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

Pedram Razavi, Bob T. Li, David N. Brown, Byoungsok Jung, Earl Hubbell, Rongli Shen, Wassim Abida, Krishna Juluru, Ino De Brujin, Chenlu Hou, Oliver Venn, Raymond Lim, Aseem Anand, Tara Maddala, Sante Gnerre, Ravi Vijaya Satya, Qinwen Liu, Ling Shen, Nicholas Eattcock, Jeanne Yue, Alexander W. Blocker, Mark Lee, Amy Sehnert, Hui Xu, Megan P. Hall, Angie Santiago-Zayas, William F. Novotny, James M. Isbell, Valerie W. Rusch, George Plitas, Alexandra S. Heerdt, Marc Ladanyi, David M. Hyman, David B. Solit, Alexander M. Aravanis and Jorge S. Reis-Filho

Accurate identification of tumor-derived somatic variants in plasma circulating cell-free DNA (cfDNA) requires understanding of the various biological 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 DNA covering a large genomic region (508 genes; 2 megabases; >60,000× raw depth) in a prospective study of 124 patients with metastatic cancer, with contemporaneous matched tumor tissue biopsies, and 47 controls without cancer. The assay displayed 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 patients with cancer) had features consistent with clonal hematopoiesis. This cfDNA sequencing approach revealed that clonal hematopoiesis constitutes a pervasive biological phenomenon, emphasizing the importance of matched cfDNA-white blood cell sequencing for accurate variant interpretation.

Circulating cell-free DNA (cfDNA) in the plasma of patients with cancer constitutes a potential source of tumor-derived DNA. Massively parallel sequencing analysis of cfDNA samples from patients with cancer 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 biological features, including histology, vascularization, proliferation and apoptosis rates. ctDNA fraction is extremely low in many early-stage and some metastatic cancers, requiring methods to detect mutations at extremely low allele fractions. Most previous studies focused on analysis of patients with advanced disease using a panel of hotspot mutations or limited genomic regions of key cancer genes sequenced at high depths, a large number of genes at moderate sequencing depths or a combination of methods to define cfDNA fraction using shallow whole-genome sequencing or targeted methods followed by whole-exome analysis of samples with high cfDNA fraction.

Even when accurate cfDNA assays are utilized, cfDNA sequencing results may still be confounded by biological signals arising from somatic mosaicism. One form of somatic mosaicism is clonal hematopoiesis (CH), which results from the accumulation of somatic mutations in hematopoietic stem cells that are clonally propagated to their progeny. These somatic mutations may provide a fitness advantage to some hematopoietic stem cells and/or their descendant cells, resulting in their disproportionate expansion or arise through neutral drift. CH increases with age and occurs in up to 31% of older individuals. Additional somatic variants not present in tumor biopsies, but in cfDNA only, have also been documented; however, their nature and source (tumor derived versus other sources) 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 the 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 a Food and Drug Administration (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 in control participants without cancer (Methods). Briefly, plasma cfDNA and matched WBC genomic DNA (gDNA) from patients with MBC, NSCLC or CRPC and from controls without cancer 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 and Supplementary Table 1). In patients with cancer, 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 Clinical Laboratory Improvement Amendments-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 and Supplementary Table 1)\(^{10,33,34}\). For the purpose of comparison with tumor biopsies, only variants mapping to the intersection of the 410 genes present in the two gene panels were considered.

Of 161 potentially eligible patients with cancer (53 MBC, 53 NSCLC and 55 CRPC) enrolled between 24 September 2015 and 1 August 2016, 124 (39 with MBC, 41 with NSCLC and 44 with CRPC) were included in the concordance subset (evaluable for both tumor biopsy and ctDNA) (Extended Data Fig. 1). The clinicopathological 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 control samples from individuals without cancer, 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 the matched tumor biopsy and WBC gDNA samples from the patients with cancer (Methods and Fig. 1a). The high-intensity cfDNA sequencing approach simultaneously analyzed plasma cfDNA and WBC gDNA using a targeted assay spanning approximately 2 megabases (Mb) and utilizing unique molecular identifiers to suppress technical assay errors at a minimum raw average target depth of 60,000× (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 and Extended Data Fig. 3a,b). Together, this resulted in an assay with the performance characteristics necessary for the detection of mutations near the molecular limits (high technical sensitivity; Fig. 1b,c), a low false positive rate (<1 error in 1 million bases sequenced; Extended Data Fig. 3c–h), high reproducibility in independent biological replicates (Fig. 1d,e and 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 $1 \times 10^{-5}$ to $3 \times 10^{-5}$ (Supplementary Methods, Extended Data Figs. 2 and 3 and Supplementary Tables 3–5), which is comparable to that of other high-fidelity cfDNA sequencing assays\(^{15,25,26}\).

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 previous 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 MBC samples (95%; 95% CI: 83–99%) was significantly higher than in NSCLC samples (76%; 95% CI: 60–80%; $P=0.0258$) and comparable to that of CRPC samples (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 MBC (73%; 95% CI: 67–79%), NSCLC (71%; 95% CI: 65–76%) and CRPC samples (71%; 95% CI: 63–78%).

Next, we 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 cutoffs)\(^{14,15}\), or were neither detected in the tumor biopsy nor in WBCs (variants of unknown source (VUSOs)).

Given the low false positive rates and accuracy of the assay for measuring variant allele fractions (VAFs) (Fig. 1d–f and Extended Data Fig. 4), we quantified the VAFs of somatic mutations not present in the WBCs in controls and patients with cancer. 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 MBC, NSCLC and CRPC samples, 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).

Next, we 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 (CCFs) in the tumor biopsies (the percentage of cancer cells within a biopsy biopsies nor WBCs, may derive from multiple origins but comprise...
Tumor mutation burden (TMB) and mutational signatures. We hereby constrained biopsy. When applying high TMB by cfDNA, 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 (TMB) and mutational signatures. We performed an exploratory analysis to determine whether TMB, mutational signatures \(^{11,13}\) and microsatellite instability (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\(^{-1}\) cut-off \(^{13}\)), including two MBC samples, three NSCLC samples and one CRPC sample. Four of these cases were also classified as having high TMB by cfDNA (>22.7 mutations Mb\(^{-1}\); 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\(^{-1}\), 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). Potential explanations for this observation include spatial tumor heterogeneity between metastatic sites, with only some sites exhibiting a hypermutator phenotype. The ten hypermutated cfDNA samples accounted for 75% of
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). PPA, positive percent agreement. b. Frequency of genomic alterations 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 VAFs of somatic variants sorted by the maximum VAF in control individuals. d. Top: plasma VAFs of somatic variants in MBC, NSCLC and CRPC. Bottom: numbers 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 that were also detected in cfDNA of patients with MBC, NSCLC and CRPC, respectively, stratified by the CCF in the tumor. Error bars indicate 95% binomial CIs. The CCF was strongly associated with detection rate in cfDNA (overall P = 5.33 × 10⁻⁹). All P-values are based on two-sided X² trend test. f. Distribution of tumor-derived cfDNA fraction estimates (that is, ctDNA fraction) in patients with MBC, NSCLC and CRPC (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 cfDNA fraction estimates as a function of disease burden. For patients with MBC (n = 34) and NSCLC (n = 29), disease volume was obtained through volumetric measurements of pre-cfDNA collection computerized tomography scans. In CRPC (n = 39), the 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 (for example, visceral disease) were included in the disease burden in the patients with CRPC. Horizontal bars indicate median values, while the boxes show the IQR. The whiskers extend to 1.5x the IQR on either side.
the cfDNA biopsy-subthreshold mutations and VUSOs across the entire cohort (Fig. 3b), and displayed mutational signatures consistent with the modalities of genetic instability known to occur in patients with MBC, NSCLC and CRPC. All patients with hypermutated MBC (n = 5) and one of the three with hypermutated CRPC displayed the APOBEC (apolipoprotein B messenger RNA editing enzyme, catalytic polypeptide-like) mutational signatures known to be acquired in the evolution of MBC and CRPC, consistent with the results of previous analysis of NSCLC, the mutational signatures of the two patients with hypermutated NSCLC comprised the smoking-related signature and a combination of other signatures, including APOBEC, homologous recombination DNA repair deficiency and loss of function of mismatch repair (MMR; Fig. 3c and Extended Data Fig. 5f).

High MSI is a biomarker of response to immune checkpoint inhibitors. We therefore assessed the MSI status of the cohort utilizing MSIsensor, adjusted for the ultra-high sequencing depth of cfDNA, tumor biopsy and matched normal WBCs (Supplementary Methods and Supplementary Fig. 1). Our analysis revealed one CRPC sample with genomics features of high MSI (Fig. 3d), which was also found to display a dominant APOBEC mutational signature (Fig. 3c). This patient with CRPC received an anti-programmed death-ligand 1 inhibitor and displayed rapid and sustained tumor regression, as defined by the response evaluation criteria in solid tumors (Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1) and prostate-specific antigen 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 TMB, 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% (739/2,983 mutations) of all somatic mutations detected in the cfDNA of patients with cancer (Fig. 4a). Notably, a median of 7.27 mutations Mb−1 (range: 0.91–20.91

**Fig. 3** | TMB and mutational signatures derived from cfDNA targeted assay. a, Distribution of the somatic TMB, defined as the number of nonsynonymous mutations Mb−1, 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−1), and the horizontal dashed line indicates the threshold for samples with a high TMB in cfDNA (22.7 mutations Mb−1). b, Venn diagrams showing the total number of mutations detected in cfDNA and tumor and their overlap. Top: distribution of mutations in the ten hypermutated cases (MBC, n = 5; NSCLC, n = 2; CRPC, n = 3). Bottom: distribution of mutations in the remaining 114 patients (MBC, n = 34; NSCLC, n = 39; CRPC, n = 41). The ten hypermutated cases account for 60% of the total cfDNA variants and 75% of cfDNA-only variants (VUSOs).

c. Bottom: Bar charts displaying the fraction of mutational signatures in the hypermutated cases. Top: Pearson correlation between the observed and expected 96 base substitutions profile. All of the MBC cases and one of the CRPC cases demonstrated a dominant APOBEC signature. HRD, homologous recombination deficiency; POLE, polymerase. d, MSI scores, obtained using a modified MSIsensor algorithm, from the tumor (x axis) and cfDNA (y axis). e, A 55-year-old patient with castration- and enzalutamide-resistant prostate cancer displayed an MMR signature and high MSI score based on the smoking-related signature and a combination of other signatures, including APOBEC, homologous recombination DNA repair deficiency and loss of function of mismatch repair (MMR; Fig. 3c and Extended Data Fig. 5f).
Fig. 4 | Characterization of biological sources and composition of cfDNA variants. a, Pie charts representing the distribution of cfDNA somatic mutations for controls (n = 47), non-hypermutated cancer cases (n = 114) and hypermutated cancer cases (n = 10). b, Bar plots showing the number of somatic variants detected in plasma cfDNA per megabase (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 colors are as in a. c, Association between age (x axis) and the number of cfDNA variants per megabase (y axis). d, Orthogonal validation of VUSOs detected in cfDNA using ddPCR. e, Posterior distribution of VAFs. The scatter plot shows the distribution of VAFs in WBCs (%). f, Orthogonal validation of VUSOs detected in cfDNA using ddPCR.

Data points are color coded according to the source of origin for n = 114 patients with non-hypermutated cancer and n = 47 controls without cancer.
mutations Mb$^{-1}$ were detected in the cfDNA samples of controls without cancer (Fig. 4a). Although previous studies have suggested that these alterations probably constitute technical artifacts of ultra-high-depth sequencing analysis$^{45,44}$, 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 cases of hypermutated cancer).

First, we 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 control patients without cancer, the vast majority (81.6%; 297 of 364) of somatic mutations detected were also identified in WBCs, suggesting that these somatic genetic alterations were probably not technical artifacts, but rather a result of CH (Fig. 4a). Likewise, the majority (53.2%; 918/1,727) of mutations identified in cfDNA samples of patients with non-hypermutated cancer were also WBC matched (Fig. 4a). Importantly, the number of WBC-matched cfDNA variants in patients with cancer 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 and Extended Data Fig. 6a). As CH is related to age$^7$, 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 biopsy-matched or biopsy-subthreshold variants, significantly correlated with age (smoking-adjusted $P=3.86 \times 10^{-12}$; Fig. 4c). Based on this interpretation, the cfDNA and WBC sequencing analysis performed here suggests that 89.5% of patients with cancer and 83% of controls without cancer have evidence of CH in their cfDNA (Fig. 4b). Consistent with recent observations in a population without cancer$^9$, and with the notion that these mutations constitute CH events, the vast majority of WBC-matched somatic mutations detected in the cfDNA of patients with and without cancer involved canonical CH genes, such as DNMT3A, TET2, PPM1D and TP53 (Fig. 4d and Extended Data Fig. 6b)$^{45}$, 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 and Extended Data Fig. 6c,d), making it unlikely that they resulted from systematic sequencing errors or background noise.

Previous studies of hypermutated tumors have shown significant spatial heterogeneity resulting in numerous subclonal mutations private to each tumor site$^{46}$. Consistently, the overall proportion of tumor-matched mutations (biopsy matched and biopsy subthreshold) was significantly lower in the ten patients with hypermutation (17.2%; 216/1,210; non-WBC-matched mutations) than in the 114 non-hypermutated patients (30.3%; 523/1,727; $P=1.2 \times 10^{-16}$), whereas a higher proportion of tumor-derived variants were subclonal biopsy-subthreshold variants in hypermutated cases (41.2% (89/216) versus 15.3% (80/523), respectively; $P=1.7 \times 10^{-10}$). These findings support the notion that a single tumor biopsy may not always capture the full landscape of the tumor mutational profile in patients whose tumors harbor a hypermutator phenotype.

Given the correlation between sequencing depth and occurrence 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 $>10,000 \times$; 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 the 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.

Next, we sought to define the biological source of nonsynonymous VUSOs in cfDNA (Methods). After removing variants with a known source of origin (WBC matched; biopsy matched and biopsy subthreshold in patients with cancer), 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 patients with cancer, 77.7% (994/1,280) of the VUSOs were detected in the ten samples of hypermutated cancer. In fact, VUSOs accounted for 82.1% (994/1,210) of the total non-WBC-matched somatic cfDNA mutations in hypermutated samples, compared with 35.4% (286/809) in non-hypermutated tumors ($P=9.3 \times 10^{-10}$). Additionally, VUSOs rarely constituted the mutation at the highest VAF in patients with cancer ($17.6, 23.1$ and $19.5\%$ of patients with cancer; $17.6, 23.1$ and $19.5\%$ of patients with MBC, NSCLC and CRPC, respectively). These findings indicate that a large proportion of the VUSOs probably 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 VUSOs further, we evaluated the genes harboring variants classified as such in patients with cancer (Fig. 2b and Extended Data Fig. 5a). A subset of VUSOs 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 and Extended Data Fig. 5a)$^8$. Furthermore, the VAF distribution for these mutations mostly mirrored that of biopsy-matched variants (Fig. 4e and Extended Data Fig. 6c,d). In hypermutated cases, however, a significant correlation between the size of the sequenced coding region of a gene harboring VUSOs and the number of VUSOs affecting the given gene was observed ($P=4.4 \times 10^{-16}$; Extended Data Fig. 5b). To determine the accuracy of the cfDNA assay for detecting VUSOs, we performed orthogonal validation of our results utilizing ddPCR assays targeting VUSOs (Methods and Fig. 4f) and found complete agreement between the two assays. Based on these results, we posit that these VUSOs 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 VUSOs also included canonical CH genes (Extended Data Fig. 5c). Consistent with the notion that at least a subset of VUSOs arise from CH or other sources of somatic mosaicism not present in matched WBC samples, VUSOs were weakly associated with age at sample collection ($P=0.141$; Fig. 4c) and affected canonical CH genes in both patients with cancer and controls (Extended Data Fig. 5c–e), with some having similar allele frequencies to WBC-matched variants (Fig. 4e and Extended Data Fig. 6c,d).

Taken together, this high-intensity cfDNA sequencing assay identified that CH mutations are the most likely origin of non-tumor-derived mutations detected in cfDNA, CH is probably more prevalent than was previously reported with lower-depth WBC sequencing approaches$^{4,16,20,22–24}$, and 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 and Methods). At least one CH mutation was
detected in 99.1% of the WBCs of the patients with cancer analyzed, and in 93.6% of the controls without cancer. 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=6.09×10^{-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 controls without cancer and patients with metastatic cancer (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 patients with cancer than in controls (age-adjusted P=0.0115; Fig. 5c).

Fig. 5 | Characterization of WBC variants. a, Direct analysis of somatic variants in WBCs. Top: bar plot showing the number of somatic variants detected across 1.1 Mb of genome, grouped by age category and ordered by increasing TMB. Bottom: VAFs of all of the somatic variants in 15 canonical genes associated with CH, together with the variant occurring at maximal VAF in WBCs. b, Association of age (x axis) and number of somatic variants in WBCs per megabase (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 the covariate. The analysis included 47 controls and 114 cases of cancer. c, Bar plot showing the percentage of patients with cancer (n=114) and control individuals (n=47) harboring a mutation with maximal VAF in a given CH gene (*P=0.0115; **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. Chemo, chemotherapy; RT, radiotherapy. d, Frequency of mutations in CH genes as a function of the number of patients with and without cancer in the given arm, colored according to the percentage of truncating mutations, including frameshifting indels, and nonsense and nonstop mutations. Note that some of these patients have ≥1 variant affecting the same canonical CH gene (for example, DNMT3A, TET2, PPM1D and ASXL1). Therefore, the sum of the size of the circles can exceed 100%. In all panels, the cohort consists of n=114 patients with cancer and n=47 controls without cancer. In b, only cases with a non-zero number of CH variants are displayed.
In addition, mutations affecting PPM1D—in particular truncating variants preferentially affecting the carboxy-terminal domain (Fig. 5d and Extended Data Fig. 8)—were significantly more common in patients who received chemotherapy and/or radiation therapy than in those who had no previous history of such treatments (age- and smoking-adjusted \( P = 0.0008 \); Fig. 5c).

**Gene copy number variation (CNV) detection.** As an exploratory analysis, we sought to define whether the high-intensity cfDNA assay would be able to detect CNVs de novo. We observed a relatively good concordance between CNVs detected in tumor biopsies and cfDNA only in cases where the cfDNA fractions were \( \geq 10\% \) (Extended Data Fig. 9). Despite this limitation, in five patients with actionable CNVs (four patients with MBC and amplified ERBB2, and one patient with NSCL and amplified MET), 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 cfDNA fractions were 1.3 and 1.9% (Extended Data Fig. 10). None of the remaining samples tested harbored amplifications of these two genes, thereby showing the specificity. However, the performance of the cfDNA assay to detect gene amplifications was found to be highly dependent on the cfDNA 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 to broaden 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\(^{25,26}\). 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 unique molecular identifiers 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 (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 (for example, 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\(^{28-31}\), 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 (Fig. 1d,e), this cfDNA assay also allowed for the de novo assessment of TMB and mutational signatures, including MSI (Fig. 3a–c), 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 patients with and without cancer. These WBC-matched mutations preferentially affected genes previously implicated in CH\(^{22,24,26}\), 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 patients with metastatic disease. The higher prevalence of CH found in WBCs in this study (93.6% of controls without cancer and 99.1% of patients with cancer) relative to that reported in previous studies\(^{25,28}\) probably resulted from the high sensitivity of the assay employed to detect variants in the WBC samples\(^{25,28}\), and was consistent with the observation by Liu et al.\(^ {27}\) in a non-cancer population. 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\(^ {29,31}\) have shown 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 (for example, CH) cfDNA variants being interpreted as tumor derived\(^ {29,31}\), and support the need for joint analysis of cfDNA and matched WBCs, given that mutations related to CH may result in inaccurate TMB and mutational signature quantification\(^ {29}\).

We also demonstrated that VUSOs 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 VUSOs were found to be tumor derived and arose from minor tumor subclones, and 77.7% of all VUSOs in patients with cancer were identified in ten patients whose tumors harbored hypermutator mutational processes, such as APOBEC, known to amplify tumor heterogeneity and subclonal diversity. Hence, in this context, high-intensity cfDNA assays may offer a more comprehensive landscape of the 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; however, a recent study has shown the importance of cfDNA sequencing in defining the heterogeneity and mechanisms of therapeutic resistance in advanced colorectal cancers\(^ {41}\). 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\(^ {4}\). However, this caveat 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 the 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 MBC, NSCLC and CRPC and their respective subtypes. Additionally, <50 baseline samples from healthy controls 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, but even when this approach is employed, a subset of VUSOs 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 patients with cancer, and provides further evidence that CH probably constitutes a biological phenomenon and a technical pitfall more prevalent than was previously anticipated.

**Online content**

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data code availability are available at [https://doi.org/10.1038/s41591-019-0652-7](https://doi.org/10.1038/s41591-019-0652-7).

Received: 23 January 2019; Accepted: 15 October 2019; Published online: 25 November 2019

**References**

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

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

3. Diaz, I. A., Jr & Bardelli, A. Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.* 32, 579–586 (2014).

4. Wan, J. C. M. et al. Liquid biopsies come of age: towards implementation of next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* 17, 251–264 (2015).

5. Newman, A. M. et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* 34, 547–555 (2016).

6. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. USA* 108, 9530–9535 (2011).

7. Szczesny, K. et al. The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* 34, 427–438 (2018).

8. Alexandria, L. B. et al. Signatures of mutational processes in human cancer. *Nature* 500, 415–421 (2013).

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

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

11. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* 49, 1478–1486 (2017).

12. Gerstner, C. et al. Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. *Cancer Cell* 34, 996–1011.e8 (2016).

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

14. Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* 372, 2599–2620 (2015).

15. Merker, J. D. et al. Circulating tumor DNA analysis in patients with cancer: American Society of Clinical Oncology and College of American Pathologists joint review. *Arch. Pathol. Lab. Med.* 142, 1242–1253 (2018).

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

17. Schullitse, A. M. et al. Massive parallel sequencing-based clonality analysis of synchronous endometrioid endometrial and ovarian carcinomas. *J. Natl Cancer Inst.* 108, djv427 (2016).

18. Hsu, J. et al. PPM1D mutations drive clonal hematopoiesis in response to cytotoxic chemotherapy. *Cell Stem Cell* 23, 700–713.e6 (2018).

19. Bettagoada, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* 6, 224ra24 (2014).

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

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

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

23. Swanton, C. et al. Prevalence of clonal hematopoiesis of indeterminate potential (CHIP) measured by an ultra-sensitive sequencing assay: exploratory analysis of the Circulating Cancer Genome Atlas (CCGA) study. *J. Clin. Oncol.* 36, 12003 (2018).

24. Mansukhani, S. et al. Ultra-sensitive mutation detection and genome-wide DNA copy number reconstruction by error-corrected circulating tumor DNA sequencing. *Clin. Chem.* 64, 1626–1635 (2018).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

Study design. This was a prospective observational study of patients with MBC, NSCLC and CRPC designed to characterize the detection of variants in plasma cfDNA using a targeted DNA assay (GRAIL), and to evaluate the concordance of variants detected in tumor tissue and plasma as evidence of cfDNA detection. The primary objectives were to assess the tumor cfDNA detection rate based on observing at least one variant (single-nucleotide variants (SNVs) or indels) and to assess the concordance of the MSK-IMPACT variants\(^3\)^\(^4\) detected in tumor biopsy samples versus cfDNA. Secondary objectives included assessing the cfDNA detection rate based on observing at least one MSK-IMPACT variant, characterizing the cfDNA detection rate as a function of the type of variant (SNVs or indels) and 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 DNA extraction, and the review of patient medical records. Detailed demographic, pathological and treatment information under an Institutional Review Board-approved biospecimen umbrella protocol (Memorial Sloan Kettering Cancer Center (MSKCC) protocol 12-245; ClinicalTrials.gov ID: NCT01779072). At least 50 patients with 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 previous 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 were allowed, provided if enrolled before 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 (for example, hormone therapy) were eligible. Blood was drawn within 6 weeks of tissue biopsy for MSK-IMPACT analysis either before or after tissue biopsy. Whole-blood samples received outside of the stability timeframe for Streck DNA BCT (5 d) were excluded.

Fifty de-identified whole-blood samples from self-reported healthy individuals (with no diagnosis of cancer) were obtained from the San Diego Blood Bank. Limited clinical data were provided with the samples. Healthy participants were required to be at least 20 years of age, to be eligible for blood donation per Institutional Review Board-approved biospecimen umbrella protocol (Memorial Sloan Kettering Cancer Center (MSKCC) protocol 12-245; ClinicalTrials.gov ID: NCT01779072). At least 50 patients with 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 previous lines of therapy) were collected from medical records. Whole-blood samples received outside of the stability timeframe for Streck DNA BCT (5 d) were excluded.

Fifty de-identified whole-blood samples from self-reported healthy individuals (with no diagnosis of cancer) were obtained from the San Diego Blood Bank. Limited clinical data were provided with the samples. Healthy participants were required to be at least 20 years of age, to be eligible for blood donation per standardized assessment and criteria, to lack a diagnosis of cancer and to have no previous history of cancer. Participants were excluded if they had a previous history of cigarette smoking for at least 1 year, a current history of cigarette smoking, were pregnant, had a personal history of cancer or had previous medical or surgical treatment of any type of cancer. The results were not returned to any patients, health care providers or the San Diego Blood Bank.

Whole-blood sample accessioning, processing and analysis. For the 161 patients with cancer, tumor DNA was extracted from formalin-fixed paraffin-embedded biopsy samples, and matched normal DNA was extracted from mononuclear cells from peripheral blood. All specimens underwent next-generation sequencing in the NSCLC and MBC cohorts. Clinical Laboratory Improvement Amendments-certified laboratory using MSK-IMPACT—an FDA-authorized hybridization capture-based next-generation sequencing assay—which analyzes all protein-coding exons of 410 cancer genes was selected from plasma variants for plotting and comparison. Somatic variants also present in the MSK-IMPACT sequencing of the tumor biopsy were labeled as biopsy matched if they were above the limit for variant calling as required for clinical reporting. Variants not matched were labeled as VUSOs.

Tumor concordance. Overall agreement between variants in plasma and tumor tissue was measured using positive percent agreement with tumor tissue as the reference; this can be expressed as the percentage 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 also present in the MSK-IMPACT sequencing of the tumor biopsy were labeled as biopsy matched if they were above the limit for variant calling as required for clinical reporting. Variants not matched were labeled as VUSOs.

Whole-blood sample accessioning, processing and analysis. Peripheral blood from patients with metastatic cancer was collected into two 10-ml Cell-Free DNA BCTs (Streck) at MSKCC and shipped to GRAIL at room temperature. Whole blood from healthy individuals drawn into Streck BCTs were purchased from the San Diego Blood Bank and shipped to GRAIL at room temperature. Retrieved whole-blood Streck BCTs were separated into plasma anduffy coat and stored at −80°C unless processed the same day.

cDNA was extracted from two tubes of plasma (up to a combined volume of 8 ml) per subject using a modified QIAamp Circulating Nuclear Acid kit (Qiagen). Extracted cfDNA was quantified using the Fragment Analyzer High Sensitivity NGS kit (Advanced Analytical Technologies). cfDNA from fresh blood (paired plasma and Buffy coat from the same blood tube) was extracted using the Qiaqen DNeasy Blood and Tissue kit. Extracted gDNA was quantified using NanoDrop (Termo Fisher Scientific) and fragmented to a mean size of 180 base pairs using the Covaris E220 ultrasonicator. Sheared gDNA was subsequently size selected using Agencourt AMPure XP magnetic beads (Beckman Coulter), then quantified using the Fragment Analyzer Standard Sensitivity NGS kit (Advanced Analytical Technologies).

Library preparation, target enrichment and sequencing. Buffy coat gDNA (50 ng) and plasma cfDNA (≤575 ng) were used for next-generation sequencing 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 CH of indeterminate potential and other artifacts.

Source of origin of plasma variants. Somatic variants were labeled by de novo assembly from control and cancer samples and were stacked, and their sources 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 somatic. Second, variants with an allele fraction of >20% matched in WBCs 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 variation. Variants not labeled were filtered to exclude ambiguous (no positive evidence for variant alleles in WBCs, 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 VUSOs.

Tumor burden and cfDNA fraction. The cfDNA 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, and then we derived the cfDNA fraction based on the VAF in cfDNA of the biopsy-matched clonal mutations.

Of the 80 patients in the NSCLC and MBC cohorts, 77 had computed tomography scans available from which volumetric tumor measurements could be obtained. Of these, 34 of the exams were computed tomography scans of the chest, abdomen and pelvis with intravenous contrast, obtained as part of a positron emission tomography/computed tomography exam; 32 exams were computed tomography scans of the chest, abdomen and pelvis without intravenous contrast; five exams were computed tomography scans of the chest only with intravenous contrast; four exams were computed tomography scans of the chest only without intravenous contrast; and two exams were computed tomography scans of the chest and abdomen with intravenous contrast. Exams were acquired on several different scanners at slice thickness ranging from 3.75–5.00 mm.

All exams were reviewed by a board-certified radiologist specializing in imaging of the chest, abdomen and pelvis (K.J.). All metastatic lesions >1 cm in size were identified. Volumes were measured using a software tool for 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 INtution advanced visualization software (version 4.4.13.P3; TeraReaco).
Of the 77 patients with available volumetric assessment, 34 with MBC and 29 with NSCLC had evaluable ctDNA fraction and were included in this analysis. Given that the majority of patients with CRPC included in this study had extensive bone disease and had undergone bone scans before enrollment, in the study, the approach employed for the volumetric assessment of disease burden was different from that used for patients with MBC and NSCLC. For patients with prostate cancer, we generated the abSI (platform version 3.3; EXINI Diagnostics)—a fully quantitative assessment of a patient’s bony disease on a bone scan to determine the number of lesions, area, and 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.

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 these weights were calculated as the mean of all such fractions. The abSI methodology utilized in this study has been shown to be an objective measure of the quantitative change in bone disease burden scans and a prognostic biomarker in patients with CRPC.

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−1 (ref. 33) for the tumor biopsy, while the corresponding value for cfDNA was evaluated de novo from the samples of patients with cancer as the median cfDNA burden in each sample according to the distribution of the six substitution classes (C → T, T → C, A → G, C → T, T → A, T → C and T → G) and the bases immediately 5′ and 3′ of the mutated base, producing 96 possible mutation subtypes using deconstructsig35. For the analyses described here, we focused on six signatures: (1) aging (signatures 1 and 5); (2) APOBEC (signatures 2 and 13); (3) homologous recombination repair deficiency (homologous recombination DNA repair deficiency; signature 3); (4) MMR (signature 6); (5) TET2 (signature 21); (6) POLE (signature 10).

Prevalence of CH in ctDNA. 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 the 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 CH in WBCs. Candidate variants in WBCs 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: they had previously been reported as somatic in COSMIC (version 86), Kandoth et al.60 or Chang et al.61; or they were frameshifting indels or truncating SNVs and occurred identical reference and alternate alleles recurring at ≥5% recurrence (Extended Data Fig. 7), as well as bona fide somatic variants found in the WBCs; (3) variants with VAF >30% were labeled germline and filtered out unless they were frameshifting indels or truncating SNVs and occurred in one of the 15 canonical genes known to be associated with CH; (4) variants with VAF >30% were labeled 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 the Exome Aggregation Consortium or Genome Aggregation Database35 were labeled 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, CREM, ATM and CHEK2.

Summary of variants and VAFs 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 VUSOs were identified. The median VAF in cfDNA was higher for biopsy-matched variants than for WBC-matched variants or VUSOs.

Sensitivity and specificity of the targeted DNA assay. Before analysis of patient samples with the targeted DNA assay, analytical characterization was performed using titrations of DNA from cell lines, gDNA extracted from Epstein–Barr virus-immortalized lymphoblastoid cell line NA12878 was purchased from the Coriell Institute. The HD753 Structural Multiplex Reference Standard gDNA, which contains known SNVs, indels, fusions and deletions, was purchased from Horizon Discovery (Supplementary Table 8). Fifteen DNA titrations using the HD753 standard and the NA12878 gDNA were prepared in triplicate to have nominal concentrations of 100, 10, 1.0, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001 ng/µL and 0.000000001 ng/µL. DNA titrations were verified using ddPCR (Bio-Rad) to ensure dilution accuracy (Supplementary Table 9). Following ddPCR verification, DNA mixtures were shared and size selected according to the targeted DNA assay protocol. A total of 30 ng sheared, size-selected gDNA was used for library construction, resulting in a mean collapsed target coverage of 2,430×. The measurement of specificity of the targeted DNA assay at various VAFs, using a probit regression model of variant calling status of 14 known small variant 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 (that is, if the triplicates were labeled A, B and C, the three simulated samples were AB = 0.5A + 0.2B, 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 4,577×, which is similar to the median of mean collapsed target coverages for all samples from patients with cancer reported here (4,408×). The estimated 95% limit of detection was 2× for 30 ng of input DNA (mean collapsed target coverage of 2,430×), and for simulated cases it was 4×, indicating a mean collapsed target coverage of 4,577×.

Figure 1c summarizes the specificity of the targeted DNA assay using samples from controls without cancer (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 0.008 for 30 ng of input DNA (mean collapsed target coverage of 2,430×), and for simulated cases it was 0.16, indicating a mean collapsed target coverage of 4,577×.

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 (version 1 and version 2); and (2) ddPCR analysis of biopsy-matched mutations and VUSOs (Supplementary Methods, Fig. 1d–f, Extended Data Fig. 4 and Supplementary Tables 8–10).

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

Statistical analyses. The association of VAFs between technical replicates was measured using the coefficient of determination (R2) value obtained from a linear regression. The degree in difference rate between the three cohorts was assessed using two-by-two Fisher’s exact tests, and the resulting Pvalues were adjusted using Bonferroni correction for multiple testing. The association between the tumor mutation CCF and the cfDNA detection rate (overall and stratified by cancer type) was assessed using a χ2 test (Cochran–Armitage test for trend). Exact 95% confidence intervals were calculated for the detection rate by cancer type. The difference in cfDNA fraction estimate between cancer types was assessed using a two-sided Kruskal–Wallis H test. The increase in cfDNA 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 (that is, biopsy matched, biopsy subthreshold, WBC matched and VUSOs) and age, with cancer status and smoking history as covariates where appropriate. The corresponding Pvalues 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 WBCs and age was assessed using a zero-inflated Poisson regression with cancer status as the covariate, and the corresponding Pvalue was obtained using a two-sided Wald test on the coefficient of the count model. The association of CH measured in WBCs in each of the 15 canonical CH genes and cancer status and previous 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 each case and each control were performed using two-sided t-tests stratified by cohort. Pairwise comparisons of the collapsed mean coverage of cfDNA or WBCs between the different cancer cohorts and controls without
cancer were performed using two-sided Mann–Whitney U-tests and adjusted for multiple testing using Bonferroni correction. The association of input DNA for library preparation and mean collapsed coverage of cfDNA samples was assessed using the F value obtained from a linear regression. The corresponding F value was obtained using a two-sided t-test. The pairwise differences of input DNA for library preparation between the different cancer cohorts and controls without cancer were assessed using two-sided Mann–Whitney U-tests and adjusted for multiple testing using Bonferroni correction. Pairwise comparisons of the percentage of collapsed bases with SNVs or combined SNVs and indels of cfDNA or WBCs between the different cancer cohorts and controls without cancer were performed using two-sided Mann–Whitney U-tests and adjusted for multiple testing using Bonferroni correction. The association of cfDNA fraction estimate or tumor purity and the Pearson's correlation coefficient measuring the association in log2[ratio space] between patient-matched cfDNA and tumor biopsy samples was assessed using one-sided Jonckheere–Terpstra tests for increasing cfDNA fraction or tumor purity with increasing Pearson's correlation. Pairwise comparisons of areas under the curve 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 had a different mutation burden than the controls without cancer. All statistical hypothesis tests were considered positive at $\alpha = 0.05$ and carried out in R/Bioconductor unless otherwise stated.

**Data availability**

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

**References**

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

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

57. Ullert, 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).

58. Armstrong, A. J. et al. Phase 3 assessment of the automated Bone Scan Index.

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

60. Kasdith, V. C. et al. Mutation landscape and significance across 12 major cancer types. Nature 502, 333–339 (2013).

61. Chang, M. T. et al. Accelerating discovery of functional mutant alleles in cancer. Cancer Discov. 8, 174–183 (2018).

62. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature Med. 24, 565–574 (2018).

63. Tambero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics 29, 2238–2244 (2013).

**Acknowledgements**

We thank the following GRAIL and Memorial Sloan Kettering Cancer Center associates for helpful discussions and contributions to this body of work: M. Berger, N. Schultz, C. Bain, M. Chung, M. Eriksen, T. Liu, R. Mauritz, A. Mich, J. Nguyen, Y. Park, S. Ramani, E. Scott, K. Shaikhahar, C. Tom, S. Wen, D. Reales, J. Galle, R. Cambria and members of the Memorial Sloan Kettering Office of Clinical Research. This work was supported by GRAIL and National Institutes of Health awards P30 CA08748 (Memorial Sloan Kettering Cancer Center) and R01 CA234361 (D.B.S.), the Breast Cancer Alliance Young Investigator Award (P.R.), the Breast Cancer Research Foundation (J.S.R.-F.), and Connecticutedly Directed Medical Research Programs W81XWH-15-1-0547 (J.S.R.-F.) and GC229671 (J.S.R.-F.).

**Author contributions**

P.R., B.T.L., D.B.S., A.M.A. and J.S.R.-F. conceived the study. P.R., B.T.L., J.I., W.A., K.I., C.H., A.A., R.V.S., Q.L., I.S., N.E., J.Y., H.X., M.P.H., A.S.-Z., W.N., J.M.I., V.W.R., G.P., M. Ladanyi, A.S.H., M. Lee, D.M.H., D.R.J., M.M., G.J.R., H.I.S., C.M.R., M.R., I.A.D., B.S. and A.M.A. acquired the data. 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. and J.S.R.-F. analyzed and interpreted the data. P.R., D.N.B., E.H., R.S., I.D.B., O.V., S.G., A.W.B., A.M.A. and J.S.R.-F. performed the bioinformatics and genomic analyses. P.R., D.N.B., E.H., M.P.H., A.M.A. and J.S.R.-F. wrote the manuscript, with input from all authors. All authors reviewed and approved the manuscript.

**Competing interests**

P.R. reports consulting and serving on the advisory board for Novartis, as well as receiving institutional research support from Illumina and GRAIL. B.T.L. reports consulting and serving on the advisory board for Genentech, Thermo Fisher Scientific, Guardant Health, Hengrui Therapeutics, Mersana Therapeutics and Bioscience Australia, as well as receiving institutional research support from Illumina, GRAIL, Genentech and AstraZeneca. W.A. reports consulting and advising for Clovis Oncology, Janssen, ORIC Pharmaceuticals and MORE Health, as well as receiving honoraria from CABRET, institutional research support from AstraZeneca, Zenith Epigenetics, Clovis Oncology and GlaxoSmithKline, and travel, accommodation and expenses from GlaxoSmithKline and Clovis Oncology. J.M.I. holds equity in LumaCyte and has received institutional research support from GRAIL and Guardant Health. G.P. is on the Scientific Advisory Board for Tizoma Therapeutics and has consulted for Merck, Bristol-Myers Squibb and Kyowa Hakko Kirin. D.M.H. reports stock and other ownership interests in Fount, as well as consulting and advising for Chuagi Pharmaceutical, Boehringer Ingelheim, AstraZeneca, Pfizer, Bayer, Genentech and Fount. He has also received research funding from AstraZeneca, Puma Biotechnology, Loxo and Bayer, and travel, accommodation and expenses from Genentech and Chuagi Pharmaceutical. G.J.R. received consulting fees from Genentech/Roche in 2016, as well as institutional research support for clinical research from Pfizer, Roche/Genentech and Takeda. C.M.R. has consulted on oncology drug development for AbbVie, Amgen, Ascenta, AstraZeneca, Bicycle, Celgene, Chugai, Daiichi Sankyo, Genentech/Roche. Gl. Therapeutics, Loxo, Novartis, Pharmamar and Seattle Genetics. He is also on the scientific advisory boards of Harpoon Therapeutics and Elucida Oncology. L.A.D. is a member of the board of directors of Personal Genome Diagnostics (PGxD) and Jounce Therapeutics, and is a paid consultant for PGxDs and NeoPhore. He is also an uncompensated consultant for Merck, but has received research support for clinical trials from Merck. At Johns Hopkins University, he is an inventor of multiple licensed patents related to technology used for circulating tumor DNA analyses and mismatch repair deficiency for diagnosis and therapy. Some of these licenses and relationships are associated with equity or royalty payments made directly to L.A.D. and Johns Hopkins University. He also holds equity in PGxDs, Jounce Therapeutics, Thrive Earlier Detection and NeoPhore. His spouse holds equity in Amgen. The terms of all of these arrangements are being managed by Johns Hopkins University and the Memorial Sloan Kettering Cancer Center in accordance with their conflict of interest policies. D.B.S. received honoraria and/or consulted for Pfizer, Loxo Oncology, Illumina, Interynne and Vividion Therapeutics. J.S.R.-F. reports receiving personal/consultancy fees from VolitionRx, Paige.AI, Goldman Sachs, REPARE Therapeutics, GRAIL, Ventana Medical Systems, Roche, Genentech and Invicro outside of the scope of the submitted work. B.T.L., E.H., C.H., O.V., T.M., S.G., R.V.S., Q.L., I.S., N.E., J.Y., A.W.B., M. Lee, A.S., H.X., M.P.H., W.N. and A.M.A. are (or were) GRAIL employees and hold stock and/or other ownership interests in GRAIL. A.W.B. additionally reports Foresite ownership interest. The other authors declare no competing interests.

**Additional information**

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0652-7.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0652-7.

Correspondence and requests for materials should be addressed to P.R. or J.S.R.-F.

Peer review information Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.
**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 x IQR on either side.
Extended Data Fig. 3 | see figure caption on next page.
**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 \( \mu \) (the rate of events), \( \theta \) (the type of event), \( \alpha, \beta \) 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 \( \lambda_p (\mu_p, d_p) \) were summarized by their mean \( \mu_p \) 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 \( \mu_p \) against the observed \( \lambda_p \) for samples in the training set. Note the data points at the bottom are all positions \( p \) with non-zero mean posterior \( \mu_p \) 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 \( \mu_p \) 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 | see figure caption on next page.
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 | see figure caption on next page.
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.
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 \( AD^0 = AD + 2 \) and \( DP^0 = DP + 4 \). 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 | see figure caption on next page.
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) Log2 ratios of tumor biopsy and 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 extend 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 Log2 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 OncodriveCLUST63 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) Log2 ratios estimated from the cfDNA of (a) n = 24 female and (b) n = 23 male control individuals. For each individual, the raw Log2 ratios were smoothed using a cubic spline. The two panels show the superimposed splines for all control samples according to gender. (c–e) Log2 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 Log2 ratios were used to compute the Pearson correlation coefficient comparing segments overlapping >75% in the tumor biopsies and cfDNA samples. In (a–e), the Log2 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 x 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 \( \text{ERBB2} \) on chromosome 17q are shown together with one NSCLC patient, MSK-VL-0044 with a reported \( \text{MET} \) 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 \( \text{ERBB2} \) and \( \text{MET} \) are displayed by orange arrows and labelled accordingly.
**Reporting Summary**

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size ($n$) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. $F$, $t$, $r$) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted
  - Give $P$ values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s $d$, Pearson’s $r$), indicating how they were calculated

Our web collection on **statistics for biologists** contains articles on many of the points above.

### Software and code

Policy information about availability of computer code

**Data collection**

**Tumor Tissue sequencing:**

Tumor sequencing was performed utilizing the FDA-authorized MSK-IMPACT assay (Zehir et al, Nat Med 2017). Briefly, tumor DNA was extracted from formalin-fixed paraffin-embedded (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, as previously described (Zehir et al, Nat Med 2017, Cheng et al, J Mol Diagn 2015). Average sequencing coverage across all tumors was greater than 900X. Somatic mutations, DNA copy number alterations, and structural rearrangements were identified as previously described (Cheng et al, J Mol Diagn 2015) and all mutations were manually reviewed.

**cfDNA sequencing:**

Whole blood sample collection, accessioning, and preparation: Peripheral blood from patients with metastatic cancer was collected into two 10 mL Cell-Free DNA BCT (Streck; La Vista, NE) at Memorial Sloan Kettering Cancer Center (New York, NY) and shipped to GRAIL, Inc. (Menlo Park, CA) at room temperature. Whole blood from healthy individuals drawn into Streck BCTs were purchased from 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 QI/Amp 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 DNAEasy 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: Library preparation, target enrichment, and sequencing Buffycf coat gDNA (50ng) and plasma cfDNA (≤75ng) were used for NGS library construction with a modified Illumina TruSeq DNA Nano
The adapter includes 96 (cancer samples) or 218 (healthy controls) unique molecular identifier (UMI) sequences, of 6-mer in length, used to suppress technical assay errors. Amplified libraries were cleaned up using magnetic beads and quantified using the Fragment Analyzer Standard Sensitivity NGS kit.

Quantified libraries underwent hybridization-based capture with a research cancer panel targeting 508 cancer-related genes (2.13 Mb; GRAIL, Inc.; Menlo Park, CA). The panel included full exons except for the telomerase reverse transcriptase (TERT) gene, which only included promoter regions. Additional intronic regions were included for rearrangement detection of 28 genes and copy number aberration detection of 42 genes. Up to 2ug of genomic DNA (gDNA) libraries were used for target enrichment with a modified Illumina Nextera Rapid Capture protocol. For cfDNA samples, up to 4ug of libraries (two parallel enrichment reactions) were used to maximize capture efficiencies. The two enriched libraries from the same cfDNA library were pooled and quantified using a Fragment Analyzer Standard Sensitivity NGS kit.

Three libraries per flowcell or six libraries across two flowcells were clustered (pooled and loaded across all eight lanes of each flowcell) and sequenced on a HiSeq X (Illumina; San Diego, CA) at a nominal raw target coverage of 60,000X (median collapsed target coverage [unique molecule counts] was ~4,400X). Read lengths were set to 150, 150, 8, and 8, respectively for read 1, read 2, index read 1, and index read 2. The cfDNA analysis pipeline is described in detail in Supplementary Methods.

Data analysis

All statistical analyses were performed using the R environment for statistical computing. A detailed description of the pipeline and assumptions made for the error correction steps is provided in the Methods and Supplementary Methods.

Additionally, the following softwares were used to acquire data:

1/ Aquarius iNtuition advanced visualization software, version 4.4.13.P3 (TeraRecon, Inc, Foster City, CA)
2/ aBSI platform version 3.3 (EXINI Diagnostics AB, Lund, Sweden)

The following softwares were used to process and analyze data:

1/ bcl2fastq version 2.18.0.6
2/ BWA MEM 0.7.12-r1039
3/ Samtools 1.3.1 (using htslib 1.3.1)
4/ FACETS 0.5.2
5/ CNVkit 0.9.2.dev0
6/ MSIsensor v0.5
7/ R 3.5.1
8/ Bioconductor version 3.8
9/ deconstructSigs 1.8.0
10/ clinfun 1.0.15
11/ mblm 0.12.1
12/ pscl 1.5.2
13/ copynumber 1.22.0
14/ maftools 1.8.0
15/ pROC 1.12.1
16/ Imtest 0.9-36
17/ glmperm 1.0-5
18/ Absolute 1.0.6

The MSK-IMPACT pipeline used to process and tumor biopsy and matched WBC sequencing together with the full list of dependencies and version information are available from https://github.com/rhshah/IMPACT-Pipeline

Custom codes written to generate the figures from the processed data are available from https://github.com/ndbrown6/MSK-GRAIL-TECHVAL

The full list of dependencies and version information are available from https://github.com/ndbrown6/modules/blob/master/conda/jrflab-modules-0.1.7.txt

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [x] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences
Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

**Sample size**  
Sample size was not pre-determined, given that this study focused on a method without any equivalent method available. Hence, sample size assumptions were not possible. All available specimens were utilized. 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.

**Data exclusions**  
A consort diagram explaining the detailed sample exclusions is included as Extended Data Fig.1 and the exclusion were described in the methods.

Of the 232 cancer patients who were enrolled, 71 patients were excluded for not fulfilling the initial clinical or laboratory including criteria:
- No progression or pseudoprogression: n=6
- Withdrew consent: n=2
- Wrong cohort assignment: n=3
- Blood collection error: n=16
- Blood sample not meeting inclusion criteria (outside time window, after start of new treatment, etc): n=26
- Tissue not collected or insufficient: n=4
- Tissue sequencing unavailable prior to inclusion: n=14

Of the 161 patients who met the initial inclusion criteria, 27 patients were excluded due to tumor tissue sequencing failure (n=18) or insufficient tumor tissue (i.e. specimen tumor content not meeting the MSK-IMPACT assay’s minimum requirements, n=9). Additionally, 13 patients were excluded for cfDNA assay failure (n=9) or cfDNA assay QC failure (n=4).

The final analysis included 124 patients with both cfDNA and tumor tissue sequencing results.

Of the 50 non-cancer controls included in the study 3 were excluded due to cfDNA assay failure. A total of 47 controls were include in the final analysis.

**Replication**  
Technical and biological replicates were performed and are described in the text and Methods. All attempts at replication were successful. Sequencing data reproducibility for the cfDNA assay is described in the text and the results are provided in Fig.1 and Supplementary Figs. 4 and 15.

**Randomization**  
This is a prospective observational study with the main aim of developing and validating a cfDNA sequencing assay. All the eligible cancer patients underwent paired tumor and cfDNA samples sequencing and the the eligible non-cancer controls underwent cfDNA sequencing. No randomization was required for this observational study. The groups (MBCs, CRPCs and NSCLCs) are groups defined on the basis of the disease type, whereas the healthy donors group was defined on the basis of absence of these diseases.

**Blinding**  
Blinding was not relevant for this prospective observational study. Mutation and copy number detection in plasma cfDNA was performed de novo, without a priori knowledge of the tumor-derived variants. After the detection of somatic genetic alterations in plasma cfDNA, the tumor-derived somatic mutations were unblinded for the comparisons performed.

**Reporting for specific materials, systems and methods**

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒  | Antibodies            |
| ☒  | Eukaryotic cell lines |
| ☒  | Palaeontology         |
| ☒  | Animals and other organisms |
| ☒  | Human research participants |
| ☒  | Clinical data         |

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒  | ChIP-seq              |
| ☒  | Flow cytometry        |
| ☒  | MRI-based neuroimaging |

### Eukaryotic cell lines

Policy information about [cell lines](#)  

**Cell line source(s)**  
NA12878

**Authentication**  
Coriell Institute (Camden, NJ)
Human research participants

Policy information about studies involving human research participants

Population characteristics

The baseline demographic characteristics of the cohort included in this study were consistent with those of a contemporary prospectively collected cohort of consecutive cases of these malignancies.

Among evaluable patients in the metastatic breast cancer (MBC) cohort, the median age was 60 (range 30-79), 26 (67%) were hormone receptor-positive and HER2-negative, and 32 (82%) had invasive ductal carcinoma. In the evaluable metastatic non-small cell lung cancer (NSCLC) cohort, the median age was 67 (range 33-83), 28 (68%) were female, 38 (93%) were adenocarcinomas, and 28 (68.3%) had M1b (extrathoracic metastases) disease. The median age of evaluable castration resistance prostate cancer (CRPC) patients was 67 (range 46-87), and 39 (89%) were adenocarcinomas. Overall, the majority of patients enrolled in the study received prior treatment in the (neo)adjuvant and/or metastatic settings (85% in MBC, 41% in NSCLC, 100% in CRPC): 38% of the MBC patients and 10% of the NSCLC patients had received at least three lines of therapy in the metastatic setting, and all CRPC patients had disease progression after initial castration therapy with or without androgen receptor antagonists, with 32 (78%) also receiving additional systemic therapy prior to sample collection.

Recruitment

The recruitment strategy is described in the Methods section under "Patient enrollment" section.

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

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

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.

Ethics oversight

All patients provided written informed consent for tumor, cfDNA, and WBC sequencing and review of patient medical records for detailed demographic, pathologic, and treatment information under an IRB-approved biospecimen umbrella protocol (MSKCC protocol 12-245, clinicaltrials.gov ID: NCT01775072).

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Study was performed under an IRB-approved biospecimen umbrella protocol (MSKCC protocol 12-245, clinicaltrials.gov ID: NCT01775072).

Study protocol

NCT01775072, https://clinicaltrials.gov/ct2/show/NCT01775072

Data collection

Detailed clinical data including baseline demographics, cancer history, and prior lines of therapy were collected from MSKCC medical records for all the cancer patients at the time of enrollment in the study (between September 24, 2015 and August 01, 2016).

Limited clinical data including age and gender 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.
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.

Outcomes

No clinical outcome analysis was planned for this study, whose primary aims 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 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. Therefore, outcome data were not collected for the cohort.