HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2020 May 25.

Published in final edited form as:

Nat Med. 2019 December ; 25(12): 1928–1937. doi:10.1038/s41591-019-0652-7.

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

Pedram Razavi[1,2,9,*], **Bob T. Li**[1,9], **David N. Brown**[3,9], **Byoungsok Jung**[4], **Earl Hubbell**[4], **Ronglai Shen**[5], **Wassim Abida**[1], **Krishna Juluru**[6], **Ino De Bruijn**[7], **Chenlu Hou**[4], **Oliver Venn**[4], **Raymond Lim**[3], **Aseem Anand**[1], **Tara Maddala**[4], **Sante Gnerre**[4], **Ravi Vijaya Satya**[4], **Qinwen Liu**[4], **Ling Shen**[4], **Nicholas Eattock**[4], **Jeanne Yue**[4], **Alexander W. Blocker**[4,†], **Mark Lee**[4,§], **Amy Sehnert**[4,¥], **Hui Xu**[4], **Megan P. Hall**[4], **Angie Santiago-Zayas**[1], **William F. Novotny**[4,‡], **James M. Isbell**[8], **Valerie W. Rusch**[8], **George Plitas**[8], **Alexandra S. Heerdt**[8], **Marc Ladanyi**[3], **David M. Hyman**[1], **David R. Jones**[8], **Monica Morrow**[8], **Gregory J. Riely**[1], **Howard I. Scher**[1], **Charles M. Rudin**[1], **Mark E. Robson**[1], **Luis A. Diaz Jr.**[1], **David B. Solit**[1,2,7], **Alexander M. Aravanis**[4], **Jorge S. Reis-Filho**[2,3,*]

[1]Memorial Sloan Kettering Cancer Center, Department of Medicine, and Weill Cornell Medical College, New York, NY.

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

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

[4]GRAIL, Inc. Menlo Park, CA.

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

[6]Memorial Sloan Kettering Cancer Center, Department of Radiology, New York, NY.

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

[8]Memorial Sloan Kettering Cancer Center, Department of Surgery, and Weill Cornell Medical College, New York, NY.

[9]Equal Contribution

[*]Correspondence to: razavip@mskcc.org, reisfilj@mskcc.org.
[†]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.

Author Contributions

Conceived the study: P.R., B.T.L., D.B.S., A.M.A., J.S.R-F; Data acquisition: P.R., B.T.L., B.J., W.A., K.J., C.H., A.A., R.V.S., K.L., L.S., N.E., J.Y., H.X., M.P.H., A.S.-Z, W.F.N, J.M.I., V.W.R., G.P., M.L., A.S., A.S.H., M.L., D.M.H., D.R.J., M.M., G.J.R., H.I.S., C.M.R., M.E.R., L.A.D., D.B.S., A.M.A.; Data analysis and interpretation: P.R., D.N.B., E.H., R.S., I.D.B., O.V., R.L., T.M., Q.L., A.W.B., A.M.A., J.S.R-F. Bioinformatics and genomic analysis: P.R., D.N.B., E.H., R.S., I.D.B., O.V., S.G., A.W.B., A.M.A., J.S.R-F. Manuscript first draft: P.R., D.N.B., E.H., M.P.H., A.M.A., J.S.R-F wrote the manuscript with input from all authors. Manuscript review and approval: all authors.

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 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 revealed that CH constitutes a pervasive biological phenomenon emphasizing the importance of matched cfDNA-WBC sequencing for accurate variant interpretation.

Introduction

Circulating cell-free DNA (cfDNA) in the plasma of cancer patients constitutes a potential source of tumor-derived DNA[1,2]. 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, vascularization, proliferation and apoptosis rates[3,4]. ctDNA fraction is extremely low in many early-stage and some metastatic cancers[5,6], requiring methods to detect mutations at extremely low allele fractions[7]. 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 depths[8–10], a large number of genes at moderate sequencing depths[11–13], or a combination of methods to define ctDNA fraction using shallow whole-genome sequencing or targeted methods followed by whole-exome analysis of samples with a high ctDNA fraction[6,14,15].

Even when accurate cfDNA assays are utilized, cfDNA sequencing results may still be confounded by biological signals arising from somatic mosaicism[16]. One 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 their progeny[17]. These somatic mutations may provide a fitness advantage to some HSCs and/or their descendant cells, resulting in their disproportionate expansion[8,18–20], or arise through neutral drift[21]. CH increases with age and occurs in up to 31% of older individuals[10,20,22–26], and can also be detected in cfDNA sequencing analysis[27]. In this context, it may confound the interpretation of cfDNA sequencing, particularly because a large proportion of the cfDNA fragments originate from hematopoietic cells[28].

Comparisons of 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 disease[6,9,15,29–32]. Additional somatic variants not present in tumor biopsies but in cfDNA only have also been

documented[4]; their nature and source (tumor-derived vs. other sources), however, have yet to be defined.

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-authorized targeted sequencing assay, 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 progressive metastatic 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 regions 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 in a CLIA-certified environment using the Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) assay, an FDA-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)[33,34]. 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 potentially eligible cancer patients (53 MBC, 53 NSCLC and 55 CRPC) enrolled between September 24, 2015-August 01, 2016, 124, 39 MBCs, 41 NSCLCs and 44 CRPCs were included in the concordance subset (evaluable for both tumor tissue and cfDNA analysis, Extended Data Fig. 1). The clinicopathologic features of the cohort were consistent with those of a contemporary prospectively collected cohort of consecutive cases of these malignancies (Supplementary Table 2). Of the 50 non-cancer control samples, three failed cfDNA sequencing resulting in 47 evaluable samples.

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 assay spanning approximately 2 Mb and utilizing unique molecular identifiers (UMI) to suppress technical assay errors at a minimum raw average target depth of 60,000X (Extended Data Fig. 2). A

joint-variant-calling of plasma cfDNA and WBC gDNA variants was performed utilizing a machine learning-based noise model (Supplementary Methods, Extended Data Fig. 3a–b). Together, this resulted in an assay with performance characteristics necessary for the detection of mutations near the molecular limits (high technical sensitivity; Fig. 1b–c), low false positive (<1 error in one million bases sequenced; Extended Data Fig. 3c–h), high reproducibility in independent biological replicates (Fig. 1d–e, Extended Data 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 per base error rate ranging from 1×10^{-5} to 3×10^{-5} (Supplementary Methods, Extended Data Figs. 2–3, Supplementary Tables 3–5), which is comparable to that of other high-fidelity cfDNA sequencing assays[25,35,36].

Given these assay characteristics, we first sought to define the performance of the high-intensity 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%). The large genomic footprint of the assay 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 Extended Data 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)[33,37], or were neither detected in the tumor biopsy nor WBCs (variants of unknown source; VUSo).

Given the low false positive rates and accuracy of the assay for measuring variant allele fractions (VAFs; Fig. 1d–f, Extended Data Fig. 4), 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 occurred 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 bioinformatically inferred to harbor a given mutation in a particular patient; Methods), with mutations defined as clonal being significantly more frequently detected than subclonal ones (p= 5.33e-21; Fig. 2e). Additionally, the ctDNA fraction (the fraction of tumor-derived cfDNA) in metastatic cancer patients varied significantly according to tumor type (p=0.0046; Fig. 2f). To assess the association between disease burden and ctDNA fraction, volumetric assessment of the disease burden was performed (Methods). We found a significant association between the estimated disease volume based on CT scans and the ctDNA fraction (Fig. 2g) in both MBC (n=34, p=1.03e-4) and NSCLC (n=29, p=0.042) patients. For CRPC patients, who had dominant bone metastases and undergone bone scans, an assessment of the association between the automated bone scan index (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 a high sensitivity for detection of tumor-derived somatic mutations and identifies mutations subclonally present in tumor tissues. VUSo, on the other hand, which are detected in neither tumor biopsy nor WBC may derive from multiple origins but comprise a set of alterations from which a subset may reflect ongoing tumor evolution and heterogeneity not captured in a small and anatomically constrained biopsy.

Tumor mutation burden and mutational signatures

We performed an exploratory analysis to determine whether tumor mutation burden (TMB), mutational signatures[38,39], and MSI score[40] 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[33]), 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). cfDNA analysis identified six additional cases with a high TMB not detected as hypermutators by MSK-IMPACT analysis of the tumor biopsy (Fig. 3a, a total of ten cases henceforth referred to as hypermutated). 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 known to occur in MBCs, NSCLCs, and CRPC. All hypermutated MBCs (n=5) and one of the three hypermutated CRPCs displayed the APOBEC mutational signatures[38,39], known to be acquired in the evolution of MBCs and CRPCs[41,42]. Consistent with the results of previous analysis of NSCLCs[43], the mutational signatures of the two hypermutated NSCLCs comprised the smoking-related signature and a combination of other signatures, including APOBEC, homologous recombination DNA repair-deficiency (HRD), and loss-of-function of mismatch repair (MMR; Fig. 3c and Extended Data Fig. 5f).

High microsatellite instability (MSI-H) is a biomarker of response to immune-checkpoint inhibitors[44]. We therefore assessed the MSI status of the cohort utilizing MSIsensor[40],

adjusted for the ultra-high sequencing depth of cfDNA, tumor biopsy and matched normal WBCs (Supplementary Methods, Supplementary Fig. 1). Our 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). This CRPC patient[33] 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 can 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 with resulting implications for treatment selection.

Characterization of the biological sources of cfDNA variants

Despite the specificity of the assay for somatic mutation detection, tumor-matched alterations (biopsy-matched and biopsy-subthreshold) accounted for only 24.4% (739/2983 mutations) of all somatic mutations detected in the cfDNA of cancer patients (Fig. 4a). Notably, 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 analysis[45,46], based on the specificity of our 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).

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 for which there was no evidence of the mutational event in the matched tumor biopsy. 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%, 918/1727) 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, Extended Data Fig. 6a). As CH is related to age[23], we examined the association of age with the number of somatic DNA variants in the cfDNA samples from individual participants. As expected, the number of WBC-matched variants, but not the number of the biopsy-matched or biopsy-subthreshold variants, significantly correlated with age (smoking-adjusted p=3.86e-27; 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 observations in a non-cancer population[27] and with the notion that these mutations constitute CH events, the vast majority of the WBC-matched somatic mutations detected in cfDNA of cancer patients and non-cancer controls involved canonical CH genes, such as *DNMT3A, TET2, PPM1D* and *TP53* (Fig. 4d, Extended Data Fig. 6b)[10], while some of these pathogenic alterations affected cancer genes other than the canonical CH genes (Supplementary Table 6). Additionally, the VAFs of WBC-matched cfDNA

variants were correlated with their VAFs in the WBCs (Fig. 4e, Extended Data Fig. 6c–d), making it unlikely that they resulted from systematic sequencing errors or background noise.

Previous studies of hypermutated tumors have demonstrated significant spatial heterogeneity resulting in numerous subclonal mutations private to each tumor site[47]. Consistently, the overall proportion of tumor-matched mutations (biopsy-matched and biopsy-subthreshold) was significantly lower in the 10 hypermutated patients (17.2% [216/1210] non-WBC-matched mutations) than in the 114 non-hypermutated patients (30.3% [523/1727]; p=1.2e-16), whereas a higher proportion of tumor-derived variants were subclonal biopsy-subthreshold variants in hypermutated cases (41.2% [89/216] vs. 15.3% [80/523] respectively; p=1.7e-13). These findings support the notion that a single tumor biopsy may not always capture the full landscape of tumor mutational profile in patients whose tumors harbor a hypermutator phenotype.

Given the correlation between sequencing depth and occurance of technical artifacts, we next investigated whether mutations detected in cfDNA at ultra-high sequencing depths could be attributed to residual sequencing noise. Of 215 mutations detected at a collapsed depth of >10,000X, 121 (56.3%) and 20 (9.3%) mutations were identified in two known hypermutated patients (Extended Data Fig. 7a). The variant level collapsed depth of somatic mutations was a function of the mean collapsed target coverage in cfDNA and the amount of input DNA (Extended Data Fig. 7b–c). Furthermore, there was no association between the VAF and the sequencing depth of variants irrespective of source of origin (Extended Data Fig. 7d). These mutations were also highly replicable and could not be attributed to copy number gains or amplification of the corresponding loci (Extended Data Fig. 7e–h). Overall, these results indicate that the mutations found at high sequencing depths were unlikely to constitute sequencing artifacts.

We next sought to define the biological source of nonsynonymous VUSo in cfDNA (Methods). After removing variants with known source-of-origin (WBC-matched; biopsy-matched and biopsy-subthreshold in cancer patients), 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% (994/1280) of the VUSo were detected in the 10 hypermutated cancer samples. In fact, VUSo accounted for 82.1% (994/1210) of the total non-WBC-matched somatic cfDNA mutations in hypermutated samples as compared with 35.4% (286/809) 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, Extended Data Fig. 5a). 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, Extended Data Fig. 5a)[37]. Further, the VAF distribution for these mutations mostly mirrored that of biopsy-matched variants (Fig. 4e, Extended Data Fig. 6c–d). 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 the given gene was observed (p=4.4e-16; Extended Data Fig. 5b). To determine the accuracy of the cfDNA assay for detecting VUSo, we performed orthogonal validation of our results utilizing ddPCR assays targeting VUSo (Methods, Fig. 4f) and found complete agreement between the two assays. Based on these results, we posit that these VUSo 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 also included canonical CH genes (Extended Data Fig. 5c). Consistent with the notion that at least a subset of VUSo arise from CH or other sources of somatic mosaicism not present in matched WBC samples, VUSo were weakly associated with age at sample collection (p=0.141; Fig. 4c), affected canonical CH genes in both cancer patients and controls (Extended Data Fig. 5c–e), with some having similar allele frequencies as WBC-matched variants (Fig. 4e, Extended Data Fig. 6c–d).

Taken together, this high-intensity cfDNA sequencing assay identified CH mutations as the most probable origin of non-tumor derived mutations detected in cfDNA, that CH is likely more prevalent than previously reported with lower-depth WBC sequencing approaches[10,20,22–24,26], and that subclonal tumor-derived mutations absent in the tumor biopsy can be detected in cfDNA.

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). At least one CH mutation was detected in 99.1% of the WBCs of the cancer patients analyzed, and in 93.6% of the non-cancer controls. 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 significantly correlated with age (p=6.09e-64, Fig. 5b). 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. 5c).

Consistent with previous studies suggesting that therapeutic interventions may result in the acquisition of specific types of CH events[20,48], 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, mutations affecting *PPM1D*, in particular truncating variants preferentially affecting the C-terminal domain (Fig. 5d, Extended Data Fig. 8), were significantly more common in patients who received chemotherapy and/or radiation therapy than in those who had no prior history of such treatments (age- and smoking-adjusted p=0.0008, Fig. 5c).

Gene copy number variation detection

As an exploratory analysis, we sought to define whether the high-intensity cfDNA assay would be able to detect copy number variations (CNVs) *de novo*. We observed a relatively good concordance between CNVs detected in tumor biopsies and cfDNA only in cases where the ctDNA fractions were ≥10% (Extended Data Fig. 9). Despite this limitation, in five patients with actionable CNVs (n=4 *ERBB2* amplified MBCs and n=1 *MET* amplified NSCL), three of the *ERBB2* amplifications could be detected *de novo*. In the two cases where actionable CNVs were present in the MSK-IMPACT tumor biopsy but not in cfDNA, the ctDNA fractions were 1.3% and 1.9% (Extended Data Fig. 10). None of the remaining samples tested harbored amplifications of these two genes thereby demonstrating the specificity. The performance of the cfDNA assay to detect of gene amplifications in cfDNA, however, was found to be highly dependent on the ctDNA fraction.

Discussion

Most clinical cfDNA assays in current use target 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 in 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 variants[45,46]. 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 coupled with matched WBC sequencing performed at similar depths to those employed for the cfDNA analysis. Without taking into account the results of WBC sequencing, cfDNA sequencing, as often currently performed in the clinical setting, might be misleading, given that some CH mutations affecting cancer genes may be interpreted as tumor-derived mutations (e.g. *TP53* mutations).

The high-intensity cfDNA sequencing approach allowed for robust *de novo* detection of somatic mutations with a sensitivity similar to that of ddPCR (Fig. 1f) and was comparable to previous high-depth targeted sequencing efforts[49–51], allowing for the detection in cfDNA of 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), this cfDNA assay also allowed for the *de novo* assessment of tumor mutational burden and mutational signatures, including MSI (Fig. 3a–e), in 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 gDNA samples, which were present in the vast majority of non-cancer controls and cancer patients. These WBC-matched mutations preferentially affected genes previously implicated in

CH[10,20,22–24,26] and their presence was strongly associated with age at collection of the blood sample. 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% of cancer patients) relative to that reported in prior studies[10,20,22–24,26] likely resulted from the high sensitivity of the assay employed to detect variants in the WBC samples[52,53] and was consistent with the observation by Liu *et al.* in a non-cancer population[27]. In our study, however, both cfDNA and WBC samples were ultra-deep sequenced at comparable raw depths, allowing for the distinction between CH and tumor-derived mutations. Although the genes recurrently affected by these somatic genetic alterations were genes previously implicated in CH, the majority of the 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 studies[25,28,54] have demonstrated in a limited number of patients that a large proportion of somatic variants in WBCs can also be identified in cfDNA, resulting in the detection of 'false-positive' tumor-derived mutations in cfDNA. These findings provide a plausible explanation for the inconsistent results between cfDNA and tumor tissue assays, which may be due to a subset of non-tumor origin (e.g. CH) cfDNA variants being interpreted as tumor-derived[25,28,54] and support the need for joint analysis of cfDNA and matched WBC, given that mutations related to CH may result in inaccurate tumor mutation burden and mutational signature quantification[7,28].

We also demonstrated that VUSo 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. The majority of the observed VUSo were found to be tumor-derived and arose from minor tumor subclones, and 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. Hence, 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; a recent study, however, has demonstrated the importance of cfDNA sequencing in defining the heterogeneity and mechanisms of therapeutic resistance in advanced colorectal cancers[15]. 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 captured[9]. This caveat, however, would remain regardless of the number of sites biopsied. Non-cancer controls were from a different source and were processed in different batches from the tumor samples, potentially affecting results. Given that the number of samples in each tumor subgroup was relatively small, the analysis performed here may not have captured the full clinical or genomic diversity of MBCs, NSCLCs, and CRPCs, and their respective subtypes. Additionally, <50 baseline samples from healthy control were employed to train the hierarchical Bayesian model, which might be further improved through the analysis of additional samples from healthy donors 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 on the patients included in this study. Finally, the cost of this high-intensity cfDNA sequencing assay may preclude its broader adoption in the clinical context at present.

Despite these limitations, the high-intensity cfDNA sequencing assay described here represents an advancement 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.

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 variants[33,34] 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 (Memorial Sloan Kettering Cancer Center [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 the 161 cancer 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, an FDA-authorized hybridization capture-based next-generation sequencing assay, which analyzes all protein-coding exons of 410 cancer-associated genes (Supplementary Table 1), as previously described[33,34]. Average sequencing coverage across all tumors was greater than 900X. Somatic mutations, DNA copy number alterations, and structural rearrangements were identified as previously described[34] and all mutations were manually reviewed. 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 patients had complete MSK-IMPACT results (Extended Data Fig. 1).

In addition to the gene-level amplification and deletion calls generated by the MSKCC Diagnostic Molecular Pathology pipeline, genome-wide total and allele-specific DNA copy numbers were determined using the FACETS algorithm[56] 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 described[57]. 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 MSKCC (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 the 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 for variant calling as required for clinical reporting. 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 of somatic mutations detected in the matched tumor biopsy sample using the FACETS algorithm as previously described[56], and then derived the ctDNA fraction 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 active metastases are difficult to distinguish from 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 were 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. For prostate cancer patients, we generated 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[58]. 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[59].

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 and smoking history as covariates. 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[33] 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 the 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 deconstructSigs[60]. For the analyses described here, 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 of 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) technical artifacts occurring at identical genomic coordinates and representing identical reference and alternate alleles recurring at >5% were filtered out to avoid possible technical artifacts unless (a) they had previously been reported as somatic in any of COSMIC (v86), Kandoth *et al.*[61] or Chang *et al.*[62] or (b) they were frameshifting indels or truncating SNVs

and 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 they were frameshifting indels or truncating SNVs and occurred in one of the 15 canonical CH genes; (4) variants occurring at any allele frequency in ExAC or gnomAD[63] 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 highest variant level recurrences occurred in *DNMT3A*, *TET2*, *PPM1D* and *TP53* at <5% recurrence (Extended Data Fig. 7) consistent with the joint-variant-calling of plasma cfDNA where the top mutated genes harboring WBC-matched variants were identical (Fig. 4d and Extended Data Fig. 6). The 15 canonical genes known to be associated with CH were *DNMT3A, TET2, ASXL1, PPM1D, TP53, JAK2, RUNX1, SF3B1, SRSF2, IDH1, IDH2, U2AF1, CBL, ATM* and *CHEK2*.

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-immortalized 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 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 FASTQs (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. 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 (Supplementary Methods, Figs. 1d–f, Extended Data Fig. 4, Supplementary Tables 8–10).

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

A modified version of MSIsensor[40] 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. 1). 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.

Statistical analyses

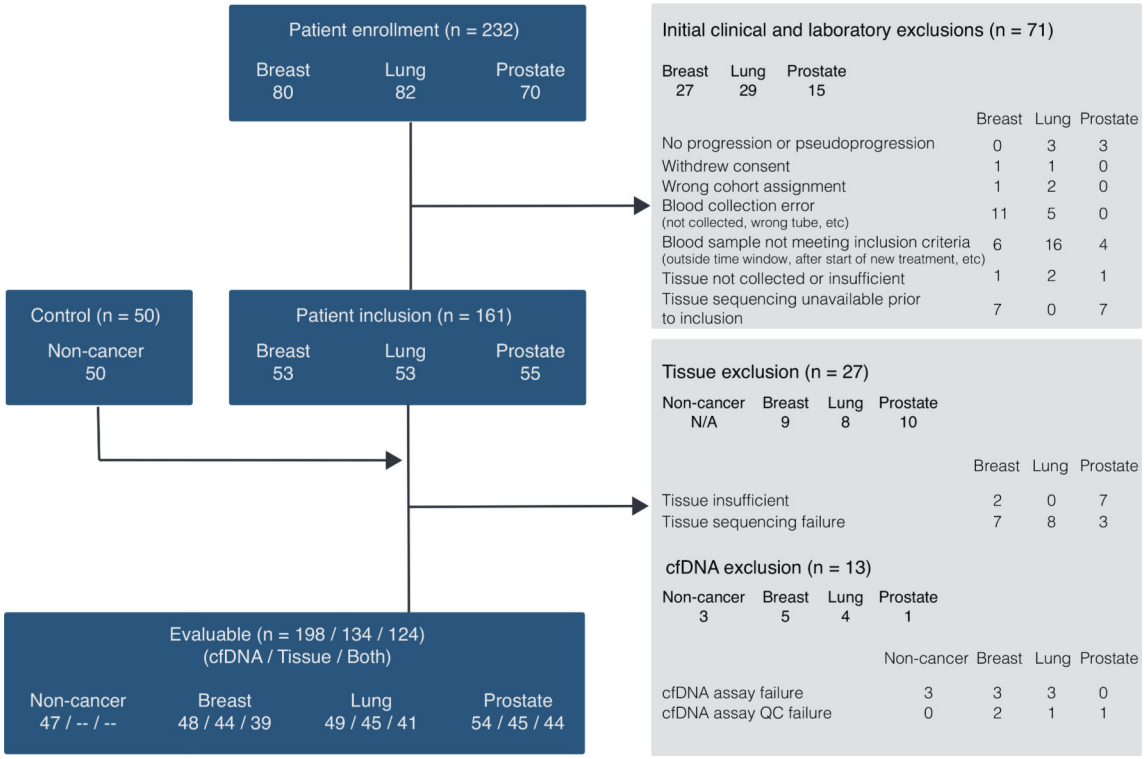
The association of variant allele fractions between technical replicates was measured using the R^2 obtained from a linear regression. The difference in detection rate between the three cohorts was assessed using two-by-two Fisher's exact tests, and the resulting p values were adjusted using the Bonferroni correction for multiple testing. The association between the tumor mutation cancer cell fraction and the cfDNA detection rate (overall and stratified by cancer type) was assessed using a χ^2 trend test (Cochran-Armitage test for trend). Exact 95% confidence intervals were calculated for the detection rate by cancer type. The difference in ctDNA fraction estimate between cancer types was assessed using a two-sided Kruskal-Wallis *H*-test. The increase in ctDNA fraction estimate with increasing disease burden was assessed using a one-sided Jonckheere-Terpstra test and stratified by cancer type. We used zero-inflated Poisson regressions to assess the association between the number of cfDNA variants per subject from each biological source (i.e. biopsy-matched, biopsy-subthreshold, WBC-matched and VUSo) and age with cancer status and smoking history as covariates where appropriate. The corresponding p values were obtained using two-sided Wald tests on the coefficients of the count model. Similarly, the association between the number of CH-derived variants per subject through direct analysis of WBC and age was assessed using a zero-inflated Poisson regression with cancer status as covariate and the corresponding p value was obtained using a two-sided Wald test on the coefficient of the count model. The association of CH measured in WBC in each of the 15 canonical CH genes and cancer status and prior history of radio- or chemotherapy was assessed using a

permutation-based likelihood ratio test from the coefficients of a logistic regression with age and smoking history as covariates where appropriate. Comparisons of the uncollapsed mean coverage between cfDNA and WBC were performed using two-sided paired Mann-Whitney *U*-tests stratified by cohort. Pairwise comparisons of the collapsed mean coverage of cfDNA or WBC between the different cancer cohorts and non-cancer controls were performed using two-sided Mann-Whitney *U*-tests and adjusted for multiple testing using the Bonferroni correction. The association of input DNA for library preparation and mean collapsed coverage of cfDNA samples was assessed using the R^2 obtained from a linear regression. The corresponding p value was obtained using a two-sided *t*-test. The pairwise differences of input DNA for library preparation between the different cancer cohorts and non-cancer controls were assessed using a two-sided Mann-Whitney *U*-tests and adjusted for multiple testing using the Bonferroni correction. Pairwise comparisons of the percentage of collapsed bases with SNVs or combined SNVs and indels of cfDNA or WBC between the different cancer cohorts and non-cancer controls were performed using two-sided Mann-Whitney *U*-tests and adjusted for multiple testing using the Bonferroni correction. The association of ctDNA fraction estimate or tumor purity and the Pearson's correlation coefficient measuring the association in Log_2 ratio space between patient-matched cfDNA and tumor biopsy samples was assessed using one-sided Jonckheere-Terpstra tests for increasing ctDNA fraction or tumor purity with increasing Pearson's correlation. Pairwise comparisons of AUCs for amplifications or homozygous deletions between the different cancer cohorts were carried out using two-sided DeLong tests. For each cfDNA variant category, a two-sided Mann-Whitney *U*-test was used to test whether the cancer cohort stratified by cancer type had a different mutation burden than the non-cancer controls. All statistical hypothesis tests were considered positive at α=0.05 and carried out in R/Bioconductor unless otherwise stated.

Data Availability

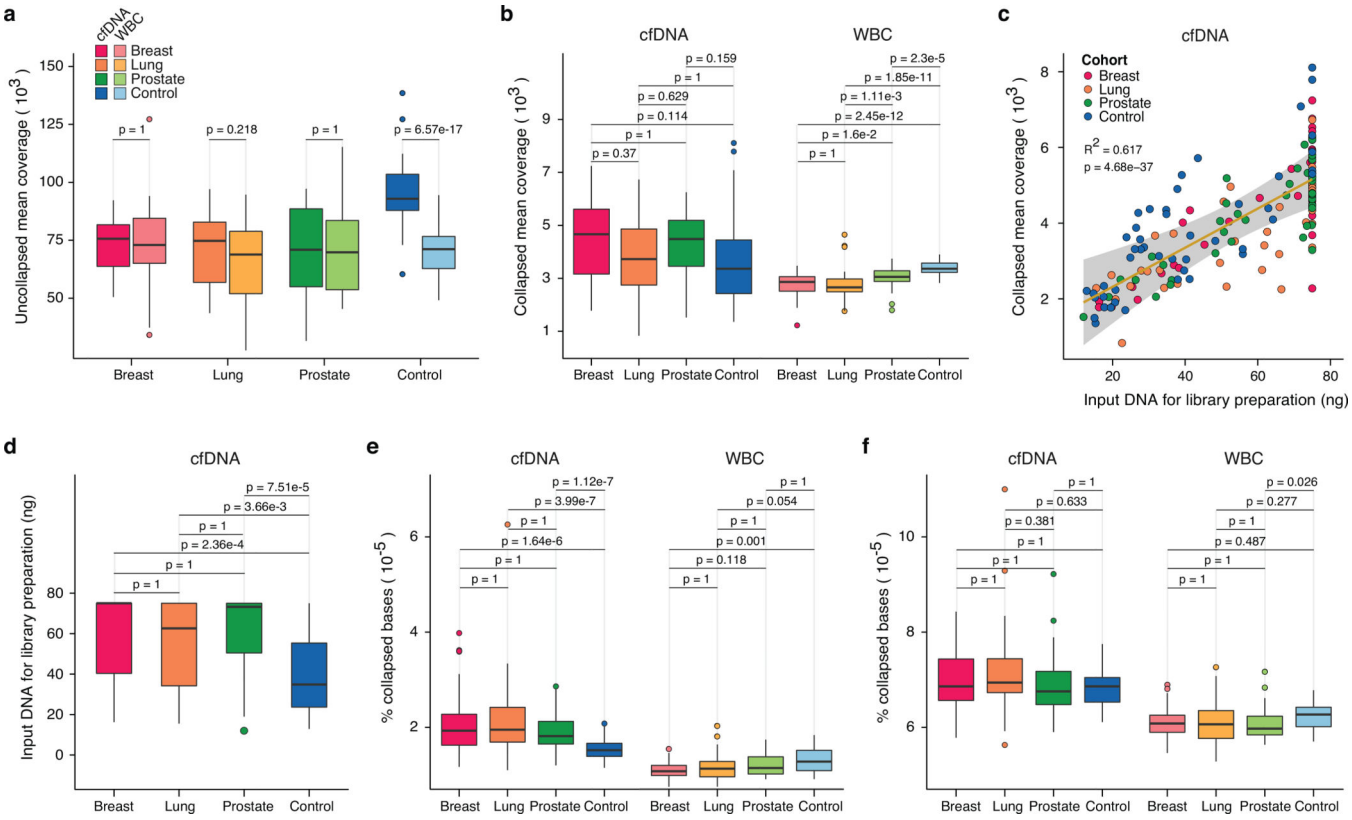
The assembled prospective somatic mutational data from cfDNA, WBC, and tumors for the entire cohort are provided as supplementary tables (Supplementary Tables 11–13) and the raw cfDNA and WBC sequencing data have been deposited in the European Genome-phenome Archive (EGA) under accession number EGAS00001003755. All code and scripts are available for academic use at https://github.com/ndbrown6/MSK-GRAIL-TECHVAL.

Extended Data



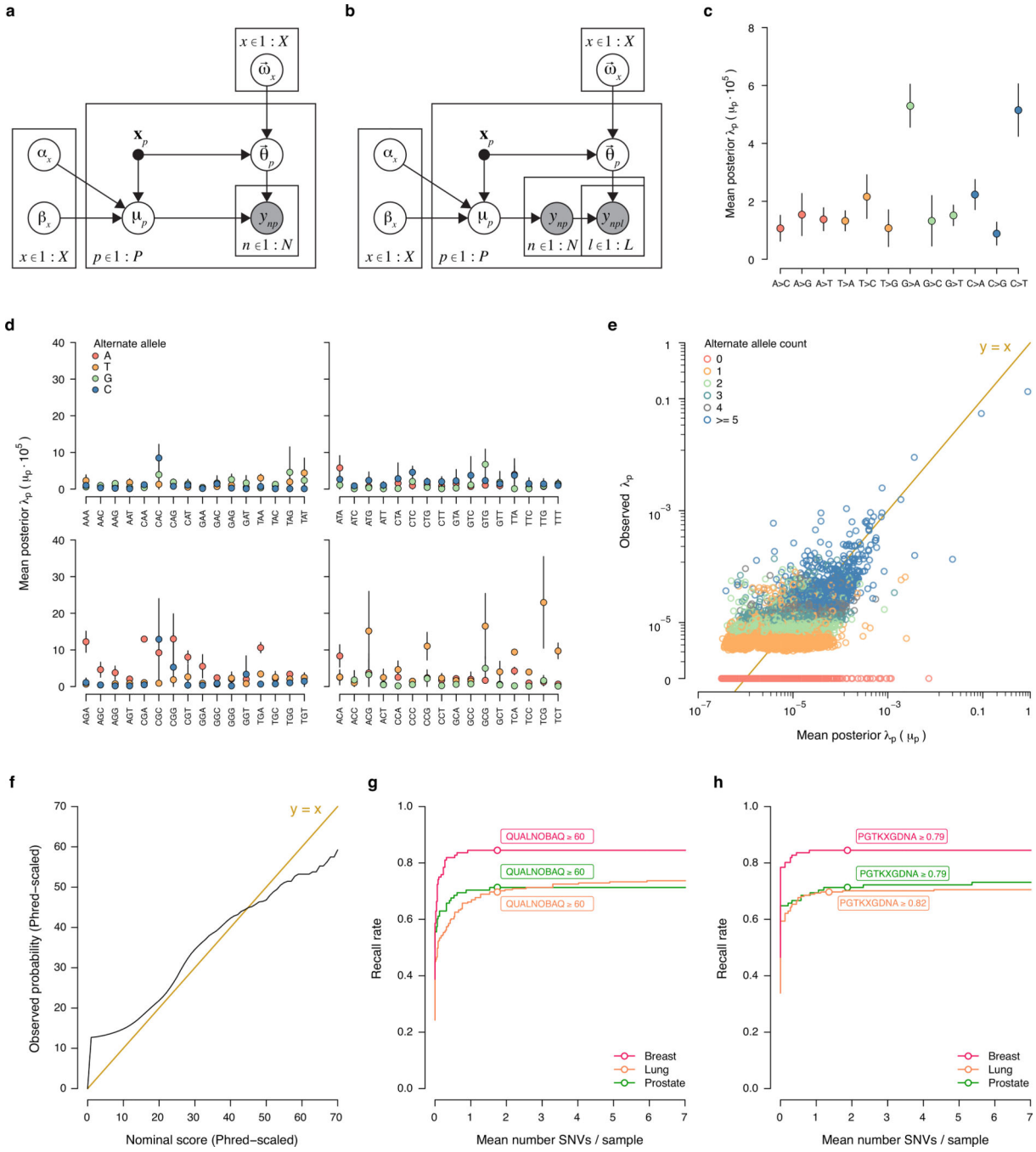
Extended Data Fig. 1. Study overview

Patient enrollment, inclusion and evaluable group are defined in the blue boxes. Detailed clinical, tissue and cfDNA exclusions are shown in the gray boxes.



Extended Data Fig. 2. Comparison of sequence depth and raw error rate distributions across cancer cohorts (n=124) and non-cancer controls (n=47)

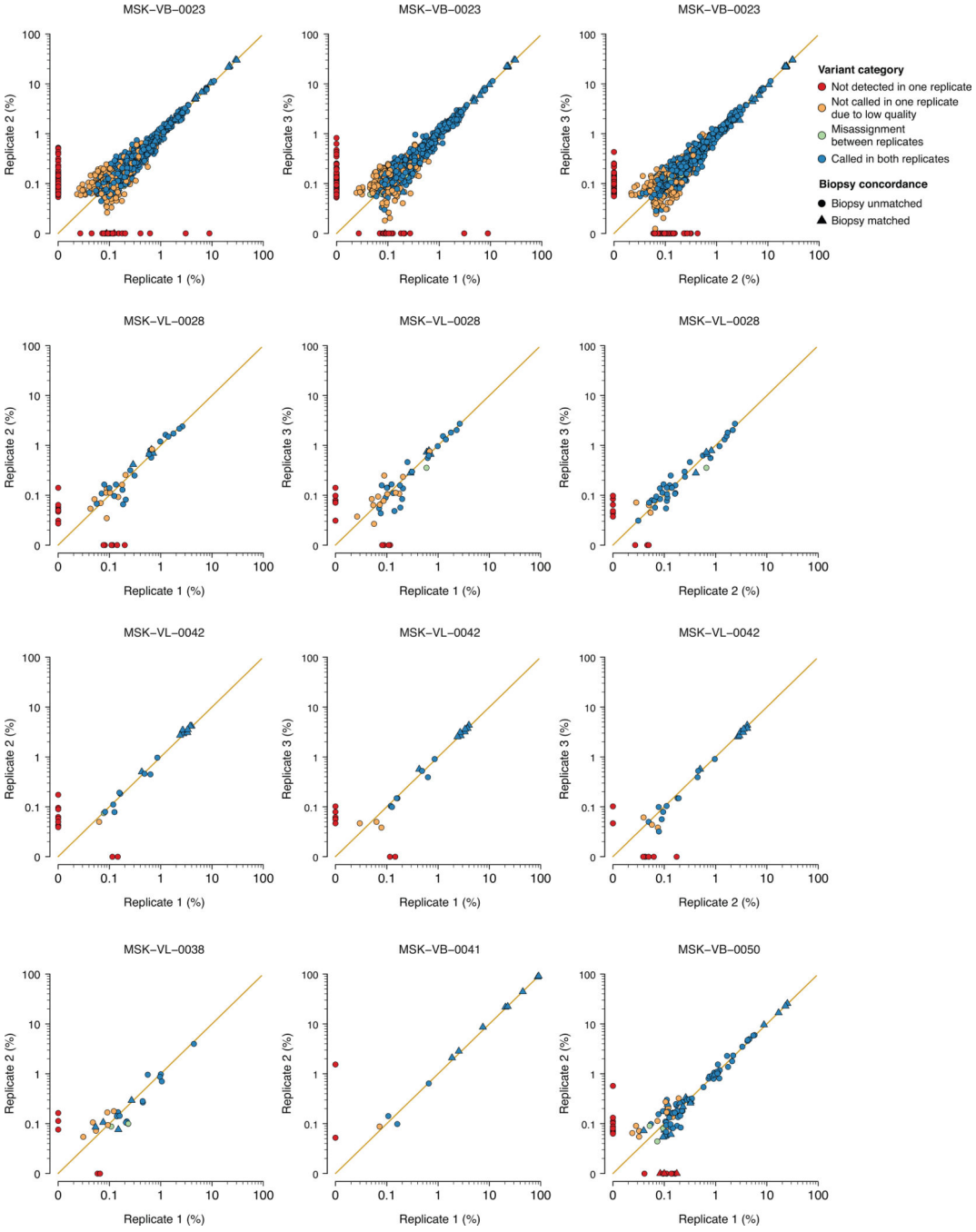
(**a**) Comparison of deduplicated and uncollapsed mean target sequence depth between cfDNA and WBC. The p values were obtained using paired two-sided Mann-Whitney *U*-tests comparing cfDNA against WBC. (**b**) Deduplicated and collapsed mean target sequence depth in cfDNA and WBC between the different cancer cohorts and non-cancer controls. (**c**) Association between the amount of cfDNA used for library preparation and the mean target deduplicated and collapsed sequencing depth. The diagonal line represents a linear regression with 99% confidence intervals. The p value was obtained using an *F*-test. (**d**) Distribution of mean target deduplicated and collapsed sequencing depth across the different cohorts. (**e,f**) Comparison of (**e**) raw substitution error rate and (**f**) raw substitution and indel error rate across the different cohorts. In (**b**) and (**d-f**), the p values were obtained from pairwise comparisons using two-sided Mann-Whitney *U*-tests and adjusted for multiple testing using the Bonferroni method. In (**e**), the substitution error rate represents the percentage of collapsed bases with non-reference base. Similarly, in (**f**) the combined error rate represents the percentage of collapsed bases with non-reference base or indels. In all panels, the cohorts consist of n=39 MBC, n=41 NSCLC and n=44 CRPC patients and n=47 non-cancer controls. In (**a-b**) and (**d-f**), the horizontal bars indicate the median and the boxes represent the interquartile range (IQR). The whiskers extend to 1.5 × IQR on either side.



Extended Data Fig. 3. Hierarchical Bayesian model for calibrated analysis of somatic cfDNA variants and performance assessment

(**a,b**) Plate models showing the hierarchy of statistical relationships for (**a**) single nucleotide variants and (**b**) small insertions and deletions influencing the observed quantity of alternate alleles y_{np} in each sample n at each position p conditional on both latent parameters μ (the rate of events), θ (the type of event), α, β as well as fixed covariates x_p (of X types) such as trinucleotide context and, separately, depth of sequencing at a position (d_p). Note that insertions and deletions have additional complexity as one must account for length of the

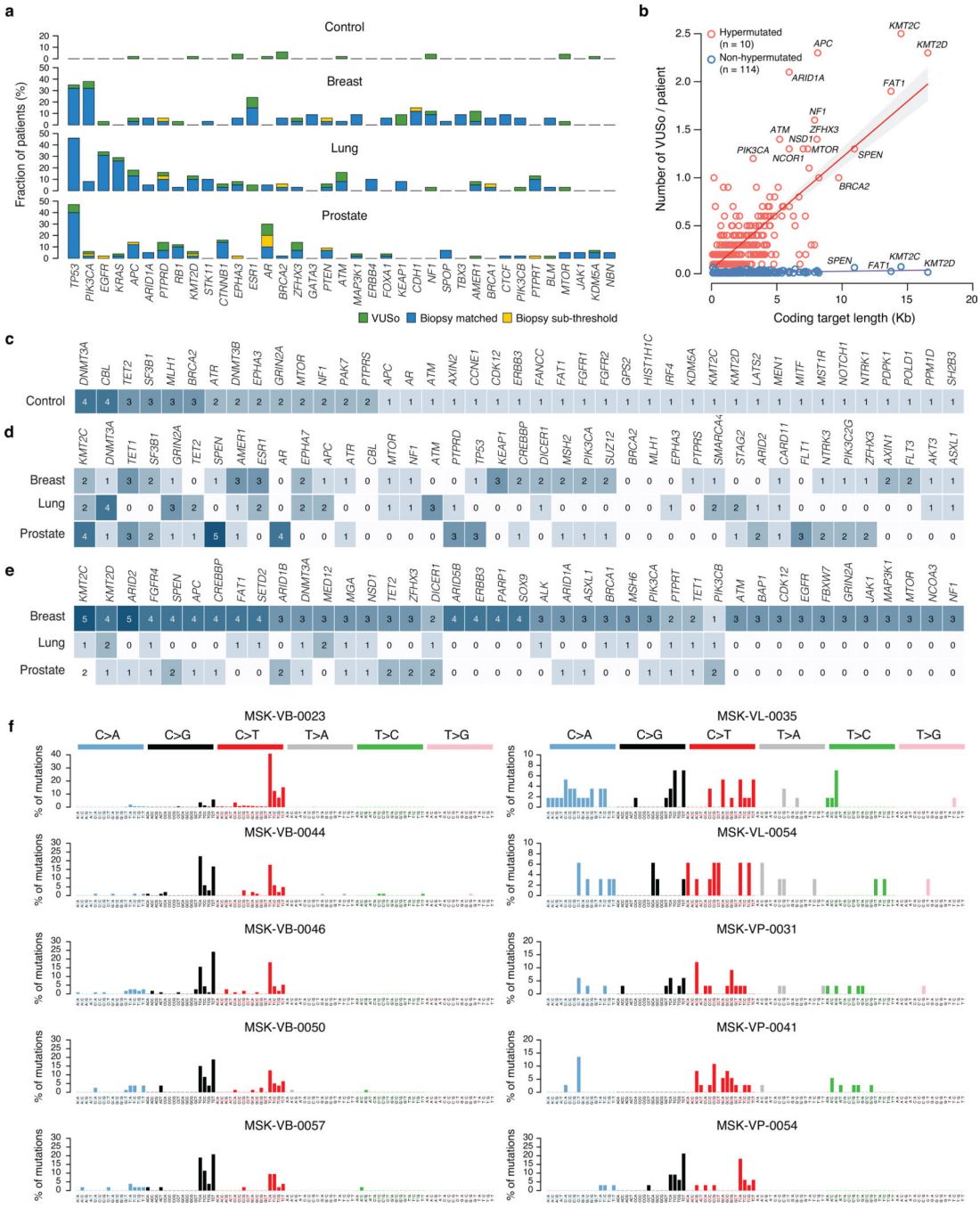
insertion/deletion event in the model as insertions and deletions of differing lengths have differing probabilities. The model was fitted to the training data consisting of n=43 unrelated non-cancer controls, estimates for the parameters were fixed and applied to new samples for scoring. (**c,d**) The posterior distributions of site-specific λ_p (μ_p d_p) were summarized by their mean μ_p and displayed for a subset of representative sites in (**c**) by type of mutation and (**d**) by trinucleotide context. In both panels, the midpoint indicates the mean and the vertical bars represent the 95% Gaussian confidence limits based on the *t*-distribution. (**e**) Estimated μ_p against the observed λ_p for samples in the training set. Note the data points at the bottom are all positions p with non-zero mean posterior μ_p and zero observed alternate allele counts. (**f**) Comparison of the estimated probability of observing an event (x-axis) with the actual empirical probability of observing such an event (y-axis). The plot was calibrated based on estimates of μ_p on chromosome 21. Note the initial sharp rise reflects the number of sites with zero observed alternate allele counts whilst the excess low probability events at the other end reflects the difficulty of stringently filtering out rare biological events such as clonal hematopoiesis. (**g**) Mean number of variants detected in healthy control individuals (x-axis) against the recall rate of biopsy-matched variants (y-axis) for the different cancer types. At Q_{60}, one can expect one false positive per million bases. Here, to exclude potentially CH derived variants, a fixed threshold of 0.8 on the posterior probability of detected variants originating from cfDNA (i.e. *PGTKXGDNA*) was adopted. (**h**) Mean number of variants detected in healthy control individuals (x-axis) against the recall rate of biopsy-matched variants (y-axis) at different probabilities for allowing variants to be assigned to cfDNA. The thresholds displayed were obtained by cross-validation holding out each cancer type and selecting a threshold which retains most of the biopsy-matched variants whilst still filtering out variants of potential hematopoietic origin. Here, to exclude variants potentially due to noise, a fixed threshold of Q_{60} was adopted.



Extended Data Fig. 4. Reproducibility of the high-intensity DNA assay

Six patient samples were selected for processing using two versions of the assay protocol (V1 and V2). These are labelled Replicate 1 and Replicate 2. A subset of three samples were further retested using version V2 and labelled Replicate 3. The panels illustrate the pairwise comparisons of measured VAF between all available replicates for each patient. In all panels, the variants are shape coded based on their origin, whether they were also detected in the matched tumor biopsy and color coded according to their category, whether they were

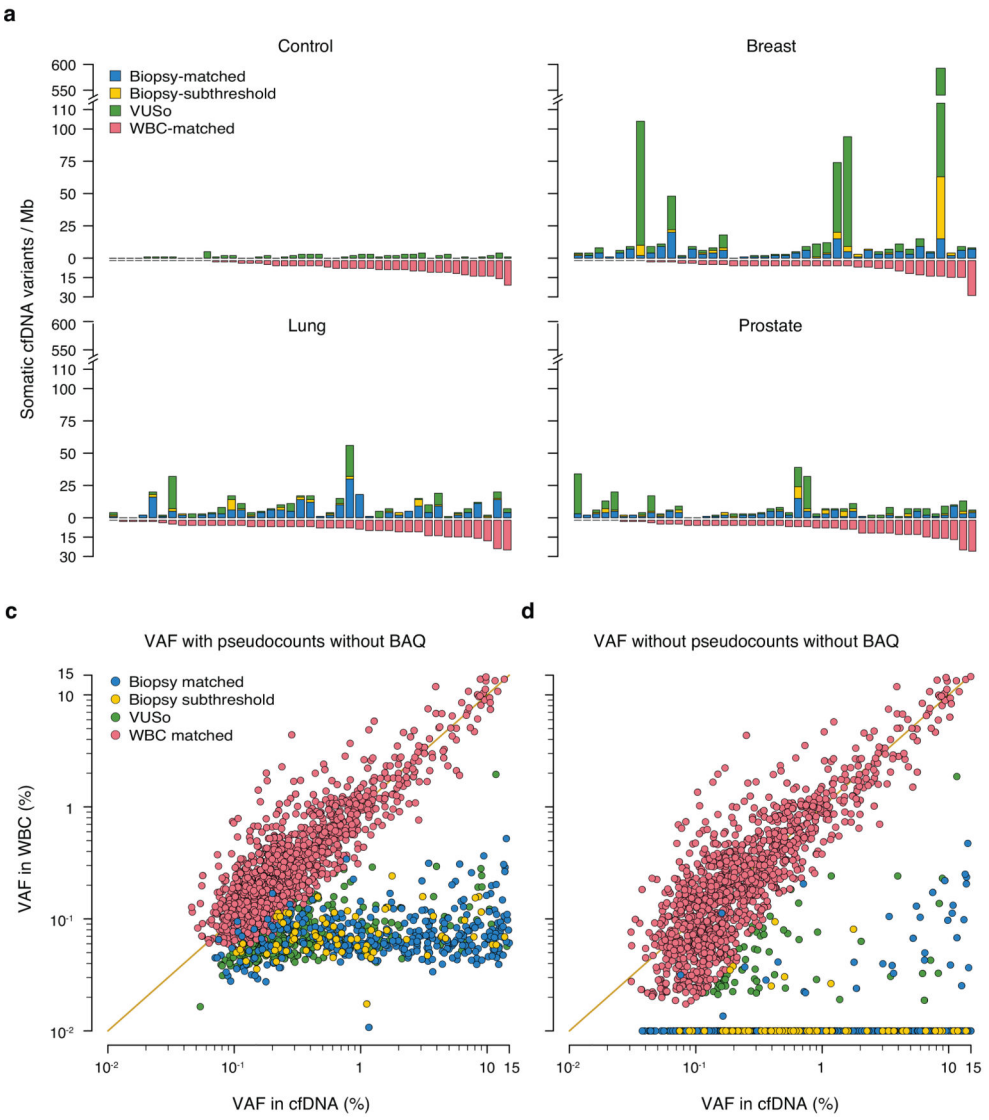
detected in both replicates and whether they were assigned to similar source categories (i.e. VUSo, WBC-matched or noise). In all panels, the samples are labelled on top.



Extended Data Fig. 5. Top mutated genes carrying VUSo and 96 base substitution profiles of ten hypermutated cfDNA samples

(**a**) Frequency of genomic alterations in cfDNA of 47 non-cancer controls and 124 cancer patients. 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). (**b**) Correlation of the number of VUSo per gene and per patient (y-axis) in the ten hypermutated and 114 non-hypermutated cancer patients

against the length of the coding region sequenced (x-axis) of each target gene. (**c**-**e**) Heat maps showing 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 (**c**) 47 non-cancer controls, (**d**) 114 non-hypermutated and (**e**) 10 hypermutated cancer patients. The numbers in the cells indicate the number of patients. (**f**) 96 base substitution profiles of the 10 hypermutated patients. 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.



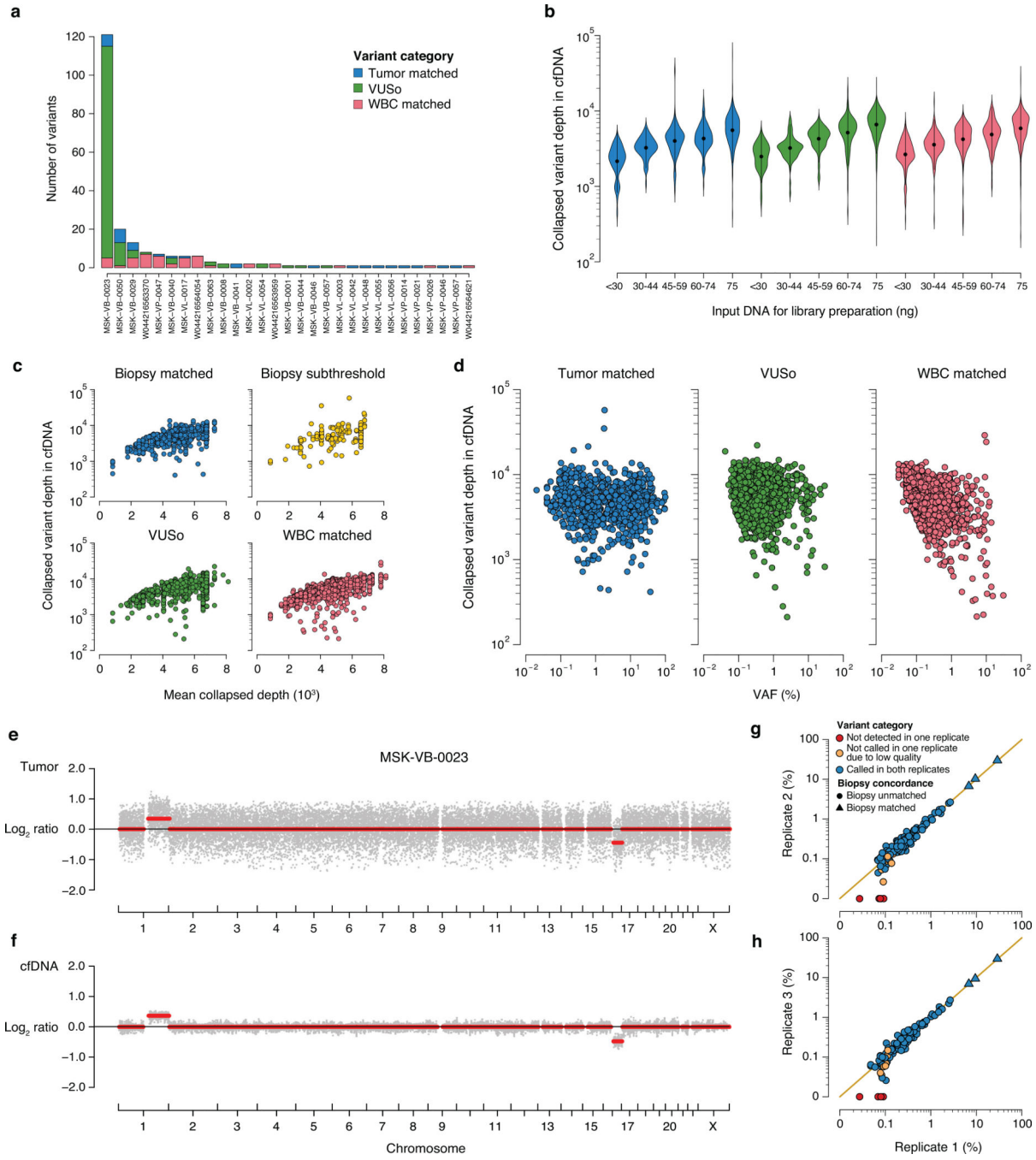
b

	Control	Breast	Lung	Prostate
DNMT3A	48	40	66	53
TET2	30	21	34	27
PPM1D	3	24	15	34
TP53	7	9	12	12
ASXL1	4	5	5	11
ATM	2	9	6	7
KMT2C	8	4	7	4
NF1	4	4	9	5
FAT1	2	9	4	6
CHEK2	1	1	5	10
KMT2D	4	4	3	5
PTPRT	3	3	6	4
CBL	2	5	5	2
SH2B3	3	4	4	1
ZFHX3	4	2	2	3
RAD21	4	0	1	5
MGA	2	3	4	0
GRIN2A	1	3	3	2
SF3B1	4	1	2	2
GNAS	3	1	2	3
EP300	4	1	3	0
AR	1	2	4	1
PTCH1	4	1	2	1
TET1	1	0	5	2
SETD2	4	0	2	2
ARID2	2	1	4	1
ARID1B	4	1	1	2
ROS1	1	5	0	1
IRS1	2	3	0	2
EGFR	2	2	2	1
NOTCH3	1	2	0	3
PGR	1	1	3	1
MED12	3	3	0	0
NCOR1	2	2	0	2
POLE	1	3	1	1
KDM5A	2	2	2	0
PTPRD	3	0	3	0
AMER1	2	0	3	1
DNMT3B	1	0	2	3
SPEN	1	0	3	2

Extended Data Fig. 6. Characterization of biological sources and composition of cfDNA variants

(**a**) 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. (**b**) 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. In (**a**,**b**), the cohorts consist of n=39 MBC, n=41 NSCLC and n=44 CRPC patients. Additionally, in (**a**) n=47, non-cancer controls are shown. (**c**,**d**) Distribution of Variant Allele Fractions (VAFs) of somatic mutations detected in cfDNA and WBC using the high-intensity sequencing assay where variants are color coded according to source of origin. Somatic variants are displayed for n=114 non-hypermutated

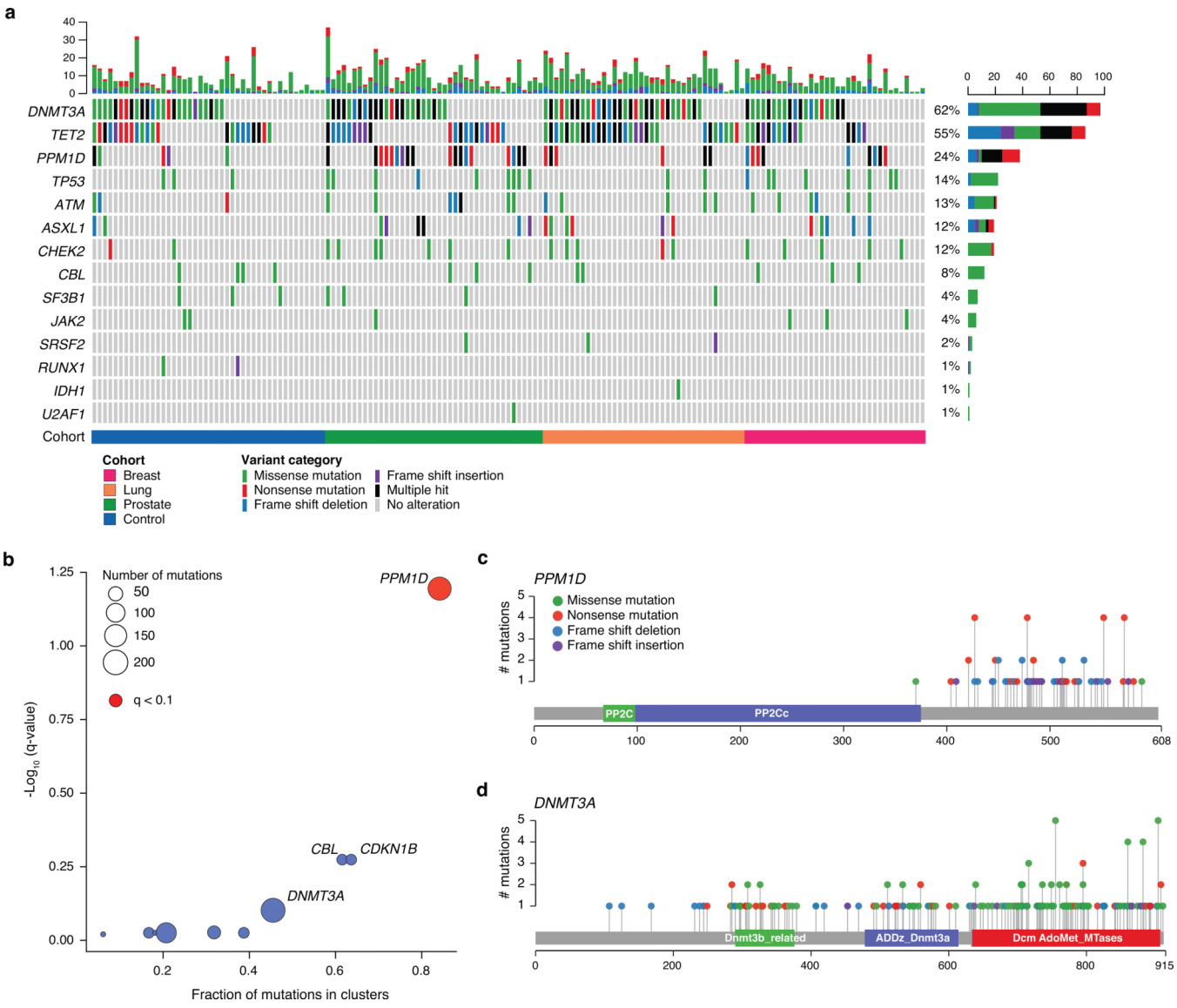
cancer patients and n=47 non-cancer controls. The allelic (AD) and total (DP) depths are obtained from raw pileups without base alignment quality filtering (BAQ). In (**c**), the VAF is smoothed with added pseudocounts to AD and DP such that $AD' = AD + 2$ and $DP' = DP + 4$. In (**d**), variants detected with zero AD in WBC were displayed as 0.01% VAF in WBC due to the logarithmic scaled axes.



Extended Data Fig. 7. Somatic mutations occurring at high sequencing depth in cfDNA

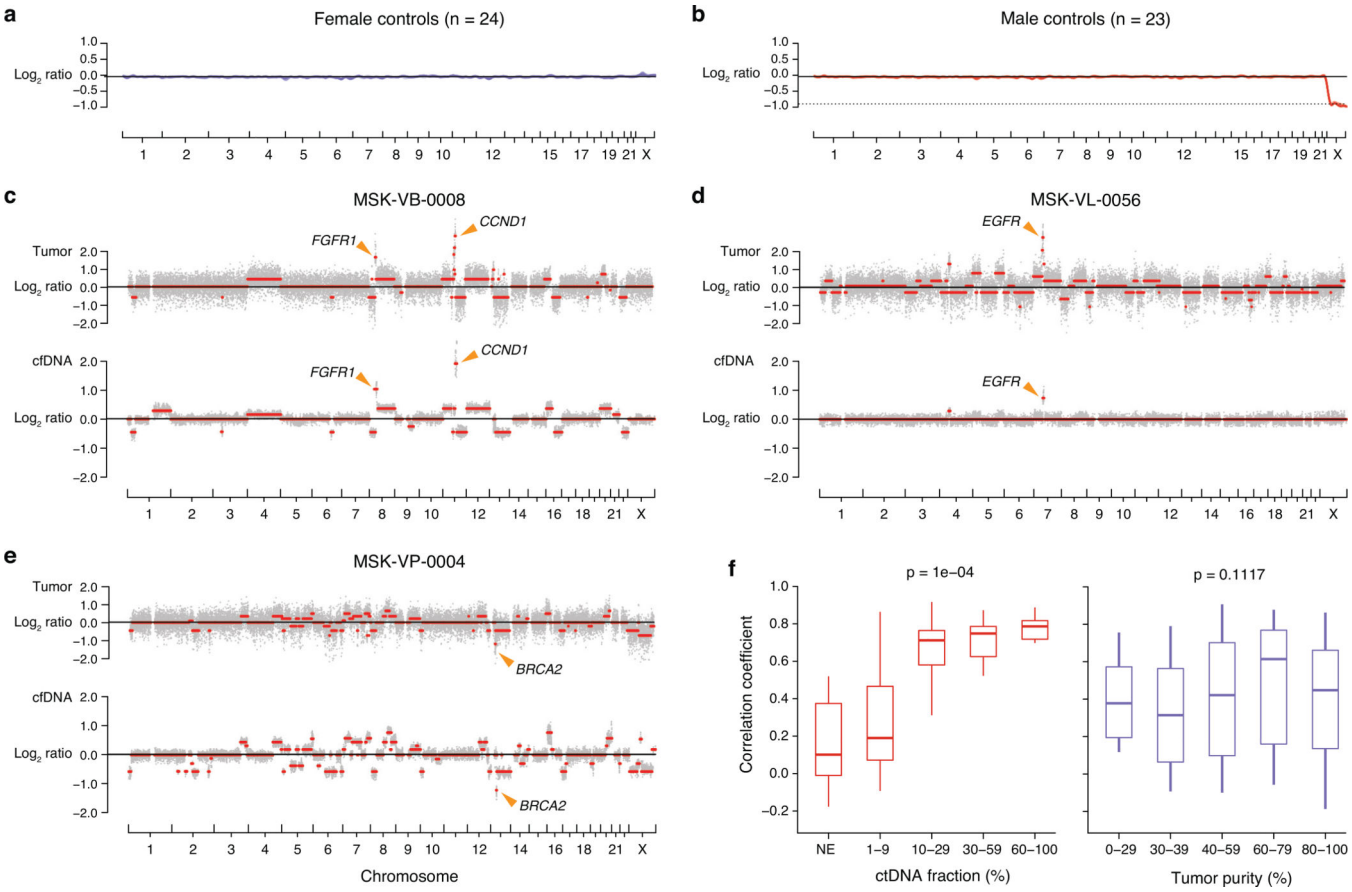
Somatic mutations detected at sequencing depth >10,000 in cfDNA occur mostly in hypermutated samples and are related to sample level mean target collapsed depth which is itself a function of the amount of input DNA used for library preparation. (**a**) Number of somatic mutations occurring at >10,000 sequence depth (n=215) per patient and categorized into WBC-matched, VUSo or Tumor-matched where the latter category is composed of Biopsy-matched and Biopsy-subthreshold mutations. (**b**) Variant level collapsed depth for all somatic mutations detected in cfDNA categorized into Tumor-matched, VUSo or WBC-

matched and grouped according to the amount of input DNA used for library preparation. (**c**) Variant level collapsed depth for all somatic mutations detected in cfDNA against sample level mean collapsed target depth. (**d**) variant level collapsed depth for all somatic mutations against modeled VAF in cfDNA. 121 of 215 (56.3%) somatic mutations detected at sequencing depth >10,000 in cfDNA occurred in the hypermutated patient MSK-VB-0023. (**e,f**) Log_2 ratios of (**e**) tumor biopsy and (**f**) cfDNA of patient MSK-VB-0023. The tumor biopsy and cfDNA showed similar copy number alterations (i.e. 1q+ and 16q-). No high-level copy number amplifications were observed in either the tumor biopsy or the cfDNA which could explain the high sequencing depth. Three replicate sequencing of cfDNA and WBC were available for that patient. (**g**) and (**h**) Pairwise comparisons of VAF for the 121 mutations detected at depth >10,000 using version V1 of the assay. In (**a**), '1' denotes hypermutated samples. In (**b**), the midpoint indicates the median whilst the violins extent to the full range of the data. In (**b**-**d**), the sequencing depths of somatic variants for the cohort of n=124 cancer patients are shown. In (**e**) and (**f**), the grey points represent the raw Log_2 ratios and are ordered according to their genomic coordinates. The solid red lines indicate the segmented values. In (**g**) and (**h**), the variants are shape coded based on their origin (i.e. whether they were also detected in the matched tumor biopsy and color coded according to their category; whether they were called in both replicates and assigned to similar source categories, namely VUSo, WBC-matched or noise).



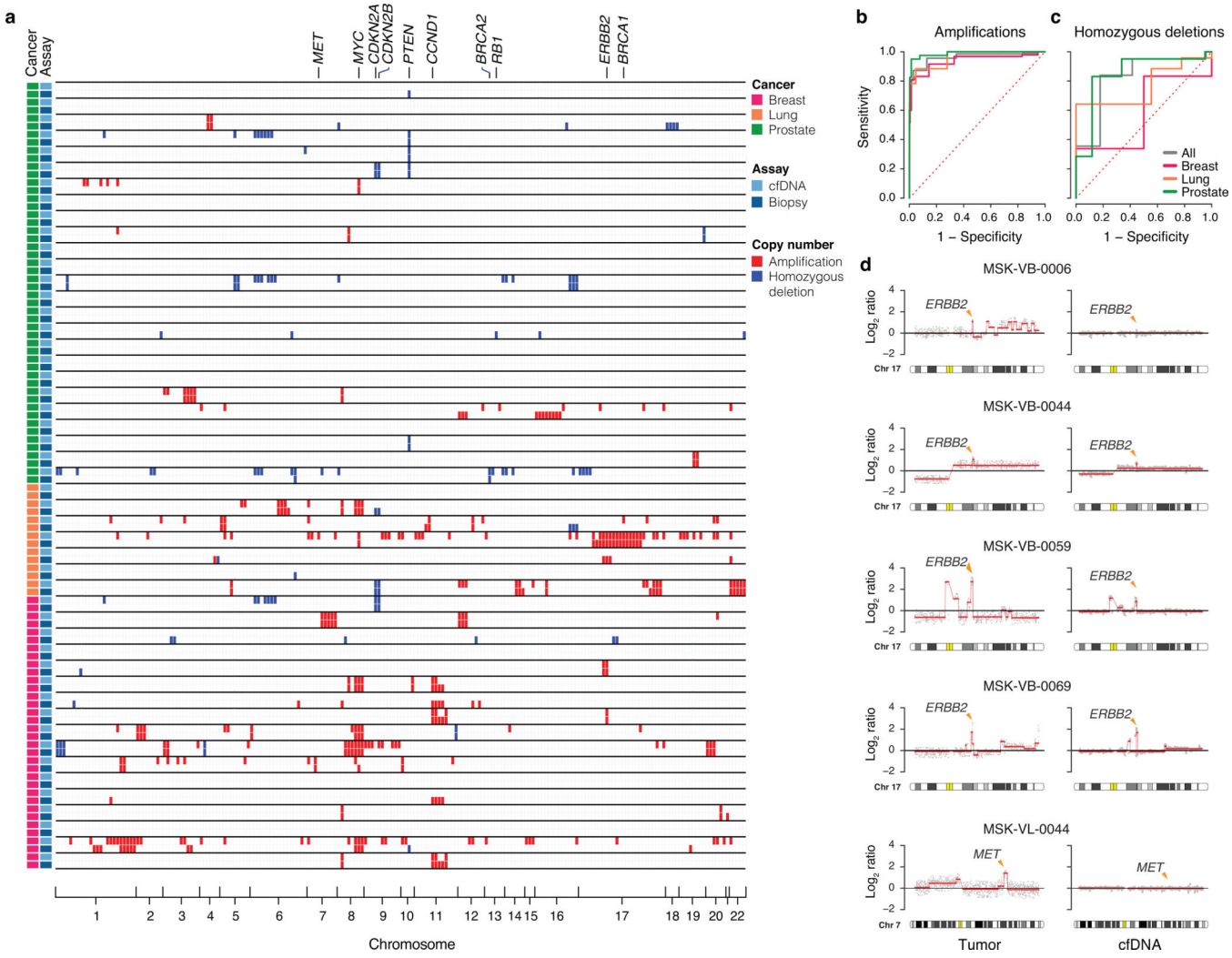
Extended Data Fig. 8. Characterization of CH derived variants through direct analysis of WBC

(**a**) CH-related somatic mutations in the top 14 mutated genes across the 114 non-hypermutated cancer patients and 47 non-cancer controls together with the marginal frequencies by patient (top) and by gene (right). *DNMT3A*, *TET2* and *PPM1D* are the top mutated genes in WBC and harbor multiple hits (i.e. two or more mutations per patient). (**b**) Clustering within genes of CH-derived mutations detected in WBC. The clusters and associated p values were computed using a modification of OncodriveCLUST[55] which assumes the number of mutations in clusters follows a Poisson distribution. The resulting p values are two-sided. (**c,d**) Distribution of mutations in *PPM1D* (**c**) according to genomic coordinates and for *DNMT3A* (**d**). Mutations detected in *PPM1D* are clustered in the C-terminus of the protein.



Extended Data Fig. 9. Copy number profile derived from cfDNA of non-cancer controls and cancer patients

(**a-b**) Log_2 ratios estimated from the cfDNA of (**a**) n=24 female and (**b**) n=23 male control individuals. For each individual, the raw Log_2 ratios were smoothed using a cubic spline. The two panels show the superimposed splines for all control samples according to gender. (**c-e**) Log_2 ratios of tumor biopsies (top panels) and their corresponding matched cfDNA (bottom panel) for three cases (**c**) MSK-VB-0008, (**d**) MSK-VL-0056 and (**e**) MSK-VP-0004 where amplification of *CCND1, FGFR1*, *EGFR* and a homozygous deletion of *BRCA2* were reported, respectively. The arrows point to the reported amplifications or deletions. The segmented Log_2 ratios were used to compute the Pearson correlation coefficient comparing segments overlapping >75% in the tumor biopsies and cfDNA samples. In (**a-e**), the Log_2 ratios are displayed according to their genomic coordinates. In (**c-e**), the grey dots show the raw estimates while the red lines represent the segmented values. (**f**) The association of the Pearson's *r* against the ctDNA fraction and purity of the tumor biopsies. The cohort consists of n=124 cancer patients with paired tumor biopsy and cfDNA samples. The p values were obtained using a permutation based one-sided Jonckheere-Terpstra test for increasing Pearson's *r* with ctDNA fraction or tumor purity. The horizontal bars indicate the median and the boxes represent the interquartile range (IQR). The whiskers extend to 1.5 × IQR on either side. NE; not evaluable.



Extended Data Fig. 10. Comparison of copy number alterations in tumor biopsies and matched cfDNA samples

(**a**) Heatmap of all genes where an amplification or a homozygous deletion was found in either the tumor biopsy or cfDNA. The samples are interleaved (i.e. tumor biopsy and cfDNA) and represented along the rows, whilst genes are ordered in columns relative to their genomic coordinates. (**b,c**) Receiver operating characteristic curves comparing (**b**) copy number amplifications and (**c**) homozygous deletions detected in the tumor biopsies with the absolute copy numbers inferred in cfDNA. Each tumor-cfDNA sample pair was used to construct individual curves. These were averaged after fitting a local polynomial regression and estimating the sensitivities over fixed intervals of specificities. In (**a**-**c**), only tumor-cfDNA sample pairs from n=49 patients with ctDNA fraction >10% were used. (**d**) Four MBC patients: MSK-VB-0006, MSK-VB-0044, MSK-VB-0059 and MSK-VB-0069 with a reported amplification of *ERBB2* on chromosome 17q are shown together with one NSCLC patient, MSK-VL-0044 with a reported *MET* amplification on chromosome 7q. The tumor biopsies are displayed on the left and the matched cfDNA are shown on the right together with the corresponding chromosome ideogram. The genomic coordinates of *ERBB2* and *MET* are displayed by orange arrows and labelled accordingly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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, S. Wen, D. Reales, J. Galle, R. Cambria and members of the MSK Office of Clinical Research. This work was supported by GRAIL, Inc. and National Institutes of Health awards P30 CA008748 (MSKCC), R01 CA234361 (D.B.S.), Breast Cancer Alliance Young Investigator Award (P.R.), and the Breast Cancer Research Foundation (J.S.R-F), Congressionally Directed Medical Research Programs W81XWH-15-1-0547 (J.S.R-F) and GC229671 (J.S.R-F).

Competing Interests

P.R. reports consulting/advisory board for Novartis and institutional research support from Illumina and GRAIL, Inc. B.T.L. 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. W.A. reports consulting or advisory role from Clovis Oncology, Janssen, ORIC phama 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. J.M.I. holds equity in LumaCyte, LLC and has received institutional research support from GRAIL, Inc. and Guardant Health. G.P. is on the Scientific Advisory Board Member for Tizona Therapeutics and has consulted for Merck, Bristol-Myers Squibb, Kyowa Hakko Kirin Pharma. D.M.H. reports stock and other ownership interests in Fount and consulting or advisory role for Chugai Pharma, Boehringer Ingelheim, AstraZeneca, Pfizer, Bayer, Genentech, and Fount. He has received research funding from AstraZeneca, Puma Biotechnology, Loxo, and Bayer and travel accommodation expenses from Genentech and Chugai Pharma. G.J.R. received consulting for Genentech/Roche in 2016 and received institutional research support for clinical research from Pfizer, Roche/Genentech, and Takeda. C.M.R. 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. L.A.D. is a member of the board of directors of Personal Genome Diagnostics (PGDx) and Jounce Therapeutics. He is a paid consultant to PGDx and Neophore. He is an uncompensated consultant for Merck but has received research support for clinical trials from Merck. He is an inventor of multiple licensed patents related to technology for circulating tumor DNA analyses and mismatch repair deficiency for diagnosis and therapy from Johns Hopkins University. Some of these licenses and relationships are associated with equity or royalty payments directly to L.A.D. and Johns Hopkins. He holds equity in PGDx, Jounce Therapeutics, Thrive Earlier Detection and Neophore. His spouse holds equity in Amgen. The terms of all these arrangements are being managed by Johns Hopkins and Memorial Sloan Kettering in accordance with their conflict of interest policies. D.B.S. received honoraria/consulted for Pfizer, Loxo Oncology, Illumina, Intezyne and Vivideon Therapuetics. J.S.R-F reports personal/consultancy fees from VolitionRx, Page.AI, Goldman Sachs, REPARE Therapeutics, GRAIL, Inc, Ventana Medical Systems, Roche, Genentech and Invicro, outside the scope of the submitted work. B.J., E.H., C.H., O.V., T.M., S.G., R.V.S., Q.L., L.S., N.E., J.Y., A.W.B., M.L. A.S., H.X., M.P.H., W.F.N., A.M.A. are or were GRAIL, Inc. employees and hold stock and/or other ownership interests in GRAIL, Inc.; A.W.B. additionally reports Foresite ownership interest. The other coauthors report no competing interests.

References

1. Stroun M, Anker P, Lyautey J, Lederrey C & Maurice PA Isolation and characterization of DNA from the plasma of cancer patients. Eur J Cancer Clin Oncol 23, 707–712 (1987). [PubMed: 3653190]
2. Leon SA, Shapiro B, Sklaroff DM & Yaros MJ Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res 37, 646–650 (1977). [PubMed: 837366]
3. Jr LAD & Bardelli A Liquid Biopsies: Genotyping Circulating Tumor DNA. Journal of Clinical Oncology 32, 579–586 (2014). [PubMed: 24449238]
4. Wan JCM, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer 17, 223–238 (2017). [PubMed: 28233803]

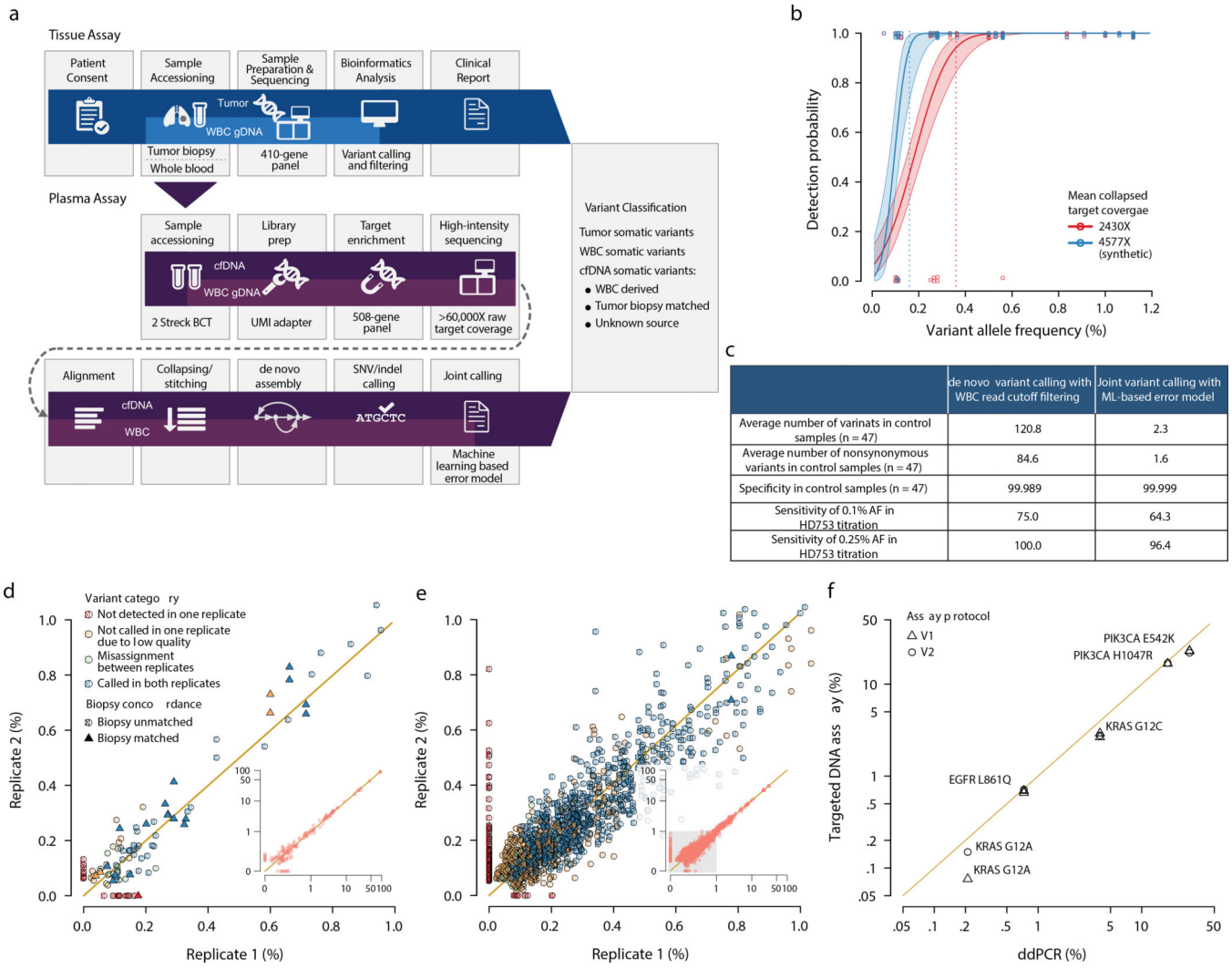
5. Lanman RB, et al. Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA. PLoS One 10, e0140712 (2015). [PubMed: 26474073]
6. Adalsteinsson VA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun 8, 1324 (2017). [PubMed: 29109393]
7. Aravanis AM, Lee M & Klausner RD Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. Cell 168, 571–574 (2017). [PubMed: 28187279]
8. 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). [PubMed: 28669404]
9. Jamal-Hanjani M, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. N Engl J Med 376, 2109–2121 (2017). [PubMed: 28445112]
10. Jaiswal S, et al. Age-related clonal hematopoiesis associated with adverse outcomes. N Engl J Med 371, 2488–2498 (2014). [PubMed: 25426837]
11. 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). [PubMed: 19861545]
12. Murtaza M, et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature 497, 108–112 (2013). [PubMed: 23563269]
13. Rothwell DG, et al. Utility of ctDNA to support patient selection for early phase clinical trials: the TARGET study. Nat Med 25, 738–743 (2019). [PubMed: 31011204]
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). [PubMed: 29463554]
15. Parikh AR, et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. Nat Med 25, 1415–1421 (2019). [PubMed: 31501609]
16. Risques RA & Kennedy SR Aging and the rise of somatic cancer-associated mutations in normal tissues. PLoS Genet 14, e1007108 (2018). [PubMed: 29300727]
17. Steensma DP, et al. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. Blood 126, 9–16 (2015). [PubMed: 25931582]
18. Bowman RL, Busque L & Levine RL Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. Cell Stem Cell 22, 157–170 (2018). [PubMed: 29395053]
19. Busque L, Buscarlet M, Mollica L & Levine RL Concise Review: Age-Related Clonal Hematopoiesis: Stem Cells Tempting the Devil. Stem Cells 36, 1287–1294 (2018). [PubMed: 29883022]
20. Coombs CC, 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). [PubMed: 28803919]
21. Zink F, et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood 130, 742–752 (2017). [PubMed: 28483762]
22. Jaiswal S, et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. N Engl J Med 377, 111–121 (2017). [PubMed: 28636844]
23. Xie M, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med 20, 1472–1478 (2014). [PubMed: 25326804]
24. Genovese G, et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N Engl J Med 371, 2477–2487 (2014). [PubMed: 25426838]
25. Phallen J, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med 9(2017).
26. Gillis NK, 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). [PubMed: 27927582]
27. Liu J, et al. Biological background of the genomic variations of cf-DNA in healthy individuals. Ann Oncol (2018).
28. Hu Y, et al. False-Positive Plasma Genotyping Due to Clonal Hematopoiesis. Clin Cancer Res 24, 4437–4443 (2018). [PubMed: 29567812]

29. 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). [PubMed: 28536309]
30. Thompson JC, 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). [PubMed: 27601595]
31. 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). [PubMed: 29325035]
32. Sacher AG, 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). [PubMed: 27055085]
33. Zehir A, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 23, 703–713 (2017). [PubMed: 28481359]
34. Cheng DT, 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). [PubMed: 25801821]
35. Newman AM, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 34, 547–555 (2016). [PubMed: 27018799]
36. Kinde I, Wu J, Papadopoulos N, Kinzler KW & Vogelstein B Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108, 9530–9535 (2011). [PubMed: 21586637]
37. Razavi P, et al. The Genomic Landscape of Endocrine-Resistant Advanced Breast Cancers. Cancer Cell 34, 427–438 e426 (2018). [PubMed: 30205045]
38. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013). [PubMed: 23945592]
39. Nik-Zainal S, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–54 (2016). [PubMed: 27135926]
40. Niu B, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics 30, 1015–1016 (2014). [PubMed: 24371154]
41. Polak P, et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nat Genet 49, 1476–1486 (2017). [PubMed: 28825726]
42. Gerhauser C, et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. Cancer Cell 34, 996–1011 e1018 (2018). [PubMed: 30537516]
43. de Bruin EC, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science 346, 251–256 (2014). [PubMed: 25301630]
44. Le DT, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med 372, 2509–2520 (2015). [PubMed: 26028255]
45. Merker JD, 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). [PubMed: 29504834]
46. Cohen JD, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 359, 926–930 (2018). [PubMed: 29348365]
47. Schultheis AM, et al. Massively Parallel Sequencing-Based Clonality Analysis of Synchronous Endometrioid Endometrial and Ovarian Carcinomas. J Natl Cancer Inst 108, djv427 (2016). [PubMed: 26832770]
48. Hsu JI, et al. PPM1D Mutations Drive Clonal Hematopoiesis in Response to Cytotoxic Chemotherapy. Cell Stem Cell 23, 700–713 e706 (2018). [PubMed: 30388424]
49. Bettegowda C, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 6, 224ra224 (2014).
50. Dawson SJ, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med 368, 1199–1209 (2013). [PubMed: 23484797]

51. Chabon JJ, et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. Nat Commun 7, 11815 (2016). [PubMed: 27283993]
52. Young AL, Challen GA, Birmann BM & Druley TE Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat Commun 7, 12484 (2016). [PubMed: 27546487]
53. 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. Journal of Clinical Oncology 36(2018).
54. 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).

Methods-Only References

55. Tamborero D, Gonzalez-Perez A & Lopez-Bigas N OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29, 2238–2244 (2013). [PubMed: 23884480]
56. Shen R & Seshan VE FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res 44, e131 (2016). [PubMed: 27270079]
57. Carter SL, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 30, 413–421 (2012). [PubMed: 22544022]
58. Ulmert D, et al. A novel automated platform for quantifying the extent of skeletal tumour involvement in prostate cancer patients using the Bone Scan Index. Eur Urol 62, 78–84 (2012). [PubMed: 22306323]
59. Armstrong AJ, et al. Phase 3 Assessment of the Automated Bone Scan Index as a Prognostic Imaging Biomarker of Overall Survival in Men With Metastatic Castration-Resistant Prostate Cancer: A Secondary Analysis of a Randomized Clinical Trial. JAMA Oncol 4, 944–951 (2018). [PubMed: 29799999]
60. Rosenthal R, McGranahan N, Herrero J, Taylor BS & Swanton C DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 17, 31 (2016). [PubMed: 26899170]
61. Kandoth C, et al. Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339 (2013). [PubMed: 24132290]
62. Chang MT, et al. Accelerating Discovery of Functional Mutant Alleles in Cancer. Cancer Discov 8, 174–183 (2018). [PubMed: 29247016]
63. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016). [PubMed: 27535533]
64. Chakravarty D, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol 2017(2017).



	de novo variant calling with WBC read cutoff filtering	Joint variant calling with ML-based error model
Average number of varinats in control samples (n = 47)	120.8	2.3
Average number of nonsynonymous variants in control samples (n = 47)	84.6	1.6
Specificity in control samples (n = 47)	99.989	99.999
Sensitivity of 0.1% AF in HD753 titration	75.0	64.3
Sensitivity of 0.25% AF in HD753 titration	100.0	96.4

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**) Analytical performance of the targeted DNA assay. The detection probability is shown as a function of increasing variant allele fraction in HD753 cell line DNA titrations. The curves correspond to the mean target coverage of 2,430X from 30 ng cell line DNA input and the mean target coverage of 4,577X obtained from simulated FASTQs. (**c**) Estimated variant calling specificity using non-cancer control samples and corresponding variant calling sensitivity using methods as described in the supplementary methods (joint variant analysis using the machine learning error model). Non-cancer controls were not used to train the model here. (**d**) Comparison of allele fraction of variants

detected using either of the two targeted DNA assay protocols in five patients. One MBC hypermutated patient, shown in (**e**), 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. (**f**) 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 sample was employed for the targeted DNA assays using two versions of the protocol (V1 and V2). One sample lacking canonical hotspot mutation in the ddPCR measurements was excluded.

a

	Patients for whom ≥ 1 of the variants detected in tumor was also detected in plasma		SNV / Indels detection	
	N patients cfDNA / tissue	Detection rate (95% CI)	N variants cfDNA / tissue	PPA (95% CI)
All	104 / 124	84% (76 - 90)	530 / 740	72% (68 - 75)
Breast	37 / 39	95% (83 - 99)	184 / 251	73% (67 - 79)
Lung	31 / 41	76% (60 - 88)	227 / 321	71% (65 - 76)
Prostate	36 / 44	82% (67 - 92)	119 / 168	71% (63 - 78)

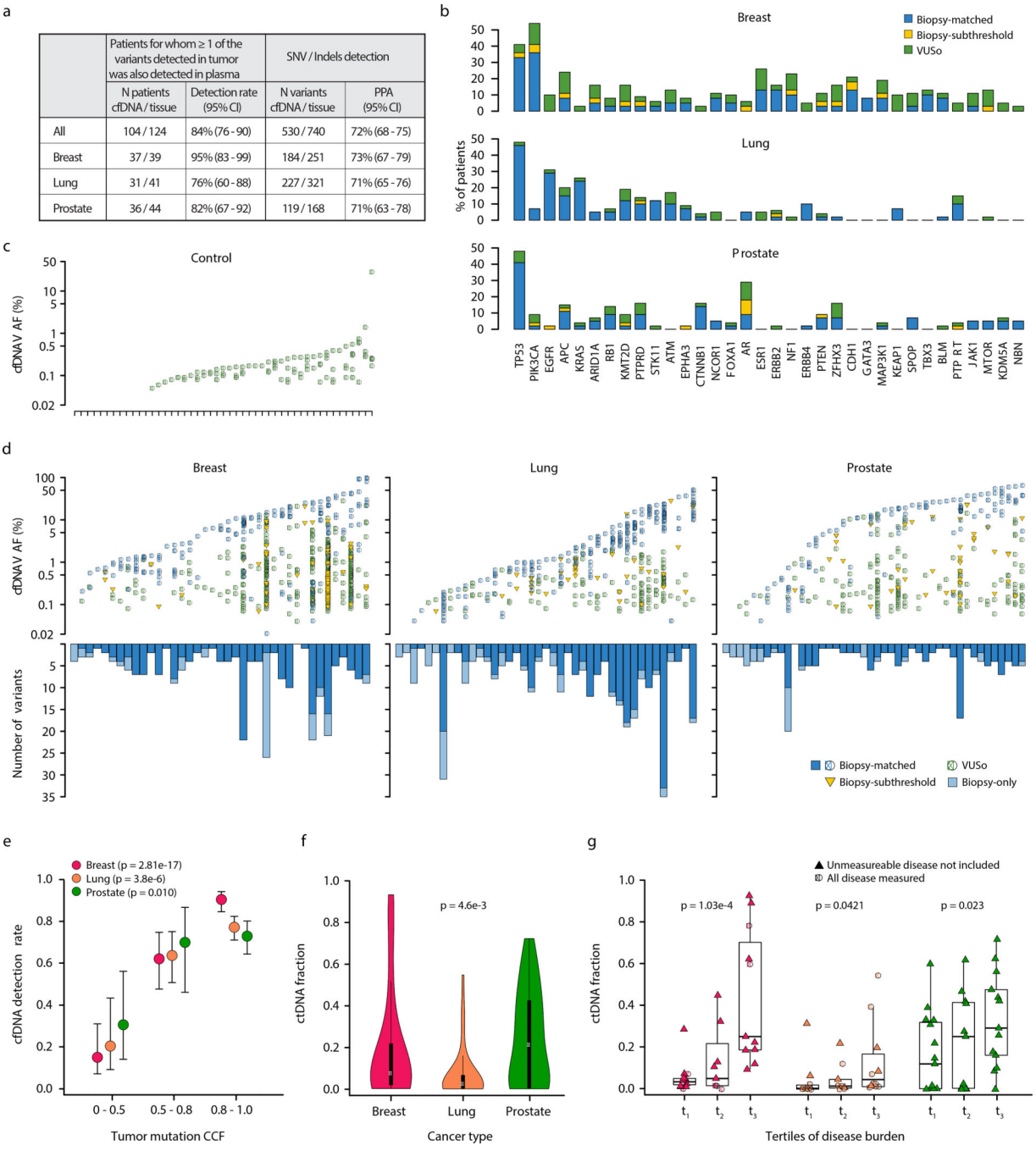


Fig. 2. Concordance of cfDNA variants with tumor biopsy.

(**a**) Summary statistics of concordance between the 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 by in cfDNA of the same patients with MBC, NSCLC, and CRPC. The genes were sorted by their frequency of alterations in the tumor tissue. (**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, biopsy-only (detected in tumor only and not in cfDNA), or VUSo. (**e**) Increasing detection rate of tumor variants in cfDNA with clonality of mutations in the tumor biopsy. The midpoint of the interval plot shows the median proportion of tumor mutations from the MSK-IMPACT assay which were also detected in cfDNA of MBC, NSCLC, and CRPC patients, stratified by the cancer cell fraction (CCF) in the tumor, whilst the error bars indicate the 95% binomial confidence intervals. The CCF was strongly associated with detection rate in cfDNA (overall p= 5.33e-21). All p values are based on two-sided X^2 trend test. (**f**) Distribution of tumor derived cfDNA fraction estimates (i.e. ctDNA fraction) in MBC, NSCLC, and CRPC patients (n=105 patients with evaluable ctDNA fraction, two-sided Kruskal-Wallis *H*-test p=0.0046). The midpoints indicate the median ctDNA fraction by cancer type and the violins extend to the full range of the data. (g) Distribution of ctDNA fraction estimates as a function of disease burden. In MBCs (n=34) and NSCLC (n=29), disease volume was obtained through volumetric measurements of pre-cfDNA collection CT scans. In CRPC (n=39), the automated bone scan index (aBSI) was used to estimate disease burden. The association between tertiles of disease burden for each cohort and the ctDNA fraction was estimated using a one-sided Jonckheere-Terpstra test for increasing ctDNA fraction. Triangles indicate patients from whom some distant metastases could not be measured and the estimates for these lesions were not included in the volumetric assessment. Note that as aBSI was employed for CRPC, not all sites of metastatic disease (e.g. visceral disease) were included in the disease burden in the CRPC patients. The horizontal bar indicates the median, while the boxes shows the interquartile range (IQR). The whiskers extend to 1.5 × IQR on either side.

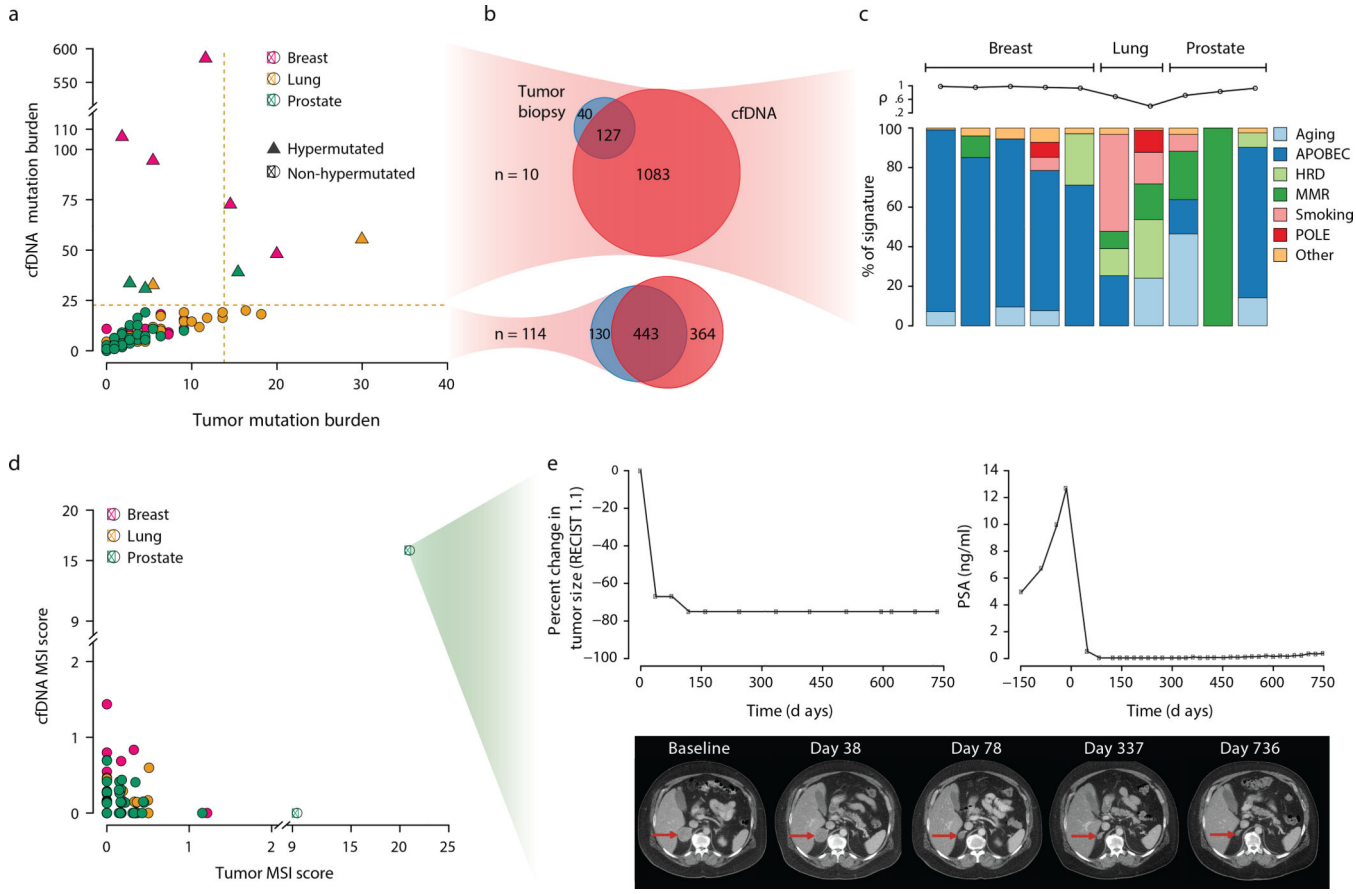


Fig. 3. Tumor mutational burden and mutational signatures derived from cfDNA targeted assay.

(**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 and tumor 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 algorithm[40] 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.

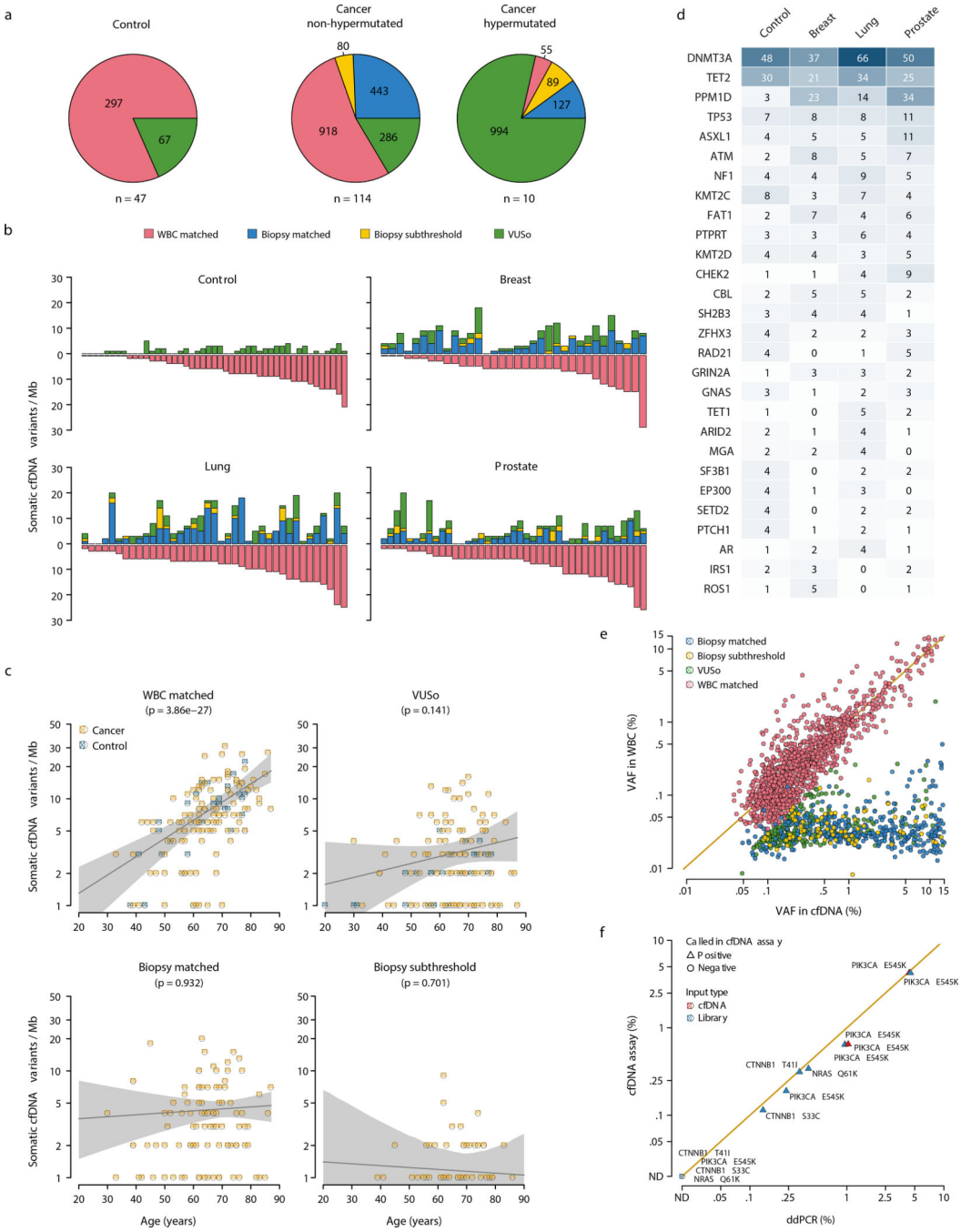


Fig. 4. Characterization of biological sources and composition of cfDNA variants.

(**a**) Pie charts representing the distribution of cfDNA somatic mutation for controls (n=47), non-hypermutated cancer cases (n=110) and hypermutated cancer cases (n=10). (**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 (x-axis)

and number of cfDNA variants per Mb categorized as WBC-matched, VUSo, tumor biopsy-matched and biopsy-subthreshold. In all panels, the p values were obtained using two-sided Wald tests on the coefficients of zero-inflated Poisson regression models with cancer status and smoking history where applicable as covariates. The cohort consists of n=114 non-hypermutated cancer patients and n=47 non-cancer controls. Only cases with a non-zero number of variants are displayed. (**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 distribution in VAFs of somatic mutations detected in cfDNA and WBC using the targeted DNA assay and color coded according to source of origin for n=114 non-hypermutated cancer patients and n=47 non-cancer controls. (**f**) Orthogonal validation of VUSo detected in cfDNA using ddPCR. The VAF measured using ddPCR (x-axis) was plotted according to the cfDNA targeted assay (y-axis). Plasma cfDNA samples and pre-enrichment libraries from seven cancer patients with hotspot mutations not detected in the matched tumor sequencing were subjected to four ddPCR assays. For one patient, only cfDNA isolated from plasma was available. For two patients, both cfDNA and pre-enrichment sequencing libraries were available whilst for the remaining four patients, only libraries were assayed. The sequencing libraries from 12 patients where the ddPCR target variants were not detected by sequencing were used as negative controls. All experiments were performed in triplicate. In (**e**) and (**f**), the diagonals represent the line $y=x$. ND; not detected.

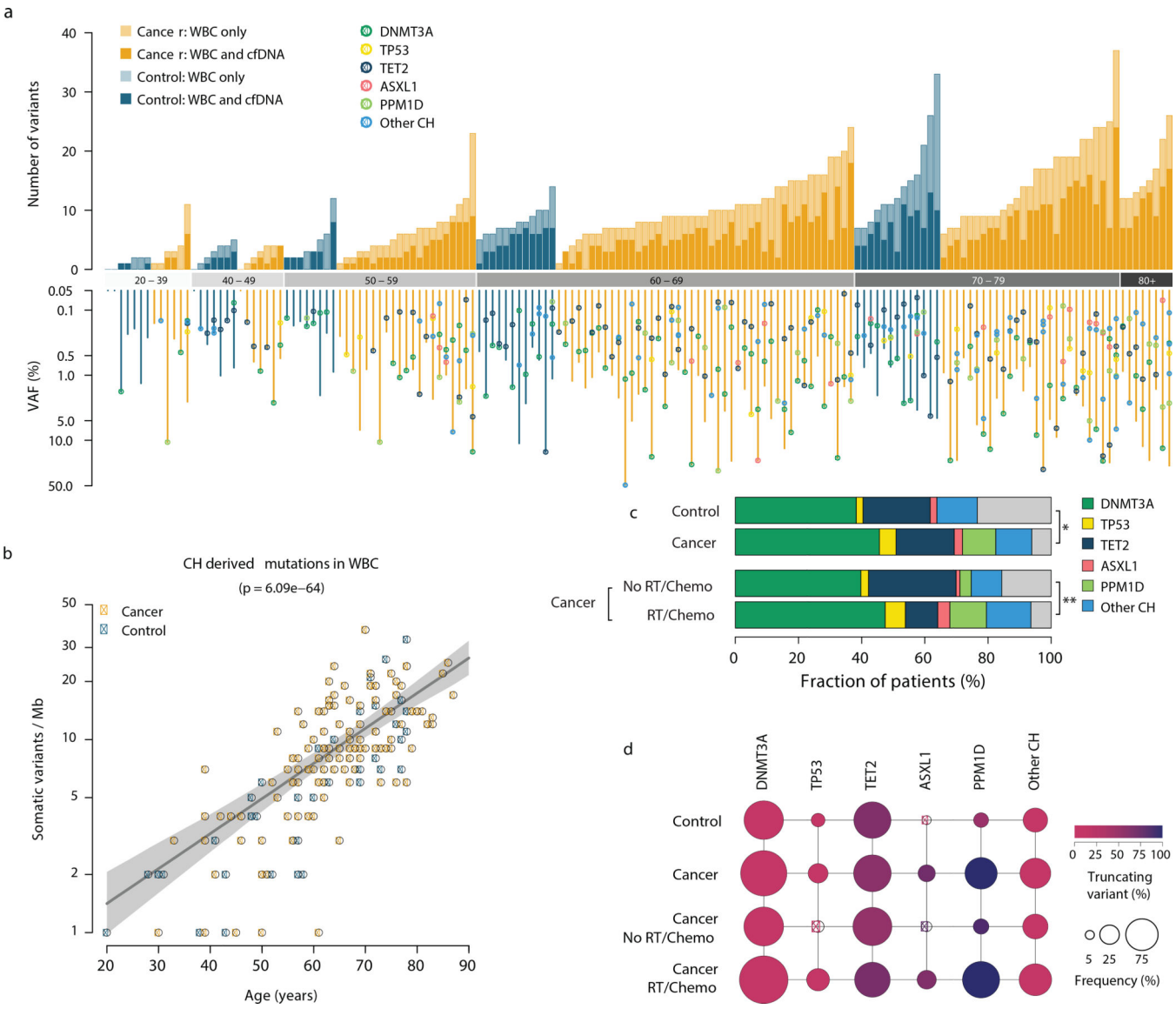


Fig. 5. Characterization of WBC variants.

(**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. The bottom panel shows the variant allele fractions (VAFs) of all somatic variants in 15 canonical genes associated with clonal hematopoiesis (CH) together with the variant occurring at maximal VAF in WBC. (**b**) Association of age (x-axis) and number of somatic variants in WBC per Mb (y-axis). The p value was obtained using a two-sided Wald test on the coefficients of a zero-inflated Poisson regression with cancer status as covariate. The analysis included 47 controls and 124 cancer cases. (**c**) Bar plot showing the percentage of cancer patients (n=114) and control individuals (n=47) harboring a mutation with maximal VAF in a given CH gene. * indicates p=0.0115 and ** indicates p=0.0008. The p values were obtained using a permutation-based likelihood ratio test to assess the significance of the coefficients of a logistic regression with age and smoking history as covariates where applicable. (**d**) Frequency of mutations in CH genes as a function

of the number of non-cancer or cancer patients in the given arm and colored according to the percentage of truncating mutations including frameshifting indel, nonsense and nonstop mutations. Note that some of these patients have ≥1 variant affecting the same canonical CH genes (e.g. *DNMT3A*, *TET2*, *PPM1D*, and *ASXL1*). The sum of the size of the circles can, therefore, exceed 100%. In all panels, the cohort consists of n=114 cancer patients and n=47 non-cancer controls. In (**b**), only cases with a non-zero number of CH variants are displayed.