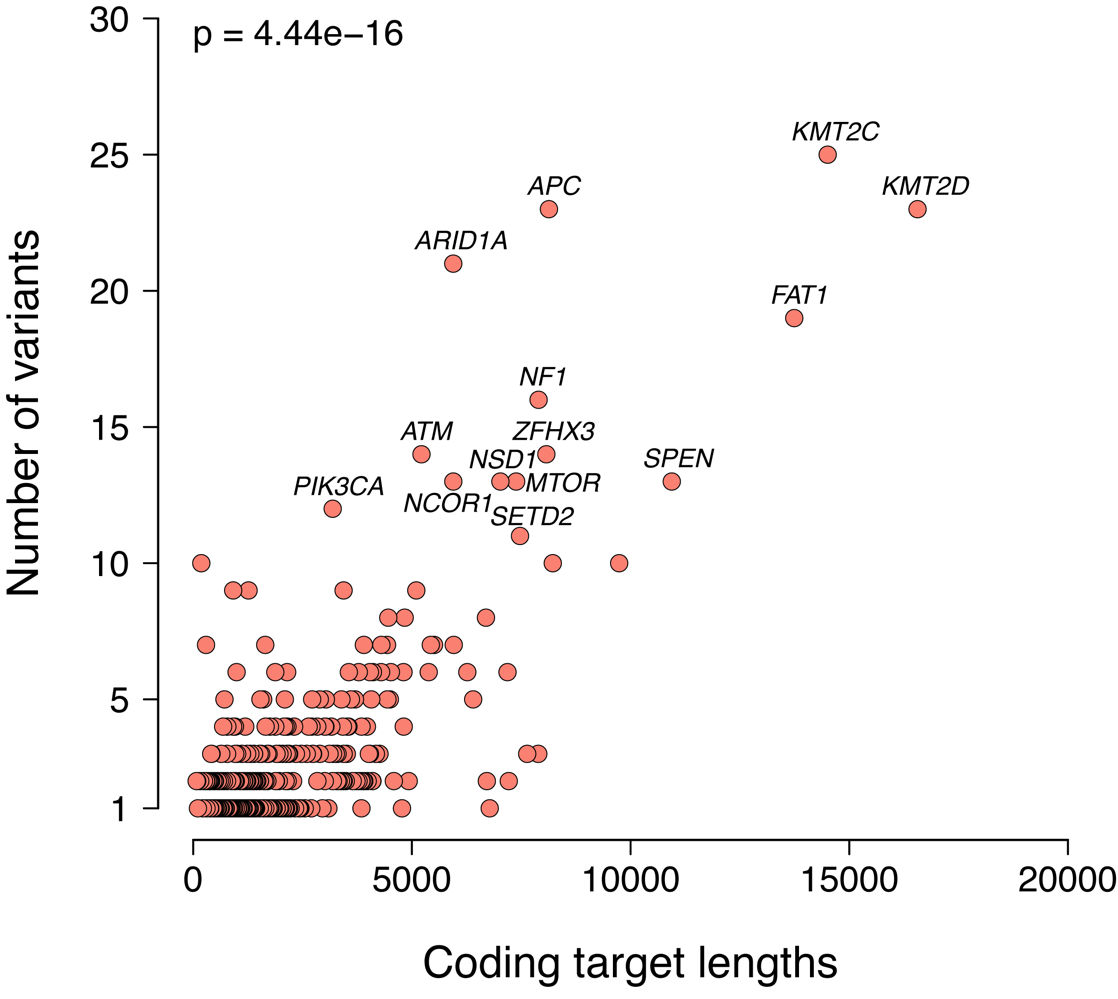
**Supplementary Fig. 6. Characterization of the biological sources and composition of cfDNA variants.** The bar plots show the number of somatic variants detected in plasma cfDNA per megabase (Mb, y-axis) for each sample (x-axis) stratified by cancer status and biological sources and ordered by increasing number of somatic WBC-matched variants. The panels show control samples (top left) and patients with MBC (top right), NSCLC (bottom left) and CRPC (bottom right). The colors indicate WBC-matched variants, tumor biopsy-matched variants, biopsy-subthreshold and VUSo.



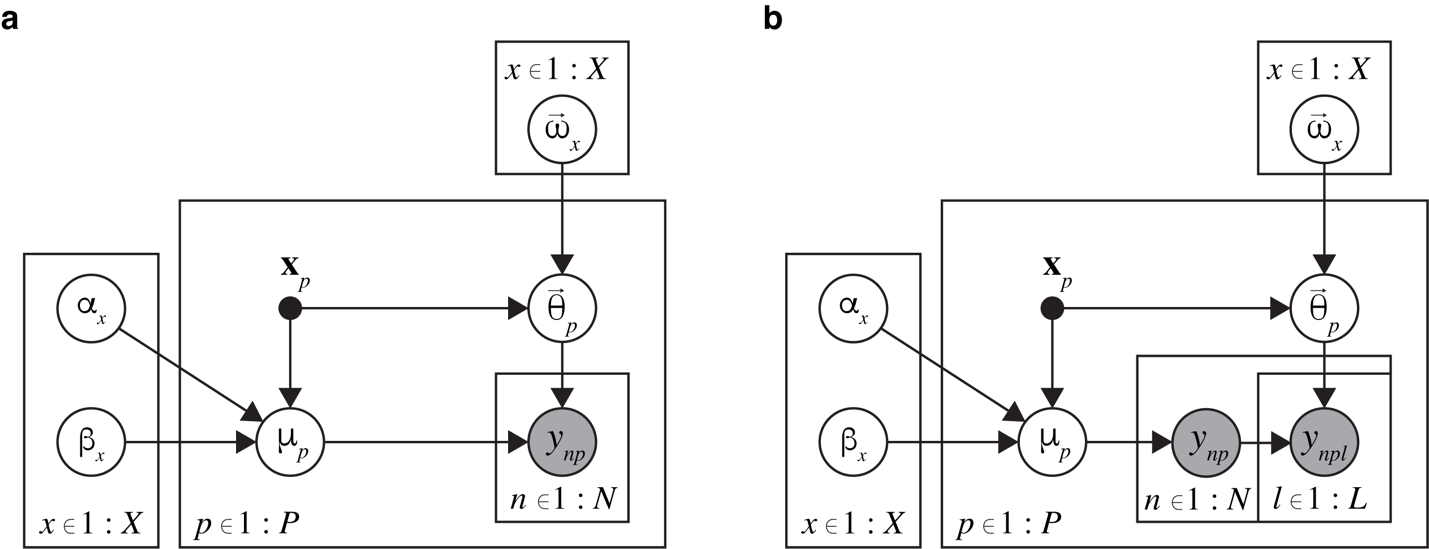
**Supplementary Fig. 7. Top mutated genes carrying WBC-matched variants.** The heat map shows the top mutated genes harboring somatic variants detected in plasma cfDNA and matched in WBC occurring in each cohort including the 10 hypermutated cases. The numbers in the cells indicate the number of somatic variants.



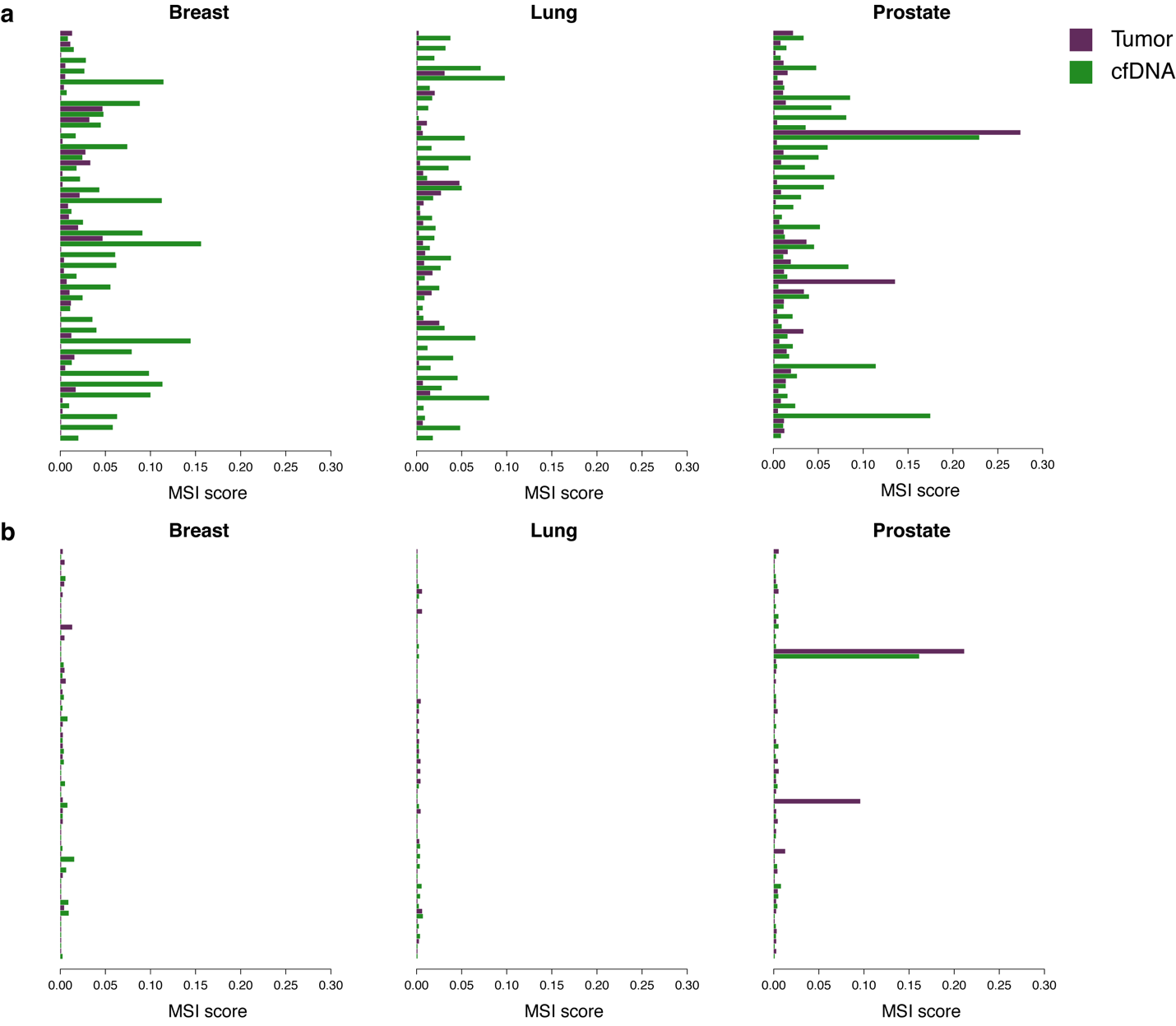
**Supplementary Fig. 8. Top mutated genes carrying VUSo.** The heat maps show the top mutated genes harboring somatic variants detected in plasma cfDNA that are neither tumor-matched (biopsy-matched or subthreshold) nor WBC-matched across each cohort in (**a**) control and non-hypermutated and (**b**) hypermutated cases. The numbers in the cells indicate the number of patients.



**Supplementary Fig. 9. Association between the number of VUSo and the size of sequenced region for each gene in hypermutated cases.** The scatter plot shows the number of VUSo per gene (y-axis) against the total length of the coding exons sequenced (x-axis) in the 10 hypermutated patients.



**Supplementary Fig. 10. Bayesian hierarchical model for calibrated analysis of somatic cfDNA variants.** (**a**) Single nucleotide variants and (**b**) small insertions and deletions. This plate model shows the hierarchy of statistical relationships influencing the observed quantity of alternate alleles (*ynp*) in each sample (n) at each position (p) conditional on both latent parameters (μ, θ, ⍺, 𝛽) as well as fixed covariates (*xp*) such as trinucleotide context, depth of sequencing at a position. Note that insertions and deletions have additional complexity as we must account for length of the insertion/deletion event in the model as insertions and deletions of differing lengths have differing probabilities. The model is fitted to the training data, estimates for the parameters are fixed and applied to new samples for scoring.



**Supplementary Fig. 11. MSIsensor analysis.** MSIsensor39 was used to compute the MSI scores of tumor biopsies and cfDNA samples using (**a**) the default algorithm settings and (**b**) parameters modified for the high depth-of-read cfDNA data. In both panels, the MSI scores are displayed on the x-axis and patients are ordered consecutively on the y-axis.

**Supplementary Tables**

**Supplementary Table 1. List of genes assayed in the cfDNA targeted panel.**

|  |
| --- |
| ABL1, *ABL2*, ACVR1, *ACVR1B*, AKT1, AKT2, AKT3, ALK, ALOX12B, ANKRD11, APC, AR, ARAF, *ARFRP1*, ARID1A, ARID1B, ARID2, ARID5B, ASXL1, ASXL2, ATM, ATR, ATRX, AURKA, AURKB, AXIN1, AXIN2, AXL, B2M, BAP1, BARD1, BBC3, BCL10, BCL2, BCL2L1, *BCL2L2*, BCL2L11, BCL6, BCOR, *BCORL1*, *BCR*, BIRC3, BLM, BMPR1A, BRAF, BRCA1, BRCA2, BRD4, BRIP1, *BTG1*, BTK, *C11orf30*, CALR, CARD11, CASP8, CBFB, CBL, CCND1, CCND2, CCND3, CCNE1, CD274, CD276, CD74, CD79A, CD79B, CDC73, CDH1, CDK12, CDK4, CDK6, CDK8, CDKN1A, CDKN1B, CDKN2A, CDKN2B, CDKN2C, CEBPA, CENPA, CHEK1, CHEK2, *CHD2*, *CHD4*, CIC, CREBBP, CRKL, CRLF2, CSF1R, CSF3R, CTCF, CTLA4, CTNNA1, CTNNB1, CUL3, CXCR4, *CYLD*, DAXX, DCUN1D1, DDR2, DICER1, DIS3, DNAJB1, DNMT1, DNMT3A, DNMT3B, DOT1L, E2F3, EED, EGFL7, EGFR, EIF1AX, EIF4A2, EIF4E, *EML4*, EP300, EPCAM, EPHA3, EPHA5, EPHA7, EPHB1, ERBB2, ERBB3, ERBB4, *ERCC1*, ERCC2, ERCC3, ERCC4, ERCC5, ERG, ERRFI1, ESR1, *ETS1*, ETV1, *ETV4*, *ETV5*, ETV6, *EWSR1*, EZH2, FAM123B, FAM175A, FAM46C, FANCA, FANCC, *FANCD2*, *FANCE*, *FANCF*, *FANCG*, *FANCI*, *FANCL*, *FAS*, FAT1, FBXW7, *FGF1*, *FGF10*, *FGF14*, FGF19, *FGF2*, *FGF23*, FGF3, FGF4, *FGF5*, *FGF6*, *FGF7*, *FGF8*, *FGF9*, FGFR1, FGFR2, FGFR3, FGFR4, FH, FLCN, *FLI1*, FLT1, FLT3, FLT4, FOXA1, FOXL2, FOXO1, FOXP1, *FRS2*, FUBP1, FYN, *GABRA6*, GATA1, GATA2, GATA3, *GATA4*, *GATA6*, *GEN1*, *GID4*, GLI1, GNA11, GNA13, GNAQ, GNAS, *GPR124*, GPS2, GREM1, GRIN2A, *GRM3*, GSK3B, H3F3A, H3F3B, H3F3C, HGF, HIST1H1C, HIST1H2BD, HIST1H3A, HIST1H3B, HIST1H3C, HIST1H3D, HIST1H3E, HIST1H3F, HIST1H3G, HIST1H3H, HIST1H3I, HIST1H3J, HIST2H3C, HIST2H3D, HIST3H3, HLA-A, HNF1A, HOXB13, HRAS, *HSD3B1*, *HSP90AA1*, ICOSLG, ID3, IDH1, IDH2, IFNGR1, IGF1, IGF1R, IGF2, IKBKE, IKZF1, IL10, IL7R, INHA, INHBA, INPP4A, INPP4B, INSR, *IRF2*, IRF4, IRS1, IRS2, JAK1, JAK2, JAK3, JUN, *KAT6A*, KDM5A, KDM5C, KDM6A, KDR, KEAP1, *KEL*, *KIF5B*, KIT, KLF4, *KLHL6*, KMT2A, *KMT2B*, KMT2C, KMT2D, KRAS, *LAMP1*, LATS1, LATS2, LMO1, *LRP1B*, *LYN*, *LZTR1*, *MAGI2*, MALT1, MAP2K1, MAP2K2, MAP2K4, MAP3K1, MAP3K13, MAP3K14, *MAP3K4*, MAPK1, MAPK3, MAX, MCL1, MDC1, MDM2, MDM4, MED12, MEF2B, MEN1, MET, MGA, MITF, MLH1, MLLT3, MPL, MRE11A, MSH2, *MSH3*, MSH6, MST1, MST1R, MTOR, MUTYH, *MYB*, MYC, MYCL1, MYCN, MYD88, MYOD1, NAB2, NBN, NCOA3, NCOR1, NEGR1, NF1, NF2, NFE2L2, NFKBIA, NKX2-1, NKX3-1, NOTCH1, NOTCH2, NOTCH3, NOTCH4, NPM1, NRAS, *NRG1*, NSD1, NTRK1, NTRK2, NTRK3, NUP93, NUTM1, PAK1, *PAK3*, PAK7, PALB2, PARK2, PARP1, *PAX3*, PAX5, *PAX7*, PAX8, PBRM1, PDCD1, *PDCD1LG2*, PDGFRA, PDGFRB, *PDK1*, PDPK1, PGR, PHOX2B, PIK3C2G, PIK3C3, PIK3CA, *PIK3C2B*, PIK3CB, PIK3CD, PIK3CG, PIK3R1, PIK3R2, PIK3R3, PIM1, PLCG2, PLK2, PMAIP1, PMS1, PMS2, PNRC1, POLD1, POLE, *PPARG*, PPM1D, PPP2R1A, *PPP2R2A*, PPP6C, *PREX2*, PRDM1, PRKAR1A, *PRKCI*, *PRKDC*, *PRSS8*, PTCH1, PTEN, PTPN11, PTPRD, PTPRS, PTPRT, *QKI*, RAB35, RAC1, RAD21, RAD50, RAD51, RAD51B, RAD51C, RAD51D, RAD52, RAD54L, RAF1, *RANBP2*, RARA, RASA1, RB1, RBM10, RECQL4, REL, RET, RFWD2, RHEB, RHOA, RICTOR, RIT1, RNF43, ROS1, RPS6KA4, *RPS6KB1*, RPS6KB2, RPTOR, RUNX1, *RUNX1T1*, RYBP, SDHA, SDHAF2, SDHB, SDHC, SDHD, SETD2, SF3B1, SH2B3, SH2D1A, SHQ1, *SLIT2*, *SLX4*, SMAD2, SMAD3, SMAD4, SMARCA4, SMARCB1, SMARCD1, SMO, *SNCAIP*, SOCS1, *SOX10*, SOX17, SOX2, SOX9, SPEN, SPOP, *SPTA1*, SRC, SRSF2, STAG2, STAT3, STAT4, STAT5A, STAT5B, STK11, STK40, SUFU, SUZ12, SYK, *TAF1*, TBX3, TCEB1, TCF3, TCF7L2, *TERC*, TERT, TET1, TET2, TFE3, *TFRC*, TGFBR1, TGFBR2, TMEM127, TMPRSS2, TNFAIP3, TNFRSF14, TOP1, *TOP2A*, TP53, TP63, TRAF2, TRAF7, TSC1, TSC2, TSHR, U2AF1, VEGFA, VHL, VTCN1, *WISP3*, WT1, XIAP, XPO1, XRCC2, YAP1, YES1, *ZBTB2*, ZFHX3, ZRSR2, *ZNF217*, *ZNF703* |

Genes in italic are specific to the cfDNA panel and were not considered to evaluate concordance with the tumor

biopsy

**Supplementary Table 2. Baseline demographic of the different cohorts.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient characteristics** | **Breast**  **(n=39)** | **Lung**  **(n=41)** | **Prostate**  **(n=44)** | **Control**  **(n=47)** |
| Mean age at enrollment (SD) | 60 (30-79) | 67 (33-83) | 67 (46-87) | 61 (30-78) |
| Age groups, N (%) |  | | | |
| ≤50 | 11 (28.2%) | 5 (12.2%) | 2 (4.5%) | 15 (31.9%) |
| >50 and ≤60 | 10 (25.6%) | 5 (12.2%) | 10 (22.7%) | 8 (17%) |
| >60 and ≤70 | 15 (38.5%) | 16 (39%) | 16 (36.4%) | 11 (23.4%) |
| >70 | 3 (7.7%) | 15 (36.6%) | 16 (36.4%) | 13 (27.7%) |
| Gender, N (%) | 0 (0%) | 13 (31.7%) | 44 (100%) | 23 (48.9%) |
| Male |
| Prior chemotherapy, any setting, N (%) | 26 (66.7%) | 12 (29.3%) | 12 (27.3%) |  |
| Prior radiotherapy, any setting, N (%) | 24 (61.5%) | 7 (17.1%) | 37 (84.1%) |
| Lines of therapies in metastatic setting, N (%) |  | | |
| 0 | 20 (51.3%) | 25 (61%) | 12 (27.3%) |
| 1 | 2 (5.1%) | 9 (22%) | 14 (31.8%) |
| 2 | 2 (5.1%) | 3 (7.3%) | 9 (20.5%) |
| ≥3 | 15 (38.5%) | 4 (9.6%) | 9 (20.5%) |
| Tumor tissue underwent sequencing, N (%) |  | | |
| Metastatic | 35 (89.7%) | 28 (63.3%) | 44 (100%) |
| Primary | 4 (10.3%) | 13 (31.7%) | 0 (0%) |
| Overall stage at diagnosis, N (%) |  | | |
| I | 9 (23.1%) | 1 |  |
| II | 8 (20.5%) | 0 |  |
| III | 8 (20.5%) | 0 |  |
| IV | 14 (35.9%) | 40 |  |
| Breast cancer receptor status, N (%) |  | | |
| HR+/HER2- | 26 (66.7%) |  | |
| HER2+ | 5 (12.8%) |
| Triple Negative | 8 (20.5%) |
| Breast cancer histology, N (%) |  | | |
| IDC | 32 (82.1%) |  | |
| ILC | 2 (5.1%) |
| Mixed | 5 (12.8%) |
| Lung cancer histology, N (%) |  | | |
| Adenocarcinoma |  | 38 (92.7%) |  |
| Other | 3 (7.3%) |
| Prostate cancer histology, N (%) |  | | |
| Adenocarcinoma |  | | 39 (88.6%) |
| Neuroendocrine/ small cell | 5 (11.4%) |

**Supplementary Table 3. Baseline demographic of the different cohorts.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Subject Type** | **Source** | **Mean** | **Median** | **5%**  **Percentile** | **95%**  **Percentile** | **p-value†** | **Mean**  **AF (%)** | **Median**  **AF (%)** | **5% quantile** | **95% quantile** |
| Control | WBC-  matched | 6 | 5 | 0 | 13 | NA | 0.60 | 0.16 | 0.06 | 2.29 |
| Breast | WBC-  matched | 6 | 5 | 0 | 14 | 0.889 | 0.62 | 0.19 | 0.05 | 1.96 |
| Lung | WBC-  matched | 8 | 6 | 2 | 17 | 0.039 | 0.69 | 0.20 | 0.06 | 3.09 |
| Prostate | WBC-  matched | 7 | 5 | 1 | 16 | 0.213 | 0.64 | 0.18 | 0.06 | 2.22 |
| Control | VUSo | 1 | 1 | 0 | 4 | NA | 0.58 | 0.14 | 0.07 | 0.36 |
| Breast | VUSo | 3 | 2 | 0 | 10 | 0.003 | 1.89 | 0.37 | 0.09 | 8.48 |
| Lung | VUSo | 4 | 3 | 0 | 11 | 0.001 | 0.90 | 0.29 | 0.08 | 1.65 |
| Prostate | VUSo | 4 | 2 | 0 | 15 | 0.002 | 1.84 | 0.28 | 0.08 | 10.79 |
| Breast | Biopsy-  matched | 4 | 3 | 1 | 8 | NA | 11.19 | 3.98 | 0.30 | 42.78 |
| Lung | Biopsy-  matched | 5 | 3 | 0 | 14 | NA | 6.65 | 3.37 | 0.15 | 22.05 |
| Prostate | Biopsy-  matched | 3 | 2 | 0 | 5 | NA | 16.20 | 13.37 | 0.15 | 42.04 |

**†** Mann-Whitney *U* test

**Supplementary Table 4. Known small variants in HD753 Structural Multiplex Reference Standard gDNA.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene** | **HGVSp** | **Variant Type** | **Stock VAF (%)** |
| *PIK3CA* | p.Glu545Lys | SNV | 5.6 |
| *PIK3CA* | p.His1047Arg | SNV | 16.7 |
| *FBXW7* | p.Ser588ValfsTer39 | Indel | 5.6 |
| *BRAF* | p.Val600Glu | SNV | 18.2 |
| *EGFR* | p.Gly719Ser | SNV | 5.3 |
| *EGFR* | p.Ala767\_Val769dup | Indel | 5.6 |
| *EGFR* | p.Glu746\_Ala750del | Indel | 5.3 |
| *MET* | p.Leu238TyrfsTer25 | Indel | 2.5 |
| *NOTCH1* | p.Pro668Ser | SNV | 5.0 |
| *KRAS* | p.Gly13Asp | SNV | 5.6 |
| *BRCA2* | p.Lys1691AsnfsTer15 | Indel | 5.6 |
| *FLT3* | p.Pro986AlafsTer27 | Indel | 5.6 |
| *AKT1* | p.Glu17Lys | SNV | 5.0 |
| *GNA11* | p.Gln209Leu | SNV | 5.6 |

**Supplementary Table 5. ddPCR verification of cell-line titrations.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***KRAS* G12D** | | ***PIK3CA* H1047R** | |
| **Dilution** | **Expected VAF (%)** | **Measured VAF (%)** | **Expected VAF (%)** | **Measured VAF (%)** |
| 20% HD753 / 80% NA12878 | 1.12 | 1.13 | 3.34% | 3.24 |
| 10% HD753 / 90% NA12878 | 0.56 | 0.61 | 1.67% | 1.54 |
| 5% HD753 / 95% NA12878 | 0.28 | 0.30 | 0.84% | 0.84 |
| 2% HD753 / 98% NA12878 | 0.11 | 0.15 | 0.33% | 0.44 |
| 0% HD753 / 100% NA12878 | 0.00 | 0.01 | 0.00% | 0.01 |

**Supplementary Table 6. Assay reproducibility in patient samples.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Patient ID** | **Cancer type** | **Number of tumor variants†** | **cfDNA yield (ng)** | **ddPCR of cfDNA††** |
| MSK-VB-0050 | Breast | 22 | 11420 | *PIK3CA* H1047R |
| MSK-VB-0041 | Breast | 8 | 6880 | None |
| MSK-VL-0028 | Lung | 4 | 1201 | *EGFR* L861Q |
| MSK-VL-0042 | Lung | 7 | 1199 | *KRAS* G12C |
| MSK-VB-0023 | Breast | 21 | 11840 | *PIK3CA* E542K |
| MSK-VL-0038 | Lung | 9 | 231 | *KRAS* G12A |

**†** Number of tumor variants identified by MSK-IMPACT; **††** Selected hotspot variants assayed in cfDNA by ddPCR

**Supplementary Table 7. Somatic cfDNA mutational data for the prospective cohort.**

Provided as an external file

**Supplementary Table 8. Somatic WBC mutational data for the prospective cohort.**

Provided as an external file

**Supplementary Table 9. Somatic tumor tissue mutational data for the prospective cohort.**

Provided as an external file