

# EXHIBIT J



# Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA

David M. Kurtz<sup>1,2,16</sup>, Joanne Soo<sup>1,16</sup>, Lyron Co Ting Keh<sup>1</sup>, Stefan Alig<sup>1</sup>, Jacob J. Chabon<sup>1,2,3,14</sup>, Brian J. Sworder<sup>1</sup>, Andre Schultz<sup>1,2</sup>, Michael C. Jin<sup>1</sup>, Florian Scherer<sup>1,15</sup>, Andrea Garofalo<sup>1</sup>, Charles W. Macaulay<sup>1</sup>, Emily G. Hamilton<sup>4</sup>, Binbin Chen<sup>1,5</sup>, Mari Olsen<sup>1</sup>, Joseph G. Schroers-Martin<sup>1,6</sup>, Alexander F. M. Craig<sup>1</sup>, Everett J. Moding<sup>1,7</sup>, Mohammad S. Esfahani<sup>1</sup>, Chih Long Liu<sup>1</sup>, Ulrich Dührsen<sup>1,8</sup>, Andreas Hüttmann<sup>1,8</sup>, René-Olivier Casanovas<sup>1,9</sup>, Jason R. Westin<sup>1,10</sup>, Mark Roschewski<sup>1,11</sup>, Wyndham H. Wilson<sup>1,11</sup>, Gianluca Gaidano<sup>1,12</sup>, Davide Rossi<sup>1,13</sup>, Maximilian Diehn<sup>1,2,3,7</sup>✉ and Ash A. Alizadeh<sup>1,2,3,6</sup>✉

**Circulating tumor-derived DNA (ctDNA) is an emerging biomarker for many cancers, but the limited sensitivity of current detection methods reduces its utility for diagnosing minimal residual disease. Here we describe phased variant enrichment and detection sequencing (PhasED-seq), a method that uses multiple somatic mutations in individual DNA fragments to improve the sensitivity of ctDNA detection. Leveraging whole-genome sequences from 2,538 tumors, we identify phased variants and their associations with mutational signatures. We show that even without molecular barcodes, the limits of detection of PhasED-seq outperform prior methods, including duplex barcoding, allowing ctDNA detection in the ppm range in participant samples. We profiled 678 specimens from 213 participants with B cell lymphomas, including serial cell-free DNA samples before and during therapy for diffuse large B cell lymphoma. In participants with undetectable ctDNA after two cycles of therapy using a next-generation sequencing-based approach termed cancer personalized profiling by deep sequencing, an additional 25% have ctDNA detectable by PhasED-seq and have worse outcomes. Finally, we demonstrate the application of PhasED-seq to solid tumors.**

**A**nalysis of somatic mutations from circulating cell-free DNA (cfDNA) has broad applications for detecting and characterizing cancer<sup>1–10</sup>. Detection of ctDNA has been shown in principle in most cancer types<sup>5</sup>. However, in practice, the current limits of detection (LOD) for ctDNA analysis are imperfect, and are inherently restricted by two key factors: low-input DNA amounts from a typical blood collection and background error rates of current sequencing technologies<sup>11</sup>.

Recent approaches have improved ctDNA minimal residual disease (MRD) performance by tracking multiple somatic mutations with error-suppressed sequencing<sup>2,12,13</sup>. This approach has allowed detection limits as low as 2–10 parts in 100,000 from limited cfDNA using off-the-shelf panels or personalized assays<sup>2,11</sup>. Detection of residual disease during or after treatment is a powerful prognostic tool, with detectable MRD representing an adverse sign even during radiographic remission<sup>14,15</sup>. However, current detection limits

are insufficient to universally detect residual disease in patients destined for disease relapse. This ‘loss of detection’ is exemplified in diffuse large B cell lymphoma (DLBCL). Here a threshold-response in ctDNA after only two cycles of curative-intent therapy is strongly prognostic for favorable clinical outcomes (major molecular response (MMR))<sup>14</sup>. Despite this, nearly one-third of patients experiencing eventual disease progression do not have detectable ctDNA at this interim landmark when using either immunoglobulin or targeted sequencing<sup>14,15</sup>, thus representing ‘false-negative’ measurements. Such high false-negative rates have also been observed using alternative methodologies to track disease in colon<sup>8</sup> and breast cancers<sup>16</sup>. More sensitive methods are therefore needed to improve residual disease detection across cancers.

Prior methods aimed at lowering the LOD have focused on somatic variants detected on both of the complementary strands of parental DNA duplexes (Extended Data Fig. 1a)<sup>2,17,18</sup>. ‘Duplex

<sup>1</sup>Division of Oncology, Department of Medicine, Stanford University, Stanford, CA, USA. <sup>2</sup>Stanford Cancer Institute, Stanford University, Stanford, CA, USA. <sup>3</sup>Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA. <sup>4</sup>Program in Cancer Biology, Stanford University, Stanford, CA, USA. <sup>5</sup>Department of Genetics, Stanford University, Stanford, CA, USA. <sup>6</sup>Division of Hematology, Department of Medicine, Stanford University, Stanford, CA, USA. <sup>7</sup>Department of Radiation Oncology, Stanford University, Stanford, CA, USA. <sup>8</sup>Department of Hematology and Stem Cell Transplantation, West German Cancer Center Essen, University Hospital Essen, Essen, Germany. <sup>9</sup>Department of Hematology, Hôpital F. Mitterrand, CHU Dijon and INSERM, Dijon, France. <sup>10</sup>Department of Lymphoma/Myeloma, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. <sup>11</sup>Lymphoid Malignancies Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. <sup>12</sup>Division of Hematology, Department of Translational Medicine, University of Eastern Piedmont, Novara, Italy. <sup>13</sup>Hematology, Oncology Institute of Southern Switzerland and Institute of Oncology Research, Bellinzona, Switzerland. <sup>14</sup>Present address: Foresight Diagnostics, Aurora, CO, USA.

<sup>15</sup>Present address: Department of Medicine I, Medical Center–University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany.

<sup>16</sup>These authors contributed equally: David M. Kurtz, Joanne Soo. ✉e-mail: [diehn@stanford.edu](mailto:diehn@stanford.edu); [arasha@stanford.edu](mailto:arasha@stanford.edu)

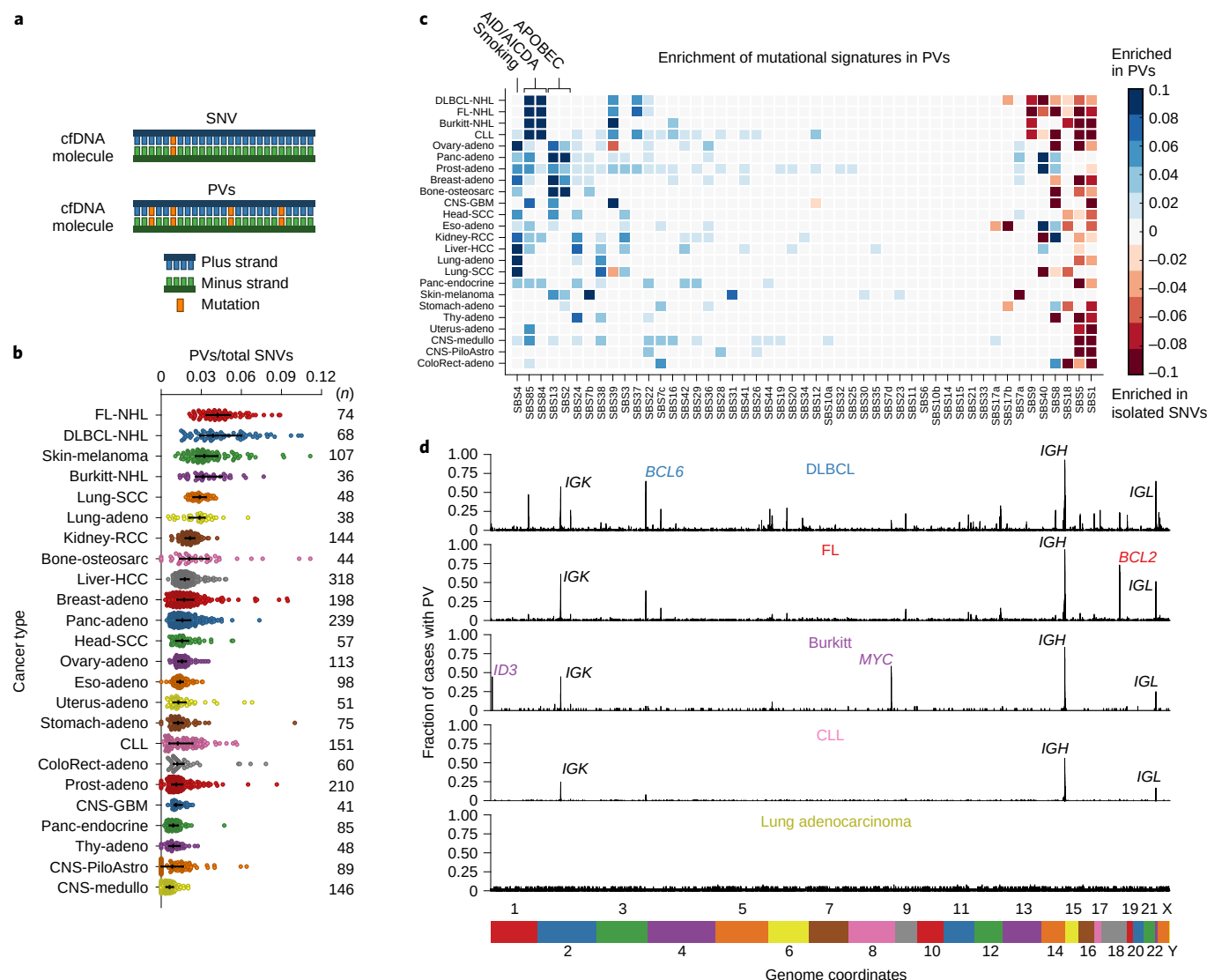


**Fig. 1 | Discovery of phased variants and their mutational signatures via analysis of whole-genome sequencing data. a**, Cartoon depicting the difference between an SNV (top) and multiple variants 'in-phase' (PVs; bottom) on DNA molecules. In theory, PVs are more specific events than isolated SNVs. **b**, Scatterplot showing the distribution of PVs from WGS data for 24 different cancer histologies, normalized by total SNVs. Bars represent the median and interquartile range (IQR). FL-NHL, follicular lymphoma; DLBCL-NHL, diffuse large B cell lymphoma; Burkitt-NHL, Burkitt lymphoma; lung-SCC, squamous cell lung cancer; lung-adenocarcinoma, lung adenocarcinoma; kidney-RCC, renal cell carcinoma; bone-osteosarc, osteosarcoma; liver-HCC, hepatocellular carcinoma; breast-adenocarcinoma, breast adenocarcinoma; panc-adenocarcinoma, pancreatic adenocarcinoma; head-SCC, head and neck squamous cell carcinoma; ovary-adenocarcinoma, ovarian adenocarcinoma; eso-adenocarcinoma, esophageal adenocarcinoma; uterus-adenocarcinoma, uterine adenocarcinoma; stomach-adenocarcinoma, stomach adenocarcinoma; colorect-adenocarcinoma, colorectal adenocarcinoma; prost-adenocarcinoma, prostate adenocarcinoma; CNS-GBM, glioblastoma multiforme; panc-endocrine, pancreatic neuroendocrine tumor; thy-adenocarcinoma, thyroid adenocarcinoma; CNS-piloastro, pilastrocytoma; CNS-medullo, medulloblastoma. **c**, Heat map demonstrating enrichment in SBS mutational signatures for PVs versus single SNVs across cancer types. Blue represents signatures enriched in PVs; red represents signatures where unphased, single SNVs are enriched. Only signatures with a significant difference between PVs and unphased SNVs after correcting for multiple hypotheses are shown; other signatures are in gray (Methods). See <https://phasedseq.stanford.edu/> for additional details. **d**, Bar plots showing the distribution of PVs in stereotyped regions across the genome in B lymphoid malignancies and lung adenocarcinoma. The genome was divided into 1,000-bp bins, and the fraction of samples with a PV in each bin was calculated. Only bins that had at least a 2% recurrence frequency in any cancer subtype are shown.

sequencing' reduces the background error rate due to the requirement of two concordant events for detection of a single-nucleotide variant (SNV). In previous studies using duplex sequencing for ctDNA analysis, an analytical LOD as low as 1 molecule in 400,000 was achievable<sup>2</sup>. However, this approach is limited by inefficient recovery of DNA duplexes—recovery of both original strands occurs in a minority (typically 20–25%)<sup>2,17</sup> of all recovered molecules. This inefficiency makes duplex sequencing suboptimal for real-world ctDNA detection, where input DNA from practical blood volumes

is limited (~4,000–8,000 genomes per standard 10-ml blood collection tube) and maximal recovery of genomes is essential<sup>19</sup>. To improve MRD detection, methods that simultaneously achieve low analytical detection limits and high molecular recovery for multiple mutations are still needed (Extended Data Fig. 1b).

Here we introduce an alternative to duplex sequencing for reducing the background error rate that involves detection of 'phased variants' (PVs), where two or more mutations occur in *cis* (that is, on the same strand of DNA; Fig. 1a and Extended Data Fig. 1). Similarly

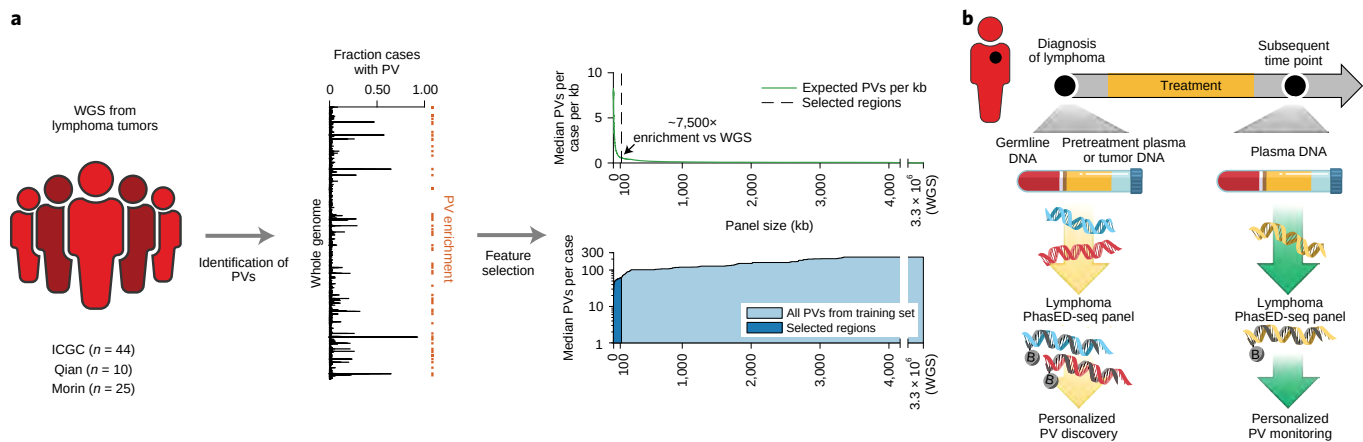


**Fig. 2 | Design of phased variant enrichment sequencing. a**, Cartoon showing the design of PhasED-seq. WGS data from DLBCL tumors were aggregated (left), and recurrent PVs were identified (middle). A panel capturing regions recurrently containing PVs was designed (right). Top right, the expected number of PVs per kb for increasing panel sizes. The dashed line shows the selected regions. Bottom right, the median total expected PVs per case for increasing panel sizes. The dark area shows the selected regions. **b**, Schematic for the use of PhasED-seq in participants with B cell malignancies. At the time of diagnosis, a tumor or plasma sample, along with matched germ line, is sequenced to identify a personalized set of PVs. These PVs can then be tracked in future plasma samples.

to duplex sequencing, this method provides lower error profiles due to the concordant detection of two separate non-reference events in individual molecules. However, unlike duplex sequencing, both events occur on the same sequencing read pair, thereby increasing the efficiency of genome recovery. We find that phased mutations are present in diverse cancer types, but occur in stereotyped portions of the genome in B cell malignancies, likely due to physiological and aberrant somatic hypermutation (aSHM)<sup>20,21</sup>. We therefore develop and apply PhasED-seq, a new method to detect ctDNA through PVs to tumor fractions on the order of ppm. We demonstrate that PhasED-seq can meaningfully improve detection of ctDNA in clinical samples both during therapy and before disease relapse. Finally, we apply PhasED-seq to participants with solid tumors through personalized identification of PVs, demonstrating the broad applicability of PhasED-seq for improved MRD detection.

## Results

**Genomic distribution of phased variants.** To identify malignancies where PVs could potentially improve disease detection, we assessed the frequency of PVs across cancer types. We analyzed publicly available whole-genome sequencing (WGS) data to identify sets of variants occurring at a distance of <170 bp apart, which represents the typical length of a single cfDNA fragment<sup>22</sup> consisting of a single core nucleosome and associated linker. We identified and summarized the frequency of these ‘putative phased variants,’ (Methods) controlling for the total number of SNVs, from 2,538 tumors across 24 cancer histologies including solid tumors and hematological malignancies (Fig. 1b, Extended Data Fig. 2 and Supplementary Table 1). Interestingly, PVs were most significantly enriched in two B cell lymphomas (DLBCL and follicular lymphoma (FL);  $P < 0.05$  versus all other histologies), a group of diseases with known hypermutation driven by AID (AICDA).

**Mutational mechanisms underlying phased variants.** To investigate the origin of PVs, we compared the single-base substitution (SBS) mutational signatures<sup>23</sup> contributing to SNVs occurring within 170 bp of another SNV, and SNVs occurring in isolation (Methods). As expected, PVs were highly enriched in several mutational signatures associated with clustered mutations<sup>24</sup>. Signatures of clustered mutations associated with activity of AID (SBS84 and SBS85) were significantly enriched in PVs from B cell lymphomas

and chronic lymphocytic leukemia (CLL), while signatures associated with activity of APOBEC3B (SBS2 and SBS13)—another known mechanism of kataegis hypermutation—were significantly enriched in PVs from multiple solid cancer histologies (Fig. 1c)<sup>24,25</sup>. Interestingly, PVs from multiple tumor types were also associated with SBS4, a signature associated with tobacco use. In contrast, aging-associated mutational signatures such as SBS1 and SBS5 were significantly enriched in isolated SNVs<sup>26</sup>.

**Phased variants occur in stereotyped genomic regions in lymphoid cancers.** To assess the genomic distribution of putative PVs, we binned these events into 1-kb regions across the genome. We observed a strikingly stereotyped distribution of PVs in individual lymphoid neoplasms (that is, DLBCL, FL, Burkitt lymphoma (BL) and CLL; Fig. 1d and Extended Data Fig. 3). In contrast, nonlymphoid cancers did not exhibit substantial recurrence of PVs in stereotyped regions. This lack of stereotype in PV-containing loci was true even when considering melanomas and lung cancers, diseases with frequent PVs.

Across lymphomas, the majority of hypermutated regions were shared between lymphoma subtypes, with the highest densities in known targets of SHM<sup>27,28</sup> including *BCL2*, *BCL6* and *MYC*, as well as the immunoglobulin loci *IGH*, *IGK* and *IGL* (Supplementary Table 2 and Fig. 1d), with DLBCL harboring the most recurrently affected regions (Supplementary Fig. 1a). Among 1,639 1-kb regions recurrently containing PVs, nearly one-third fell into genomic areas previously associated with physiological or aberrant SHM in B cells. Specifically, 19% (315/1,639) were located in immunoglobulin regions, while 13% (218/1,639) were in portions of 68 previously identified targets of aSHM (Supplementary Table 2). While most PVs fell into noncoding regions of the genome, we also identified additional recurrently affected loci not previously described as targets of aSHM, including *XBPI*, *LPP* and *AICDA*, among others.

Interestingly, the distribution of PVs within each lymphoma correlated with known oncogenic features associated with the corresponding disease, including PVs in *BCL2* in FL<sup>29</sup> and PVs in *MYC* and *ID3* in BL<sup>30</sup> (Fig. 1d and Supplementary Fig. 1b–d). DLBCL molecular subtypes associated with distinct cell of origin<sup>31</sup> also demonstrated distinct distributions of PVs (Supplementary Table 2), including significant enrichments in PVs in *BCL2* in germinal center B cell-like (GCB) DLBCLs and *IGH* class-switch regions

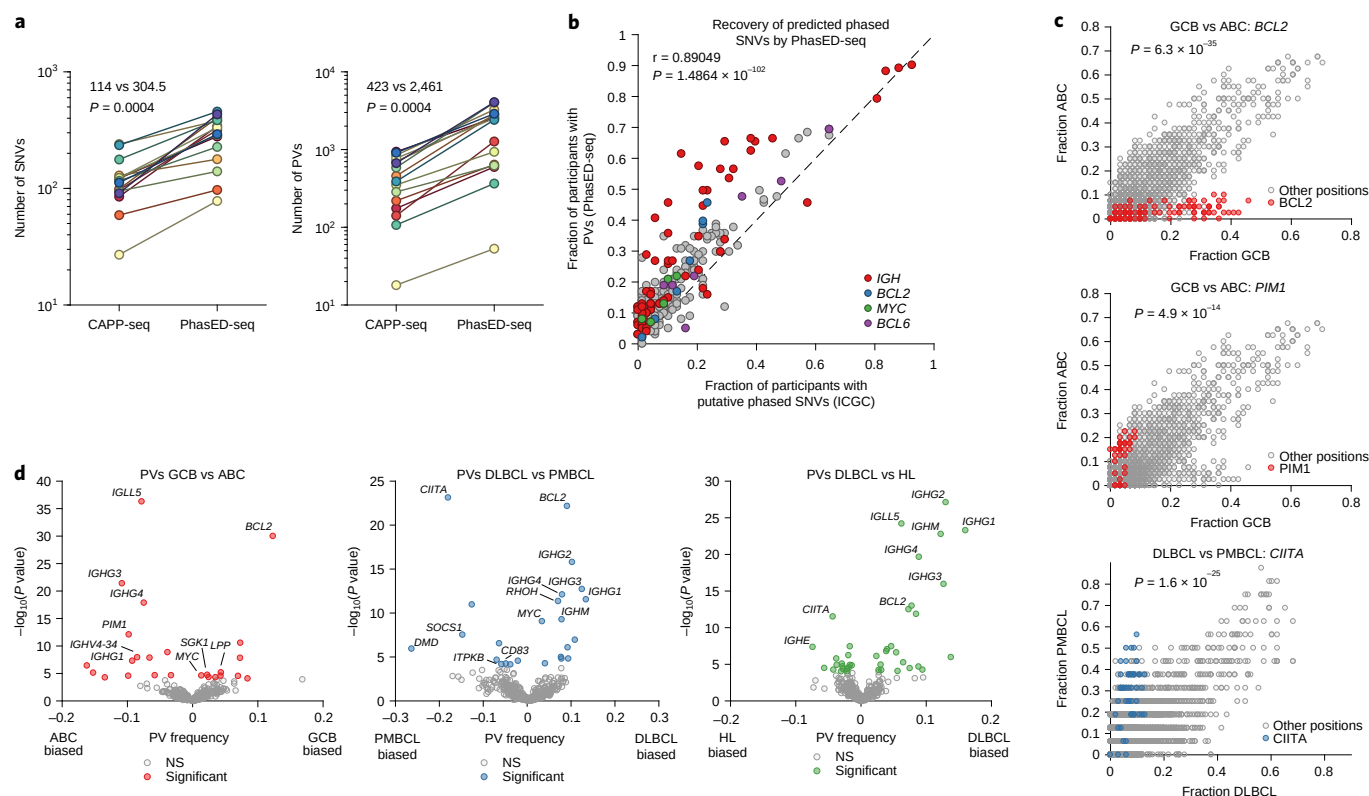


**Fig. 3 | Validation and application of phased variant enrichment sequencing.** **a**, Comparison of SNV (left) and PV (right) yield for sequencing tumor and/or cfDNA and matched germ line by a previously established lymphoma CAPP-seq panel or PhasED-seq (two-sided Wilcoxon signed-rank test). PVs included doublet, triplet and quadruplet phased events. **b**, Scatterplot showing the frequency and Pearson correlation of PVs in 1,000-bp bins for participants with DLBCL, identified by either WGS or PhasED-seq. **c**, Scatterplots comparing the frequency of PVs by location (in 50-bp bins) for subtypes of lymphoma. The colored circles show the frequency of PVs in 50-bp bins from a gene of interest; the other (gray) circles show the frequency of PVs in 50-bp bins from the remainder of the PhasED-seq sequencing panel. Statistical analysis was performed by two-sided Wilcoxon rank-sum test of all 50-bp bins in a given gene against all other bins (Methods). See <https://phasedseq.stanford.edu/> for additional details. **d**, Volcano plots summarizing the difference in relative frequency of PVs in specific loci between types of lymphoma, including ABC-DLBCL versus GCB-DLBCL (red, left); PMBCL versus DLBCL (blue, middle); and HL versus DLBCL (green, right; Methods). NS, not significant.

(Sy1 and Sy3) in activated B cell-like (ABC) DLBCLs, consistent with previous reports<sup>32</sup> (Supplementary Fig. 1e).

**Design and validation of PhasED-seq panel for lymphoma.** To validate these PV-rich regions and assess their utility for disease detection from ctDNA, we designed a sequencing panel targeting putative PVs identified within WGS from three independent cohorts of individuals with DLBCL, as well as in individuals with CLL<sup>33–36</sup> (Fig. 2a and Methods). This final PhasED-seq panel targeted ~115 kb of genomic space focused on PVs, along with an additional ~200 kb targeting genes known to be recurrently mutated in B cell non-Hodgkin lymphomas (B-NHLs); this single panel was used for both identification of PVs from tumor and/or plasma samples and tracking residual disease (Fig. 2b). While the 115 kb of space dedicated to PV capture targets only 0.0035% of the human genome, it captures 26% of PVs observed by WGS (Extended Data Fig. 4a), yielding a ~7,500-fold enrichment over WGS.

We next compared expected SNV and PV recovery to a previously reported 228-kb lymphoma ctDNA panel designed to maximize SNVs for each participant from WES data (that is, cancer personalized profiling by deep sequencing (CAPP-seq)) in B cell lymphomas (Extended Data Fig. 4a–c)<sup>3</sup>. Compared to this prior panel, PhasED-seq covered 3 times more SNVs and 2.9 times more PVs in available WGS data<sup>33–35</sup>. To validate these yield improvements, we profiled 16 pretreatment tumor or plasma DNA samples

from participants with DLBCL (Supplementary Table 3). We applied both panels to each specimen and sequenced to high unique molecular depths (Fig. 3a and Extended Data Fig. 4d,e). Compared to the expected enrichment, we observed similar improvements in yield of SNVs by PhasED-seq compared to the prior panel (2.7 times; median 304.5 versus 114). However, we found a larger improvement in the number of PVs recovered by PhasED-seq (5.8 times; median 2,461 versus 423 PVs per participant; Fig. 3a). We also observed a robust correlation between the frequency of PVs in WGS and targeted data across 101 DLBCL samples (Fig. 3b).

**Differences in phased variants between lymphoma subtypes.** Having validated the PhasED-seq panel, we sought to assess the biological differences in PVs between various B cell malignancies, including DLBCL ( $n=101$ ), primary mediastinal B cell lymphoma (PMBCL;  $n=16$ ), FL ( $n=13$ ), classical Hodgkin lymphoma ( $n=23$ ) and mantle cell lymphoma (MCL;  $n=13$ ; Supplementary Table 4). We evaluated the quantity of SNVs and PVs in each of these lymphomas, as well as the number of ‘independent tumor reporters’ for possible disease monitoring from cfDNA, defined as PVs separated into 150-bp islands across our panel (Extended Data Fig. 5f–j and Methods). The number of SNVs identified per sample was not significantly different between lymphoma subtypes, with the exception of fewer SNVs in MCL. However, DLBCL, PMBCL and FL had a higher burden of PVs than classical Hodgkin lymphoma or

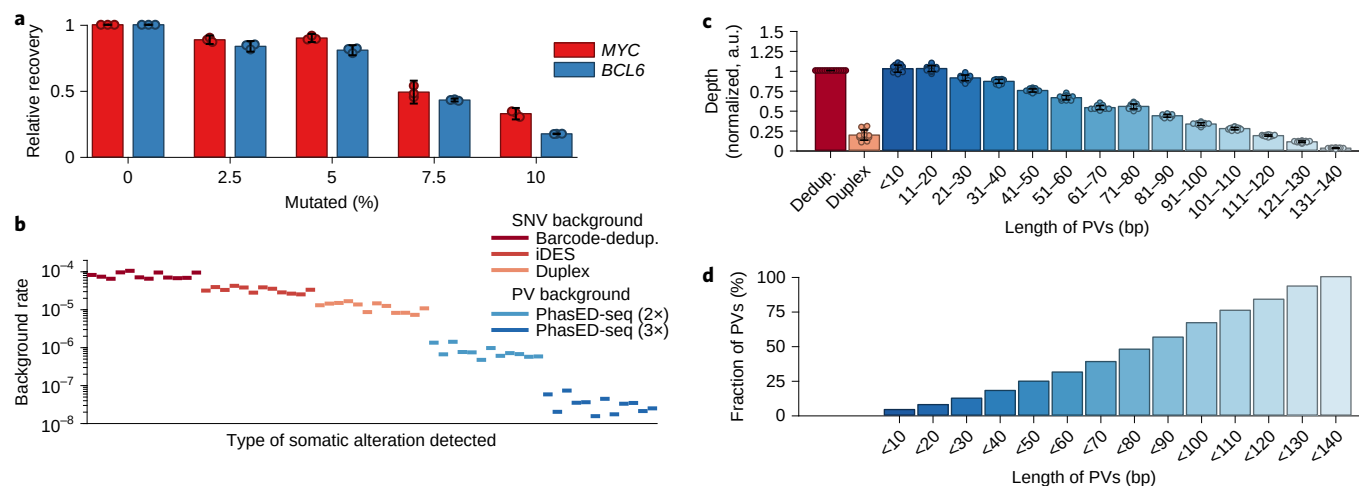


**Fig. 4 | Technical performance of PhasED-seq.** **a**, Bar plot showing the performance of hybrid-capture sequencing across  $n=3$  replicates for recovery of synthetic 150-bp oligonucleotides from two loci (*MYC* and *BCL6*; Supplementary Table 6) with increasing degree of mutation. Data are presented as the mean  $\pm$  95% confidence interval normalized to the unmutated condition. **b**, Plot demonstrating the background rate (Methods) for different sequencing methods from 12 control cfDNA samples. **c**, Bar plot showing the unique molecular depth of sequencing from  $n=12$  independent cfDNA samples for single-stranded and duplex deduplication, and recovery of PVs of increasing distance between SNVs in-phase. Data are presented as the mean  $\pm$  s.d. values. a.u., arbitrary units. **d**, Bar plot showing the cumulative fraction of PVs with a maximal distance between SNVs of less than a given number of base pairs.

MCL, consistent with MCL having a lower activity of AICDA. We also observed differences in the genomic locations of PVs between lymphoma subtypes (Fig. 3c,d and Supplementary Fig. 2), including frequent PVs in *BCL2* in GCB-DLBCL, *PIM1* PVs in ABC-DLBCL and *CIITA* PVs in PMBCL<sup>28,37,38</sup>. Additional loci enriched in specific lymphoma types are provided in Supplementary Table 5.

**Recovery of phased variants through PhasED-seq.** To facilitate detection of ctDNA using PVs, efficient recovery of DNA molecules is required. Hybrid-capture sequencing is potentially sensitive to DNA mismatches, and AID hotspots can contain a 5–10% or higher local mutation rate<sup>39</sup>. Using in silico simulation of 150-mers, we observed the expected decrease in predicted binding energy with an increasing number of mutations (Extended Data Fig. 5a). We assessed the effect on hybrid-capture sequencing efficiency using synthetic 150-bp DNA oligonucleotides with 0 to 10% difference from the reference sequence in *MYC* and *BCL6*, known targets of aSHM (Supplementary Table 6 and Methods). When subjecting an equimolar mixture of these oligonucleotides to capture and sequencing, molecules with as high as a 5% mutation rate were captured with nearly the same efficiency as their wild-type counterparts (85% versus 100%), while molecules with a 10% mutation rate were captured with only 27% relative efficiency (Fig. 4a). Notably, only 7% of cases had any region exceeding 10% mutation frequency across the panel (Methods and Extended Data Fig. 5b,c), and in all cases, the 90th percentile mutation rate was <5%, suggesting the majority of phased mutations are recoverable by hybrid capture.

**Background profile and limit of detection of PhasED-seq.** Previous methods for error-suppressed sequencing of cfDNA have used either a combination of molecular and in silico methods for error suppression, for example, integrated digital error suppression (iDES)<sup>3</sup>, or duplex sequencing to reduce background signal<sup>17</sup>. However, each of these methods has limitations, either for detecting ultra-low tumor fractions or for efficient recovery of original DNA molecules. We compared the background profile and recovery of input genomes from plasma cfDNA samples from 12 healthy adults by PhasED-seq with both iDES-enhanced CAPP-seq and duplex sequencing. As expected, while iDES-enhanced CAPP-seq had a

lower background profile than barcode deduplication alone, duplex sequencing offered even lower background signal for non-reference single-nucleotide substitutions (Fig. 4b;  $3.3 \times 10^{-5}$  versus  $1.2 \times 10^{-5}$ ,  $P < 0.0001$ ). However, the rate of phased errors—that is, multiple non-reference bases occurring on the same sequencing fragment—was significantly lower than the rate of single errors in either iDES-enhanced CAPP-seq or duplex sequencing data. This was true for the incidence of both two (2 $\times$  or ‘doublet’ PVs) and three (3 $\times$  or ‘triplet’ PVs) substitutions on the same DNA molecule (Fig. 4b;  $8.0 \times 10^{-7}$  and  $3.4 \times 10^{-8}$ , respectively,  $P < 0.0001$ ), even without the use of barcode-mediated error suppression (Extended Data Fig. 5d). The background rate of PVs was also lower than that of small insertions and deletions (Extended Data Fig. 5d). Phased errors containing C-to-T or T-to-C transitions were more common than other types of PVs (Extended Data Fig. 5e). The rate of background PVs also decreased with increasing distance between positions (Extended Data Fig. 5f). When considering unique molecular depth, duplex sequencing recovered only 19% of all unique cfDNA fragments (Fig. 4c). In contrast, the unique depth of reads covering PVs within a genomic distance of <20 bp was nearly identical to the overall sample depth. Similarly, PVs of up to 80 bp in size had a depth greater than 50% of the median unique depth for a sample. Importantly, almost half (48%) of all PVs were less than 80 bp in length (Fig. 4d).

To compare PhasED-seq to alternative methods for ctDNA detection, we generated limiting dilutions of ctDNA from three participants with lymphoma into healthy control cfDNA, resulting in expected tumor fractions between 0.1% and 0.00005% (1:2,000,000; Methods and Supplementary Table 7). We compared the expected and observed tumor fractions in each of these dilutions using PhasED-seq and SNV-based methods (that is, CAPP-seq or duplex sequencing; Fig. 5a). All methods performed equally well down to tumor fractions of 0.01% (1 part in 10,000); however, below this level (for example, 0.001%, 0.0002%, 0.0001% and 0.00005%), both PhasED-seq and duplex sequencing significantly outperformed single-strand unique molecular identifier (UMI)-based SNV detection ( $P < 0.0001$  for duplex, ‘2 $\times$ ’ PhasED-seq and ‘3 $\times$ ’ PhasED-seq; Fig. 5a). In addition, when compared to detection of SNVs from duplex molecules, tracking either two or three variants in-phase

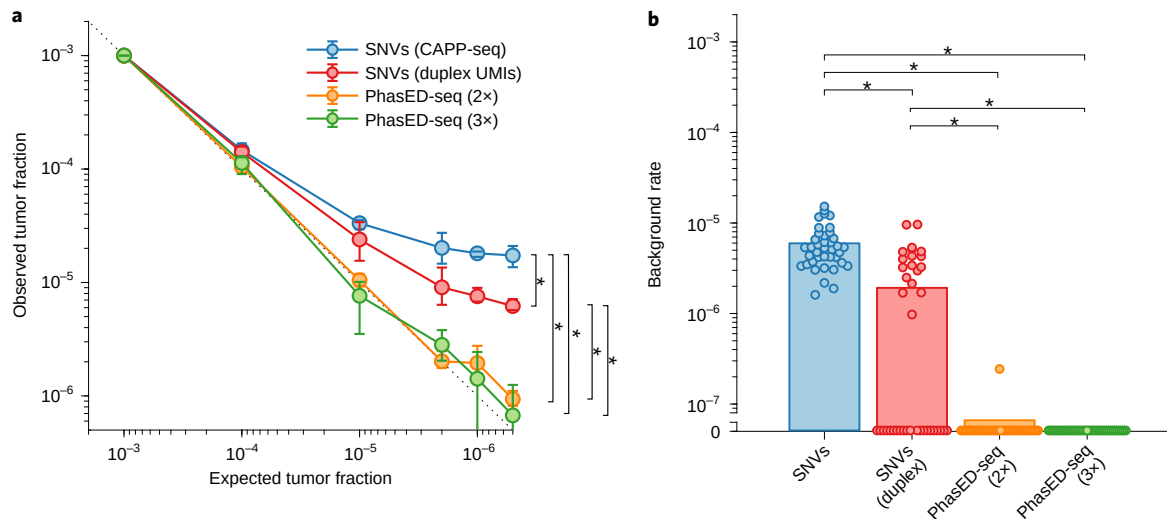


**Fig. 5 | Dilution series to determine detection limits. a**, A limiting dilution series simulating cfDNA containing participant-specific tumor fractions of  $1 \times 10^{-3}$  to  $0.5 \times 10^{-6}$ ; cfDNA from  $n = 3$  independent participant samples were used in each dilution. We analyzed the same sequencing data using multiple methods, including iDES-enhanced CAPP-seq, duplex sequencing and PhasED-seq (both for recovery of doublet and triplet molecules). Data are the mean and range across three independent participant samples. The difference between observed and expected tumor fractions for sample  $< 1:10,000$  were compared via paired *t*-test. \* $P < 0.05$ ; CAPP-seq versus duplex,  $P = 3.2 \times 10^{-5}$ ; CAPP-seq versus PhasED-seq (2x),  $P = 1.1 \times 10^{-4}$ ; CAPP-seq versus PhasED-seq (3x),  $P = 1.9 \times 10^{-5}$ ; duplex versus PhasED-seq (2x),  $P = 0.0047$ ; duplex versus PhasED-seq (3x),  $P = 0.0016$ . **b**, Plot demonstrating the background signal of tumor-specific alleles in 12 unrelated control cfDNA samples, and the control cfDNA sample used for limiting dilution series ( $n = 13$  total samples). In each sample, we assessed for tumor-specific SNVs or PVs from three participant samples used in the limiting dilution experiment, for a total of 39 assessments. Bars represent the mean across all 39 assessments; statistical comparison was performed by Wilcoxon rank-sum test. \* $P < 0.05$ ; CAPP-seq versus duplex,  $P = 3.7 \times 10^{-8}$ ; CAPP-seq versus PhasED-seq (2x),  $P = 4.4 \times 10^{-16}$ ; CAPP-seq versus PhasED-seq (3x),  $P = 2.9 \times 10^{-16}$ ; duplex versus PhasED-seq (2x),  $P = 9.0 \times 10^{-6}$ ; duplex versus PhasED-seq (3x),  $P = 3.1 \times 10^{-6}$ .

(that is, 2x or 3x PhasED-seq) more accurately identified expected tumor content, with superior linearity down to 1 part in 2,000,000 ( $P = 0.005$  for duplex versus 2x PhasED-seq,  $P = 0.002$  for 3x PhasED-seq; Methods). This was also true even without the use of molecular barcodes (Extended Data Fig. 5g). We also assessed specificity of PVs by looking for evidence of tumor-derived SNVs or PVs in cfDNA samples from 12 unrelated healthy controls and the healthy control used for our limiting dilution. Here PhasED-seq showed the lowest background signal across methods (Fig. 5b). This lower error rate and background from PVs improves the detection limit for ctDNA disease detection. The detection limit could be accurately modeled as a binomial sampling process down to the ppm range, with increased detection from increasing number of independent reporters (Extended Data Fig. 5h–j and Methods).

**Improvements in detection of low-burden minimal residual disease.** To test the utility of the lower detection limits from PhasED-seq, we sequenced serial cfDNA from a participant undergoing front-line therapy for DLBCL (Fig. 6a). Using SNV-based methods (that is, CAPP-seq), ctDNA was undetectable after only one cycle of therapy. Subsequently, detectable ctDNA re-emerged  $> 250$  d later, with eventual clinical and radiographic progression 5 months later, indicating falsely negative measurements with SNVs. Strikingly, all four of the plasma samples that were undetectable using SNVs during and after treatment had detectable ctDNA levels by PhasED-seq, with mean allelic fractions as low as 4 in 1,000,000. This increased sensitivity improved the lead time of disease detection by ctDNA compared to radiographic surveillance from 5 to 10 months with PhasED-seq.

We next assessed the performance of PhasED-seq in a cohort of 107 participants with large B cell lymphomas receiving standard immuno-chemotherapy (Supplementary Table 8). In total, we assessed 442 tumor, germline and cfDNA samples, including

cfDNA before therapy ( $n = 107$ ) and after one or two cycles of treatment ( $n = 82$  or 88). Before therapy, participant-specific PVs were detectable by PhasED-seq in 98% of samples, with 97% specificity (Extended Data Fig. 6a). Importantly, ctDNA levels measured by PhasED-seq were highly correlated with those measured by prior SNV-based methods (Spearman  $\rho = 0.93$ ; Extended Data Fig. 6b). We next compared ctDNA levels measured by PhasED-seq and SNV-based methods from samples after initiation of therapy. In total, 79% (77/98) of samples with detectable ctDNA by PhasED-seq after one or two cycles were also detected by SNV-based methods (Fig. 6b). Among 98 samples detected by PhasED-seq, disease burden was significantly lower for those with undetectable (21%) versus detectable (79%) ctDNA levels using SNVs, with a  $> 10$ -fold difference in median ctDNA levels (tumor fraction  $2.9 \times 10^{-4}$  versus  $1.5 \times 10^{-5}$ ,  $P < 0.0001$ ; Fig. 6b). In total, an additional 10% (8/82) of samples after one cycle of therapy and 15% (13/88) of samples after two cycles of therapy had detectable ctDNA when comparing PhasED-seq with SNV-based CAPP-seq (Fig. 6c).

We previously described ctDNA molecular response criteria for DLBCL participants using SNV-based CAPP-seq, including MMR, defined as a 2.5-log reduction in ctDNA after two cycles of therapy<sup>14</sup>. However, many participants had undetectable ctDNA when measured by SNVs at this landmark (Fig. 6d,e). Indeed, in the 88 participants with a sample available from this time point, 59% (52/88) had undetectable ctDNA by SNV-based CAPP-seq at their interim MMR assessment, after completing two of six planned cycles of therapy. Using PhasED-seq, 25% (13/52) of samples not detected by SNVs had evidence of ctDNA as confirmed by PVs, with levels as low as  $\sim 3:1,000,000$  (Extended Data Fig. 7)—these 13 samples additionally detected by PhasED-seq represent potential false-negative tests by prior SNV-based methods. Similar results were seen at the early molecular response time point (that is, after one cycle of therapy; Extended Data Fig. 8).

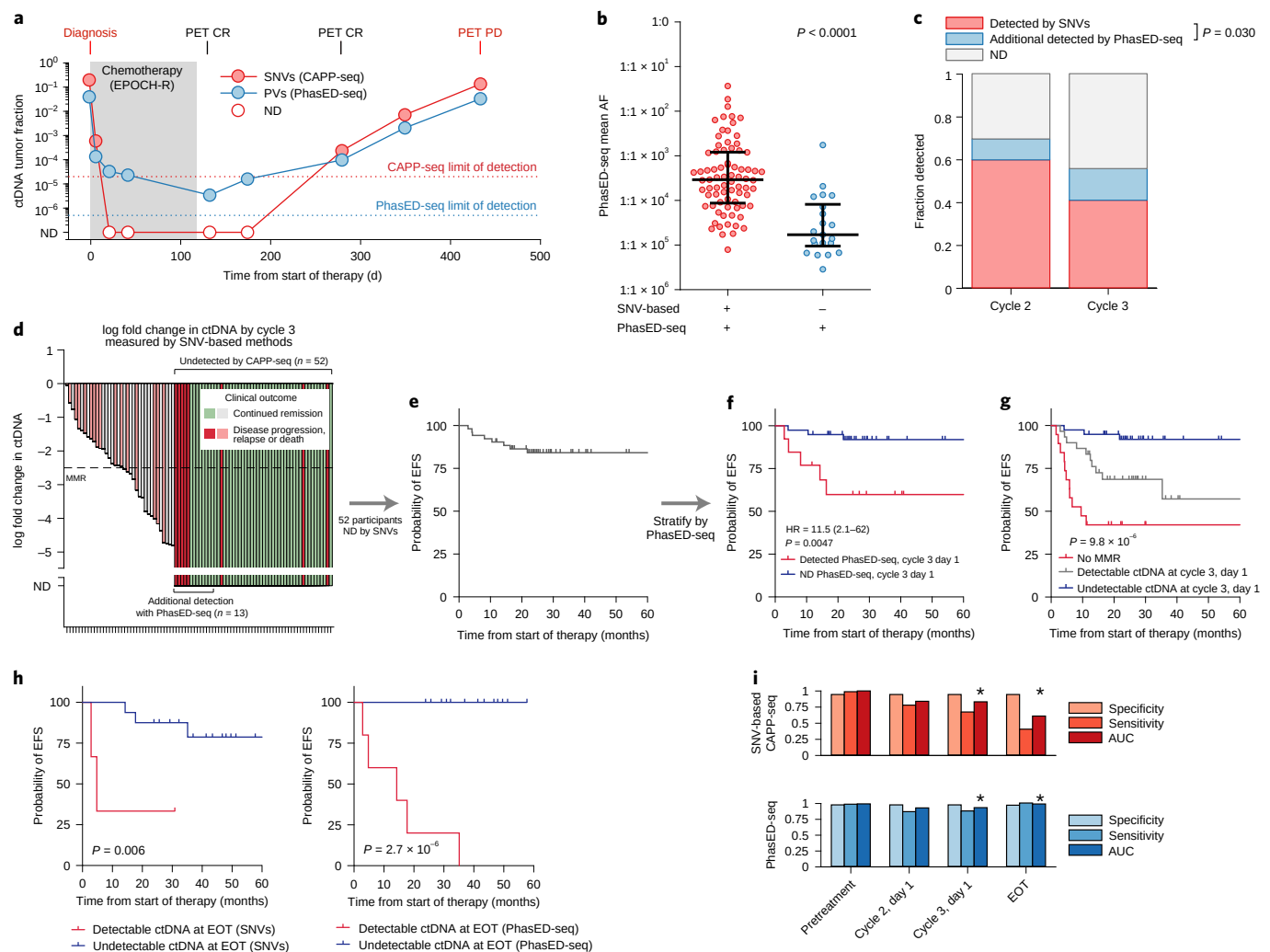


**Fig. 6 | Clinical application of PhasED-seq for ultra-sensitive disease detection and response monitoring in DLBCL.** **a**, Plot showing ctDNA levels for a participant with DLBCL undergoing first-line therapy measured by both CAPP-seq and PhasED-seq. Open circles represent undetectable levels by CAPP-seq. CR, complete response; ND, not detected; PD, progressive disease; PET, positron emission tomography. **b**, Univariate scatterplot showing the mean tumor allele fraction (AF) of  $n = 98$  independent clinical samples measured by PhasED-seq after one or two cycles of therapy for DLBCL. The plot is divided by detection with CAPP-seq;  $P$  value derived from Wilcoxon rank-sum test. Bars show the median and IQR. **c**, Bar plot showing the fraction of DLBCL participants who were ctDNA positive by CAPP-seq after one or two treatment cycles (red), and the fraction of additional participants who were ctDNA positive with addition of PhasED-seq (blue).  $P$  value derived from Fisher's exact test across 170 total samples. **d**, Waterfall plot showing the change in CAPP-seq ctDNA after two cycles in DLBCL participants. **e**, Kaplan-Meier plot showing the EFS for 52 DLBCL participants who were ctDNA negative by CAPP-seq after two cycles. **f**, Kaplan-Meier plot and two-sided log-rank test showing the EFS of the 52 participants in **e** stratified by ctDNA detection with PhasED-seq. HR, hazard ratio. **g**, Kaplan-Meier plot and two-sided log-rank test showing the EFS for 88 participants with DLBCL stratified by ctDNA at cycle 3, day 1 separated into three strata—participants failing to achieve an MMR (red), participants achieving MMR with detectable ctDNA by PhasED-seq and/or CAPP-seq (gray), and participants with undetectable ctDNA by PhasED-seq and CAPP-seq. **h**, Kaplan-Meier plots and two-sided log-rank test showing EFS for 19 DLBCL participants with >24 months of follow-up stratified by EOT ctDNA detection by CAPP-seq (left) or PhasED-seq (right). **i**, Bar plots summarizing the performance of ctDNA by CAPP-seq (top, red) and PhasED-seq (bottom, blue) at various time points. True-positive values included participants with a new diagnosis of lymphoma pretreatment ( $n = 107$ ), or participants known to have eventual disease progression for detection at cycle 2, cycle 3 or EOT ( $n = 22, 24$  or 5 respectively). True-negative values included healthy control cDNAs compared to participant-specific sets of PVs (4,280 total tests).  $*P < 0.05$  for the area under the curve (AUC) comparison by two-sided DeLong test; cycle 3,  $P = 0.043$ ; EOT,  $P = 0.022$ .

While detection of ctDNA in DLBCL after one or two cycles of therapy is a known adverse prognostic marker<sup>14,15</sup>, outcomes for participants with undetectable ctDNA at these time points are heterogeneous (Fig. 6e and Extended Data Fig. 8e). Importantly, even in participants with undetectable ctDNA by SNVs after one or two cycles of therapy, detection of ultra-low ctDNA levels by PhasED-seq was strongly prognostic for outcomes including event-free survival (EFS; Fig. 6f and Extended Data Fig. 8f). When combining detection by PhasED-seq with our previously described MMR threshold,

participants could be stratified into three groups: (1) participants not achieving MMR, (2) participants achieving MMR but with persistent ctDNA and (3) participants with undetectable ctDNA (Fig. 6g). Interestingly, while participants not achieving MMR were at especially high risk for early events despite additional planned first-line therapy (that is, within the first year of treatment), participants with persistent low levels of ctDNA appeared to have a higher risk of later relapse or progression events. In contrast, participants with undetectable ctDNA after two cycles of therapy by PhasED-seq

had overwhelmingly favorable outcomes. Similar results were seen at the early molecular response time point after one cycle of therapy (Extended Data Fig. 8g).

Given its enhanced sensitivity for detection of residual disease during therapy, we sought to explore disease detection via PhasED-seq at the time point of lowest disease burden, that is, at end of therapy (EOT). We assessed SNV-based CAPP-seq and PhasED-seq ctDNA detection in a cohort of 19 participants after therapy, of whom 5 had eventual disease progression. While only 2/5 participants who had eventual progression at this landmark had detectable disease using SNVs, PhasED-seq detected all 5/5 participants, resulting in improved stratification of outcomes (Fig. 6h). PhasED-seq also correctly identified all participants without clinical relapse as having no residual disease, including one participant with abbreviated therapy after only one cycle of treatment (Extended Data Fig. 9). Taken together, this suggests the performance of PV-based PhasED-seq outperformed SNV-based detection for residual disease, particularly during and after therapy (Fig. 6i).

**Extension of PhasED-seq to solid tumors.** While PVs are concentrated in stereotyped genomic loci B cell malignancies, this is not the case in most other cancer types (Extended Data Fig. 3). Despite this, most cancer types considered in the Pan-Cancer Analysis of Whole Genomes (PCAWG) dataset have a sufficient number of PVs in their genome to improve on the limit of detection for ctDNA from SNV-based or alternative approaches using a typical sequencing depth (Fig. 7a, Extended Data Fig. 10a–c and Methods). However, these PVs must first be identified and targeted on an individualized basis. To extend PhasED-seq beyond B cell cancers, a ‘personalized’ approach is therefore needed, including (1) up-front WGS from tumor–normal pairs to identify candidate PVs, (2) design of a personalized panel to select loci containing PVs and (3) targeted sequencing of cfDNA in these regions to monitor the previously defined PVs (Fig. 7b). To explore the potential utility of this approach, we performed personalized PhasED-seq for six participants with solid tumors (lung cancer,  $n=5$ ; breast cancer,  $n=1$ ). In all six cases, multiple PVs were identified from paired tumor–normal WGS data and subsequently validated through targeted resequencing (median 118 validated PVs, range 14–622). Similarly to our lymphoma PhasED-seq panel, the background rate of PVs was lower than the background rate for SNVs across all six personalized panels, even compared to duplex molecules (Fig. 7c).

We next assessed the performance of this approach for MRD detection in 24 plasma samples from these six patients. In nearly all samples, we recovered a sufficient number of cfDNA fragments evaluable for tumor content<sup>40</sup> to detect tumor fractions at or below the limit of detection for SNV-based methods (Fig. 7c). Using SNV-based methods, ctDNA was detected in 9/24 of the plasma

samples. As expected, PhasED-seq also detected ctDNA in these samples (Fig. 7d). Importantly, PhasED-seq also detected ctDNA in 6 additional samples that were negative when assessed by SNVs (Fig. 7d and Extended Data Fig. 10d). These additional samples had very low tumor fractions (median 0.0005%), below the background error rate of SNV-based methods. The lowest measurable tumor fraction was 0.000094%, or less than 1:1,000,000, from a pretreatment sample from a participant with stage I non-small-cell lung cancer (LUP649).

We analyzed serial samples from a participant with stage III lung adenocarcinoma treated with chemoradiotherapy (LUP814; Fig. 7e). CAPP-seq and PhasED-seq detected similar ctDNA levels before therapy; however, three samples after treatment initiation had undetectable ctDNA by CAPP-seq before ctDNA re-emerged at the time of biopsy-confirmed recurrent disease. Using PhasED-seq, we observed molecular residual disease in 3/3 (100%) of samples that were undetected by SNVs, with tumor fraction as low as 0.00016%. Furthermore, the trend in ctDNA quantification mirrored the participant’s disease, with an initial response to chemoradiotherapy but disease progression during immunotherapy. Importantly, this participant’s ctDNA remained detectable at all time points, including after completion of chemoradiotherapy 8 months before disease progression. Similar improvements for detection of low-burden disease after treatment were seen in LUP831 (Extended Data Fig. 10e), as well as BRCA001 for disease detection in anticipation of clinical diagnosis (Fig. 7f).

## Discussion

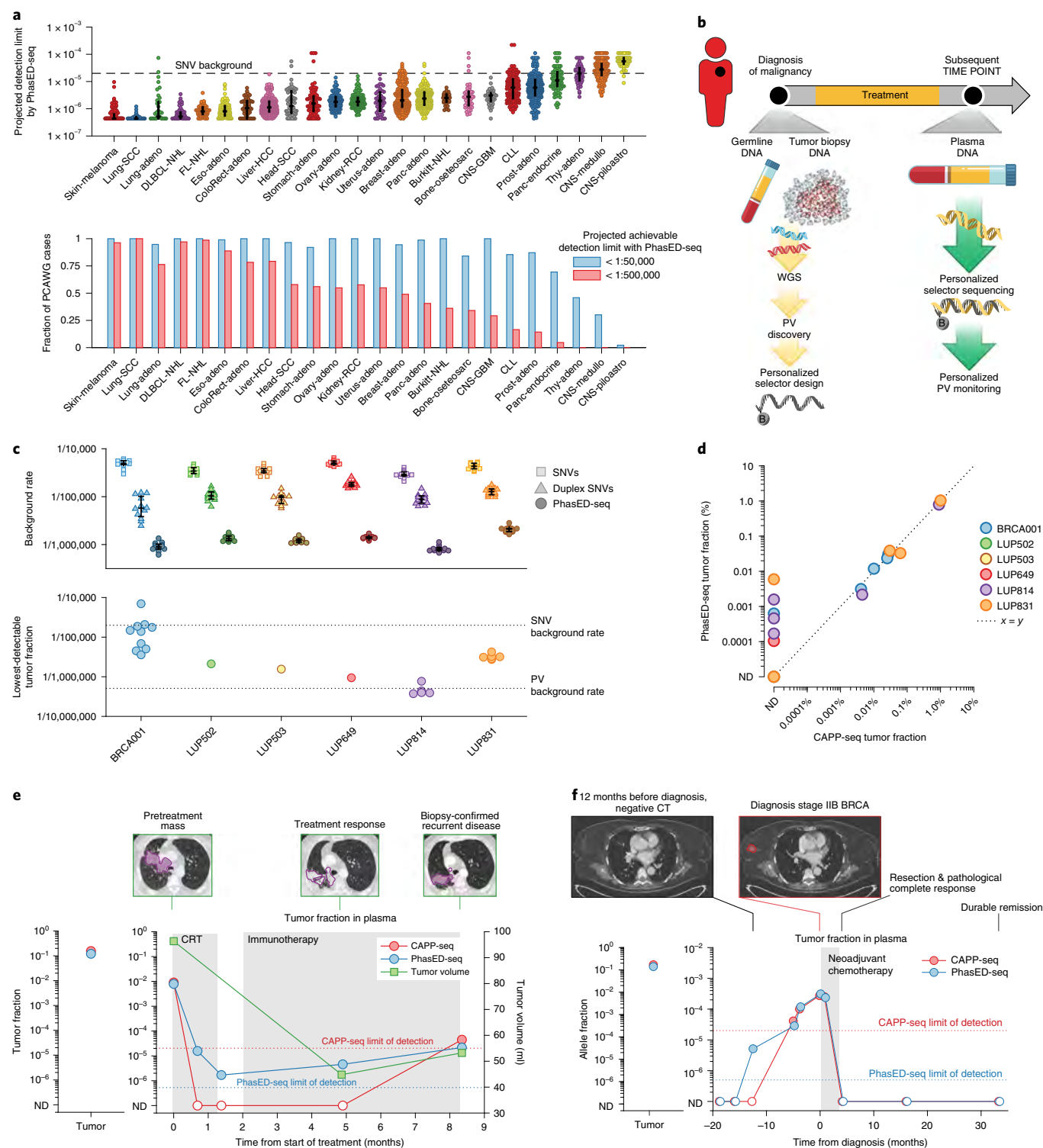
The emergence of cfDNA in oncology has signaled a revolution in precision medicine. While ctDNA-based mutational genotyping and MRD detection are gaining adoption in the clinic, current approaches for disease surveillance have suboptimal sensitivity. We present a method to leverage PVs to improve on these limitations. Using WGS data, we identified putative PVs amenable to capture on a single cfDNA fragment (<170 bp) in multiple cancer subtypes. These PVs were associated with established mutational signatures associated with clustered mutations driven by mechanisms of hypermutation such as AID/AICDA and APOBEC3B.

We observed that cancers derived from B lymphocytes contain PVs in stereotyped genomic regions, while other tumor types contain such events throughout their genome. We therefore designed PhasED-seq, a hybrid-capture approach for disease characterization and MRD detection, and applied it to both lymphomas and solid tumors. We demonstrated in technical and biological experiments the superior error profile of PhasED-seq, establishing an improved sensitivity for PhasED-seq over alternative approaches for ctDNA detection, including error-suppressed SNV-based methods and duplex sequencing. This resulted in detection down to below 1 part per 1,000,000 in a limiting dilution series and clinical samples.

**Fig. 7 | Extension of PhasED-seq for disease monitoring to participants with solid tumors.** **a**, Plot showing the projected detection limit using PVs for ctDNA in cases from the PCAWG dataset. This represents the lowest tumor fraction predicted to be detectable with 95% sensitivity, determined by assuming a personalized PhasED-seq panel and inferring the number of DNA fragments evaluable for tumor content assuming 64 ng of input and 20 million sequencing reads (Methods). The maximum analytical sensitivity is assumed to be 1:2,000,000. Top, participant-level data; bars represent the median and IQR. Bottom, the fraction of participants with at least 1:50,000 and 1:500,000 sensitivity. The background rate for SNVs is shown at  $2 \times 10^{-5}$  (1 in 50,000). **b**, Schematic for personalized PhasED-seq. At the time of diagnosis, tumor and germline WGS were performed to identify a personalized set of PVs. A personalized panel targeting these PVs was then designed. Future cfDNA samples can then be captured and sequenced using this personalized panel. **c**, The performance of personalized PhasED-seq across six patients. Top, the background rate of SNVs (squares), duplex SNVs (triangles) or PVs (circles); bars represent the median and IQR. Bottom, the lowest detectable tumor fraction for each sample. The background rate for SNVs is shown at  $2 \times 10^{-5}$  and for PVs at  $5 \times 10^{-7}$ . **d**, Comparison between the recovered tumor fraction by CAPP-seq (x axis) and PhasED-seq (y axis) for all samples from the six patients with solid tumors. **e**, The ctDNA profile of a participant with stage III lung adenocarcinoma (LUP814). Left, the measured tumor fraction in the tumor biopsy; right, the tumor fraction from cfDNA. While CAPP-seq failed to detect multiple samples with low-burden MRD, PhasED-seq successfully measured disease in all samples. The measured tumor volume and representative computed tomography (CT) scan images are also shown. **f**, The ctDNA profile of a participant with stage II breast adenocarcinoma (BRCA001). Samples were banked before the diagnosis of breast cancer as part of a biomarker study, including at the time of a CT scan for unrelated disease 12 months before diagnosis. At this time point, ctDNA was not detected using CAPP-seq (red), but was detected using PhasED-seq.

The lower background rate of PVs compared with SNVs for tracking known genotypes was expected, as PVs require multiple non-reference bases. We initially expected the background rate of SNVs from duplex sequencing and the rate of PVs to be similar, as both require two independent observations (in *trans* and *cis*, respectively). However, we found that the background rate of SNVs in duplex data was higher than that of PVs, even when considering all molecules (that is, single-stranded and double-stranded recovery) or without using molecular barcodes. We hypothesize that this is most likely due to a biological source of background SNVs. While

'technical' background from PCR error requires two events to create signal for both PhasED-seq and duplex sequencing, SNV 'biological' background—that is, true somatic mutations occurring in non-tumor cells—can be created from a single *in vivo* mutational event and would not be suppressed by duplex barcoding. In contrast, a PV of biological origin still requires two separate mutational events in the same cell, within <170 bp, to result in PV background signal. Consistent with this, while theoretical estimates of the error rate in duplex sequencing were <1 in  $1 \times 10^9$ , the observed rate of background signal from biological samples is  $\sim 1:1 \times 10^5$ – $1 \times 10^6$



(ref. <sup>17</sup>), with reported background SNV rates from duplex sequencing of cfDNA also in this range<sup>2</sup>. Indeed, multiple recent reports of error-suppressed sequencing from cfDNA have demonstrated frequent low-level mutations that can be found in paired leukocyte samples, likely representing low-burden clonal hematopoiesis<sup>41,42</sup>, creating background for duplex sequencing.

To demonstrate the utility of the lower background from PhasED-seq, we studied samples from participants with DLBCL with undetectable ctDNA assessed by prior methods. We demonstrated improved disease detection by PhasED-seq, where participants with occult disease only detectable by PhasED-seq had significantly inferior outcomes to participants without detectable disease. PhasED-seq also improved on disease detection in DLBCL participants at the EOT landmark. Finally, we extended PhasED-seq beyond B cell lymphomas. To apply PhasED-seq to solid cancers that do not have stereotyped locations for PVs, a two-step approach is required, involving up-front tumor WGS to identify PVs, followed by development of a personalized panel to capture PVs. Notably, this type of personalized approach has been successfully applied to ctDNA for SNV-based detection in previous studies<sup>12,43,44</sup>. Applying this to participants with lung or breast cancer, we demonstrate improved disease detection as compared with SNV-based approaches (for example, CAPP-seq), including tumor fractions as low as ~1 in 1,000,000.

We note that this personalized approach could be used for any type of somatic alteration with intrinsically low background error rates, such as fusions or indels, similarly to the implementation here for PVs. Prior approaches using fusions for ctDNA<sup>45,46</sup> have had limited sensitivity due to use of only a few fusions for each participant. However, recent improvements in genome-wide identification of structural variants<sup>47</sup> makes these a potentially attractive alternative for PV-poor cancers. However, most cancer types have a sufficient number of PVs to enable PhasED-seq.

This technical improvement in the performance of ctDNA detection at the time point of lowest disease burden potentially unlocks several new precision medicine approaches. For example, participants achieving molecular remissions without detectable ctDNA by PhasED-seq after receiving chemotherapy might benefit from 'dose de-escalation' or an abbreviated course of therapy to mitigate toxicity and reduce healthcare costs. Similarly, a lower false-negative rate for ctDNA detection after the completion of a planned course of therapy in participants destined for disease relapse could allow for additional consolidative or adjuvant treatment approaches, including new immunotherapy options<sup>48</sup>.

While better prognostication of outcomes for individual patients is useful by itself, for improved ctDNA methodologies such as PhasED-seq to make a significant impact on patient outcomes, changes in treatment paradigms are required. Prospective clinical trials implementing PhasED-seq to select patients for personalized therapies will therefore be necessary to establish its utility before wide clinical adoption.

### 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 and code availability are available at <https://doi.org/10.1038/s41587-021-00981-w>.

Received: 14 March 2020; Accepted: 11 June 2021;

Published online: 22 July 2021

### References

- Diehl, F. et al. Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **14**, 985–990 (2008).

- Newman, A. M. et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **34**, 547–555 (2016).
- Scherer, F. et al. Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. *Sci. Transl. Med.* **8**, 364ra155 (2016).
- Chabon, J. J. et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. *Nat. Commun.* **7**, 11815 (2016).
- Bettegowda, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra224 (2014).
- Sozzi, G. et al. Analysis of circulating tumor DNA in plasma at diagnosis and during follow-up of lung cancer patients. *Cancer Res.* **61**, 4675–4678 (2001).
- Thierry, A. R. et al. Clinical validation of the detection of *KRAS* and *BRAF* mutations from circulating tumor DNA. *Nat. Med.* **20**, 430–435 (2014).
- Tie, J. et al. Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer. *Sci. Transl. Med.* **8**, 346ra392 (2016).
- Kalinich, M. & Haber, D. A. Cancer detection: seeking signals in blood. *Science* **359**, 866–867 (2018).
- Phallen, J. et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* <https://doi.org/10.1126/scitranslmed.aan2415> (2017).
- Abbosh, C., Birkbak, N. J. & Swanton, C. Early stage NSCLC—challenges to implementing ctDNA-based screening and MRD detection. *Nat. Rev. Clin. Oncol.* **15**, 577–586 (2018).
- Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446–451 (2017).
- Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
- Kurtz, D. M. et al. Circulating tumor DNA measurements as early outcome predictors in diffuse large B cell lymphoma. *J. Clin. Oncol.* **36**, 2845–2853 (2018).
- Roschewski, M. et al. Circulating tumour DNA and CT monitoring in patients with untreated diffuse large B cell lymphoma: a correlative biomarker study. *Lancet Oncol.* **16**, 541–549 (2015).
- Garcia-Murillas, I. et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Sci. Transl. Med.* **7**, 302ra133 (2015).
- Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).
- Kennedy, S. R. et al. Detecting ultralow-frequency mutations by duplex sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
- Medina Diaz, I. et al. Performance of Streck cfDNA blood collection tubes for liquid biopsy testing. *PLoS ONE* **11**, e0166354 (2016).
- de Yébenes, V. G. & Ramiro, A. R. Activation-induced deaminase: light and dark sides. *Trends Mol. Med.* **12**, 432–439 (2006).
- Pasqualucci, L. et al. Hypermutation of multiple proto-oncogenes in B cell diffuse large-cell lymphomas. *Nature* **412**, 341–346 (2001).
- Jiang, P. et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl Acad. Sci. USA* **112**, E1317–E1325 (2015).
- Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013).
- Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
- Khodabakhshi, A. H. et al. Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308–1319 (2012).
- Schmitz, R. et al. Genetics and pathogenesis of diffuse large B cell lymphoma. *N. Engl. J. Med.* **378**, 1396–1407 (2018).
- Rowley, J. D. Chromosome studies in the non-Hodgkin's lymphomas: the role of the 14;18 translocation. *J. Clin. Oncol.* **6**, 919–925 (1988).
- Schmitz, R. et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **490**, 116–120 (2012).
- Alizadeh, A. A. et al. Distinct types of diffuse large B cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
- Lenz, G. et al. Aberrant immunoglobulin class switch recombination and switch translocations in activated B cell-like diffuse large B cell lymphoma. *J. Exp. Med.* **204**, 633–643 (2007).
- Morin, R. D. et al. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood* **122**, 1256–1265 (2013).
- Qian, J. et al. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**, 1524–1537 (2014).
- Richter, J. et al. Recurrent mutation of the *ID3* gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.* **44**, 1316–1320 (2012).
- Puente, X. S. et al. Noncoding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).

37. Pasqualucci, L. et al. Analysis of the coding genome of diffuse large B cell lymphoma. *Nat. Genet.* **43**, 830–837 (2011).
  38. Steidl, C. et al. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature* **471**, 377–381 (2011).
  39. Nakamura, N. et al. Analysis of the immunoglobulin heavy chain gene variable region of CD5-positive and -negative diffuse large B cell lymphoma. *Leukemia* **15**, 452–457 (2001).
  40. Wan, J. C. M. et al. ctDNA monitoring using patient-specific sequencing and integration of variant reads. *Sci. Transl. Med.* <https://doi.org/10.1126/scitranslmed.aaz8084> (2020).
  41. Chabon, J. J. et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245–251 (2020).
  42. Razavi, P. et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat. Med.* **25**, 1928–1937 (2019).
  43. Bratman, S. V. et al. Personalized circulating tumor DNA analysis as a predictive biomarker in solid tumor patients treated with pembrolizumab. *Nat. Cancer* **1**, 873–881 (2020).
  44. Reinert, T. et al. Analysis of plasma cell-free DNA by ultradeep sequencing in patients with stages I to III colorectal cancer. *JAMA Oncol* <https://doi.org/10.1001/jamaoncol.2019.0528> (2019).
  45. Leary, R. J. et al. Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* **2**, 20ra14 (2010).
  46. Leary, R. J. et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci. Transl. Med.* **4**, 162ra154 (2012).
  47. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
  48. Neelapu, S. S. et al. Axicabtagene ciloleucel CAR T-cell therapy in refractory large B cell lymphoma. *N. Engl. J. Med.* **377**, 2531–2544 (2017).
- 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. 2021

## Methods

**Participant enrollment and clinical sample collection.** All samples analyzed in this study were collected with informed consent from participants enrolled on studies approved by an Institutional Review Board that were in accordance with the Declaration of Helsinki and complied with all relevant ethical regulations at their respective centers—Stanford University, MD Anderson Cancer Center, the National Cancer Institute, University of Eastern Piedmont, Essen University Hospital and CHU Dijon. Participants were not compensated for their contribution in this study.

Individuals with B cell lymphomas undergoing front-line therapy were enrolled on this study from six centers across North America and Europe, including Stanford University, MD Anderson Cancer Center, the National Cancer Institute, University of Eastern Piedmont (Italy), Essen University Hospital (Germany) and CHU Dijon (France). An additional six participants with a diagnosis of solid tumors (lung cancer,  $n=5$ ; breast cancer,  $n=1$ ) were enrolled at Stanford University. In total, 386 cfDNA, 109 tumor and 219 germline samples from 219 participants and 76 cfDNA samples from controls were included in this study. Cell-free, tumor and germline DNA were isolated as previously described<sup>2,3,13,14,49</sup>. All radiographic imaging was performed as part of standard clinical care.

**Whole-genome sequencing analysis. Putative phased variant identification from whole-genome sequencing.** WGS data were obtained from two sources. Data for lymphoid malignancies (DLBCL, FL, BL and CLL) were downloaded from the International Cancer Genome Consortium (ICGC) data portal on 7 May 2018. Data from all other histologies were part of the PCAWG dataset and were downloaded on 11 November 2019 (ref. <sup>30</sup>). Only cancer histologies with at least 35 available cases were considered; details of the dataset considered are provided in Supplementary Table 1. All samples had somatic mutations called from WGS using matched tumor and normal genotyping. Queries were limited to base substitutions obtained from WGS (single-, double-, triple- and oligo-nucleotide variants (SNVs, DNVs, TNVs and ONVs)). Having thus identified the cases and variants of interest, we next identified the number of putative PVs in each tumor. To function as a PV on a single cfDNA molecule, two SNVs must occur within a genomic distance less than the length of a typical cfDNA molecule (~170 bp). We therefore defined putative PVs as two variants occurring on the same chromosome within a genomic distance of <170 bp. DNVs, TNVs and ONVs were considered as the set of their respective component SNVs. The number of SNVs and identity of putative PVs for each case are detailed in Supplementary Table 1. The raw number of SNVs and putative PVs, as well as the number of putative PVs controlling for the number of SNVs, is shown in Extended Data Fig. 2.

**Mutational signatures of phased variants from WGS.** To assess the mutational processes associated with phased and non-phased mutations across different cancer types/subtypes, we enumerated the mutational signatures of SBSs for each WGS case described above using the R package 'deconstructSigs'<sup>51</sup>. We began by first dividing the list of SNVs for each participant into two groups: (1) SNVs contained within a possible PV; that is, with an adjacent or 'nearest neighbor' SNV < 170 bp away, and (2) isolated SNVs (that is, non-phased), defined as those occurring ≥ 170 bp in distance from the closest adjacent SNV. We then applied 'deconstructSigs' using the 49 SBS signatures described in COSMIC<sup>24</sup> (excluding signatures linked to possible sequencing artifacts) to assess the contribution of each SBS signature to both candidate phased SNVs and unphased SNVs for each participant. To compare the contribution of each SBS signature to phased and isolated SNVs, we performed a Wilcoxon signed-rank test to compare the relative contribution of each SBS signature between these two categories for each cancer type. To account for multiple hypotheses, we applied Bonferroni's correction, by considering any SBS signature that differed in contribution to phased versus unphased SNVs to be significant if the Wilcoxon signed-rank test resulted in a  $P$  value < 0.05/49 or 0.001. The distributions of these comparisons for each of the 49 SBS signatures, along with significance testing, are depicted in figures available at <https://phasedseq.stanford.edu/>. A summary of this analysis is also shown in Fig. 1c using a heat-map display, where the 'heat' represents the difference between the mean contribution of the SBS signature to PVs to the mean contribution to isolated/unphased variants.

**Genomic distribution of phased variants from WGS.** We assessed the recurrence frequency for PVs in each cancer type across the genome within each tumor type. Specifically, we first divided the human genome (build GRCh37/hg19) into 1-kb bins (3,095,689 total bins); then, for each sample, we counted the number of PVs (as defined above) contained in each 1-kb bin. For this analysis, we included any PV with at least one of its constituent SNVs falling within the 1-kb bin of interest. We then calculated the fraction of participants whose tumors harbored a PV for each cancer type within each genomic bin. To identify 1-kb bins recurrently harboring PVs across participants, we plotted the fraction of participants containing PVs in each 1-kb bin versus genomic coordinates (Fig. 1d and Extended Data Fig. 3); for this analysis, we only plotted bins where at least 2% of samples contained a PV in at least one cancer subtype.

**Identification of recurrent 1-kb bins with phased variants.** To identify 1-kb bins that recurrently contain PVs in B lymphoid malignancies, we utilized WGS data

from the following diseases: DLBCL, FL, BL and CLL. Any 1-kb bin in which >1 sample from these tumor types contained a PV was considered to recurrently contain PVs from B lymphoid malignancies. The genomic coordinates of 1-kb bins containing recurrent PVs in lymphoid malignancies are enumerated in Supplementary Table 2.

**Design of PhasED-seq panel for B lymphoid malignancies. Identification of recurrent PVs from whole-genome sequencing at high resolution.** Given the prevalence of recurrent putative PVs from WGS data in B cell malignancies, we designed a targeted sequencing approach for their hybridization-mediated capture, PhasED-seq, to enrich these specific PV events from tumor or cfDNA. In addition to the ICGC data described above, we also used WGS data from other sources in this design, including both B cell NHLs and CLL<sup>33–36</sup>. We also considered our previous experience with targeted sequencing from cfDNA in NHLs<sup>5</sup>. We began by identifying pairs of SNVs occurring at a distance of <170 bp apart in each B cell tumor sample. We then identified genomic 'windows' that contained PVs as follows: for each chromosome, we sorted the PVs by genomic coordinates relative to the reference genome. We then identified the left-most position for any PV in any participant; this defined the left-hand (5') coordinate seeding a desired window of interest, to be captured from the genome. We then extended this window by growing its 3' end to capture successive PVs until a gap of ≥340 bp was reached, with 340 bp chosen as capturing two successive chromosomal-sized fragments of ~170 bp. When such a gap was reached, a new window was started, and this iterative process of adding neighboring PVs was repeated again until the next gap of ≥340 bp was reached. This resulted in a BED file of genomic windows containing all possible PVs from all samples considered. Finally, each window was additionally padded by 50 bp on each side, to enable efficient capture from flanking sequences in rare scenarios when repetitive or poorly mapping intervening sequences might preclude their direct targeting for enrichment.

Having identified the regions of interest containing putative PVs, we then divided each window into 170-bp segments (that is, the approximate size of a chromosomal cfDNA molecule). We then enumerated the number of cases containing a PV in each case. For each 170-bp region, we included the region in our final sequencing panel design if one or more of the following criteria was met: (1) at least one participant contained a PV in the 170-bp region in three of our five independent datasets, (2) at least one participant contained a PV in the region in two of five independent datasets if one dataset was our prior CAPP-seq experience<sup>5</sup>, or (3) at least one participant contained a PV in the region in two of five independent datasets, with a total of at least three participants containing a PV in the region. This resulted in 691 'tiles', with each tile representing a 170-bp genomic region. These tiles, along with an additional ~200 kb of genomic space targeting driver genes recurrently mutated in B-NHL, were combined into a unified targeted sequencing panel as previously described for both tumor and cfDNA genotyping<sup>2,13</sup> using NimbleDesign (Roche NimbleGen).

**Comparison of phased variant yield from PhasED-seq and CAPP-seq.** To evaluate the performance of PhasED-seq for capturing both SNVs and PVs compared to our previously reported CAPP-seq selector for B cell lymphomas, we quantified the predicted number of both SNVs and PVs that would be recovered with each panel by limiting WGS in silico prediction to the capture targets of each approach (Extended Data Fig. 4a–c). The predicted number of variants was then compared using the Wilcoxon signed-rank test. We also performed both CAPP-seq and PhasED-seq on 16 samples from participants with DLBCL. In these samples, tumor or plasma DNA, along with matched germline DNA, was sequenced. The resulting number of variants were again compared by the Wilcoxon signed-rank test (Fig. 3a and Extended Data Fig. 4d,e). The sequencing depths for the samples included in this analysis are provided in Supplementary Table 3.

**Identification of phased variants from targeted sequencing. Library preparation and sequencing.** To generate sequencing libraries and targeted sequencing data, we applied CAPP-seq as previously described<sup>2,13</sup>. Briefly, cell-free, tumor, and germline DNA were used to construct sequencing libraries through end repair, A-tailing and adaptor ligation using the KAPA Hyper Prep Kit according to the manufacturer's instructions with ligation performed overnight at 4°C. CAPP-seq adaptors with UMIs were used for barcoding of unique DNA duplexes and subsequent deduplication of sequencing read pairs. Hybrid captures were then performed for lymphoma samples using (SeqCap EZ Choice; NimbleGen) at 47°C for 48 h and (xGen Lockdown Probe Pools; IDT) for solid tumors at 65°C for 16 h. Lymphoma samples utilized an off-the-shelf panel, while solid tumor samples utilized a customized panel as described in the main text. Affinity capture was performed according to the manufacturer's protocol, with the respective hybridization conditions conducted on an Eppendorf thermal cycler. Following enrichment, libraries were sequenced using an Illumina HiSeq 4000 instrument with 2 × 150-bp paired-end reads.

**Preprocessing and alignment.** FASTQ files were demultiplexed and UMIs were extracted using a custom pipeline as previously described<sup>2,13</sup>. FASTQ files were then processed using fastp<sup>52</sup> without global trimming, with an unqualified percentage limit of 50% and a required length of 35. All other parameters were kept as default. For more efficient parallelization and to minimize barcode collisions, sample

FASTQ files were then split into multiple files based on fragment UMIs. Each FASTQ file was then aligned to the human genome (build GRCh37/hg19) using the BWA ALN algorithm<sup>53</sup> without fragment size estimation, and deduplicated using both samtools 'markup' and our previously described in-house algorithm<sup>2</sup>. UMI-specific files were then merged to generate final sample alignment files. Molecular barcode-mediated error suppression and background polishing (that is, iDES) were then performed as previously described<sup>2</sup>.

**Identification of phased variants.** For detailed methods of identification of PVs and MRD detection, see the Supplementary Methods.

**Assessment of hybridization bias.** To assess the effect of mutations on hybridization efficiency, we first estimated the affinity of mutated molecules to wild-type capture baits in silico by considering DNA fragments harboring 0–30% mutations across the entire fragment. For each mutation condition across this range, we first randomly sampled 10,000 regions, each 150 bp in length, from across the whole genome. We then mutated these 150-mers in silico to simulate the desired mutation rate in three different ways: (1) mutating 'clustered' or contiguous bases starting from the ends of a sequence, (2) mutating clustered bases started from the middle of the sequence, or (3) mutating bases selected at random positions throughout the sequence. We then used the 'energy.c' package to calculate the theoretical binding energy (kcal mol<sup>-1</sup>) between the mutated and wild-type sequences, in relying on a nearest-neighbor model using established thermodynamic parameters<sup>54–56</sup>.

We then replicated this in silico experiment by testing the effects of same mutation rates in vitro. Specifically, we synthesized and annealed oligonucleotides (IDT) to form DNA duplexes harboring 0–10% mutations at defined positions relative to the human reference genome sequence. We then captured these synthetic DNA molecules together at equimolar concentrations and quantified the relative capture efficiency of mutated duplexes compared to the wild-type, unmutated species (Fig. 4a). Two sets of oligonucleotide sequences were selected from the coding regions of *BCL6* (ref. <sup>57</sup>) and *MYC*<sup>58</sup> to capture AID-mediated aSHMs associated with each gene<sup>35</sup>; we ensured the preserved mappability of the mutated species by BWA ALN (Supplementary Table 6). These synthetic oligonucleotide duplexes were then subjected to library preparation and then captured and sequenced using PhasED-seq, performed in triplicate using distinct samples. This allowed assessment of the relative efficiency of hybrid capture and molecular recovery as directly compared to wild-type molecules identical to the reference genome.

**Assessment of limit of detection with dilution series.** To empirically define the analytical sensitivity of PhasED-seq, we utilized a limited dilution series of ctDNA from three participants that were spiked into healthy control ctDNA at defined concentrations. Our dilution series contained samples with an expected mean tumor fraction of 0.1%, 0.01%, 0.001%, 0.0002%, 0.0001% and 0.00005% or ranging from 1 part in 1,000 to 1 part in 2,000,000. The sequencing characteristics and ctDNA quantification via CAPP-seq, duplex sequencing and PhasED-seq are provided in Supplementary Table 7. Tumor fractions were normalized to the measured concentration of the highest sample (that is, the 1:1,000 sample). To compare the performance of each method, we calculated the difference,  $\delta$ , between the observed and expected tumor fraction for each participant  $i$  at each dilution concentration  $j$ , given by equation (1):

$$\delta_{ij} = \widehat{\text{tumorfrac}}_{i,j} - \text{tumorfrac}_{i,j} \quad (1)$$

This value was calculated for participants  $I = \{1, 2, 3\}$  and concentrations  $j = \{0.001\%, 0.0002\%, 0.0001\%, 0.00005\% \}$  for each ctDNA detection method (CAPP-seq, duplex, doublet PhasED-seq and triplet PhasED-seq). The performance of each method was then compared to each other by paired  $t$ -test across this set of participants and concentrations. Results are shown in Fig. 5a, as well as in Extended Data Fig. 5g without the use of barcode deduplication (that is, samtools markup without UMI-mediated error suppression).

**Extension of PhasED-seq to solid tumors.** *Whole-genome sequencing to identify phased variants.* To demonstrate the utility of PhasED-seq for solid tumors, we first performed WGS on tumor and matched normal DNA from six tumors (lung cancer,  $n = 5$ ; breast cancer,  $n = 1$ ) to ~30× read depth (Supplementary Table 9). Sequencing reads were aligned to hg19 and deduplicated with samtools markup, as described above. Per GATK best practices<sup>60</sup>, tumor and normal deduplicated BAM files were processed with GATK IndelRealigner and BaseRecalibrator before variant calling, using default parameters (GATK v3.8-1-0-gf15c1c3ef)<sup>60</sup>. Variant calling was performed using three methods: VarScan2 (v2.3.9)<sup>61</sup>, Mutect (v1.1.7)<sup>62</sup> and Strelka2 (v2.9.1)<sup>63</sup>. Mutect and VarScan2 VCF files were annotated by annovar (v2018Apr16), and Strelka VCF files were annotated by Oncotator (v1.9.8.0). Variants called by each method were combined and filtered according to the following criteria: (1) pass caller-intrinsic quality filters (for example, base quality, orientation bias and germline risk); (2) depth  $\geq 30\times$ ; (3) AF  $\geq 5\%$ ; and (4) variant identified by  $\geq 2$  variant callers. SNVs passing all filters were then assessed for possible phased relationships—any pair of SNVs  $\leq 170$  bp from its nearest neighbor was considered a possible PV. We also genotyped PVs directly

from WGS reads, considering any candidate PVs with at least two supporting reads, 10× depth and 5% tumor fraction. Candidate PVs were then assessed and prioritized for tumor specificity, considering the (1) presence in individual tumor reads as phased relationships, (2) absence of read support in matched normal, (3) presence of other non-reference bases on the supporting reads, (4) base quality, (5) mapping quality and (6) uniqueness of genomic positions. Candidate PVs were then selected for targeted resequencing below.

*Targeted resequencing to validated phased variants.* After identifying candidate PVs from each of these six tumors, we designed 120-bp biotinylated hybrid-capture oligonucleotides targeting the regions of interest (Integrated DNA Technologies). We then performed hybrid capture resequencing of the tumor–normal pairs to high unique molecular depth to create a validated list of PVs for tumor monitoring. A PV was considered to be validated if it was present in the tumor at higher than 5% AF and had no read support in the matched germline DNA.

*Measurement of ctDNA in participants with solid tumors.* We applied the above personalized hybrid-capture panels targeting PVs to plasma samples from each of these six participants, sequencing to high unique molecular depth. We also sequenced 24 control healthy cfDNA samples with each panel to assess specificity. Tumor fraction was defined as the number of reads containing an a priori defined PV over the total number of reads covering a PV position (that is, PV tumor fraction, defined in 'Determination of tumor fraction in a sample from phased variants' in the Supplementary Methods). Most samples had been assessed for ctDNA content using SNV-based CAPP-seq approaches previously<sup>41,64</sup>, providing comparison to PhasED-seq. The results of these experiments are shown in Fig. 7c–f and Extended Data Fig. 10d,e.

**Statistical analyses.** All  $P$  values reported in this paper are two-sided unless otherwise noted. Comparisons of matched samples and populations were performed using the Wilcoxon signed-rank test, whereas comparisons of samples drawn from unrelated populations were performed using the Wilcoxon rank-sum test. Comparisons of paired samples were performed by paired  $t$ -test. Survival probabilities were estimated using the Kaplan–Meier method; survival of groups of participants based on ctDNA levels were compared using the log-rank test. Other statistical tests are noted in the main text where used. All analyses were performed using MATLAB (v2018b), R (v3.4.1), Perl (v5.10.1) and GraphPad Prism (v8.0.2). The contribution of known mutational processes to phased and isolated SNVs from WGS was assessed with the 'deconstructSigs' R package using the COSMIC signature set (v2) as described. Calculation of AUC accounting for survival and censorship was performed using the R 'survivalROC' package (v1.0.3) with default settings. The PhasED-seq software (v1.0) was developed using C++17.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

WGS data were obtained from the PCAWG dataset available at <https://dcc.icgc.org/pcawg/>. Additional lymphoma WGS data were obtained from the ICGC website (malignant lymphoma project, MALY-DE), as well as from work by Morin et al.<sup>33</sup> and Qian et al.<sup>34</sup>. Raw data for samples used in the dilution series are available at <http://phasedseq.stanford.edu/>. This website provides users with software for phased variant monitoring (PhasED-seq v1.0). Included here are the list of SNVs and PVs used to detect tumor DNA from each of the three samples used in this dilution series, as well as the sequencing files for each concentration condition. Owing to restrictions related to dissemination of germline sequence information included in the informed consent forms used to enroll study participants, we are unable to provide access to other raw sequencing data. Reasonable requests for additional data will be reviewed by the senior authors to determine whether they can be fulfilled in accordance with these privacy restrictions. Requests for additional materials related to this work should be directed to A.A.A.

## Code availability

PhasED-seq (v1.0) was used to generate the results in this work and is freely available for academic research use at <http://phasedseq.stanford.edu/>, including a downloadable compiled version of the software. For academic users, the source code is available upon request.

## References

- Kurtz, D. M. et al. Reply to J. Wang et al. *J. Clin. Oncol.* **37**, 755–757 (2019).
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
- 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).
- Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
54. Bozdech, Z. et al. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol.* **4**, R9 (2003).
55. Jaeger, J. A., Turner, D. H. & Zuker, M. Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. USA* **86**, 7706–7710 (1989).
56. Sugimoto, N., Nakano, S., Yoneyama, M. & Honda, K. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.* **24**, 4501–4505 (1996).
57. Lu, Z. et al. *BCL6* breaks occur at different AID sequence motifs in Ig-*BCL6* and non-Ig-*BCL6* rearrangements. *Blood* **121**, 4551–4554 (2013).
58. Robbiani, D. F. et al. AID is required for the chromosomal breaks in *c-myc* that lead to *c-myc/IgH* translocations. *Cell* **135**, 1028–1038 (2008).
59. Lieber, M. R. Mechanisms of human lymphoid chromosomal translocations. *Nat. Rev. Cancer* **16**, 387–398 (2016).
60. Van der Auwera, G. A. et al. From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
61. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
62. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
63. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
64. Moding, E. J. et al. Circulating tumor DNA dynamics predict benefit from consolidation immunotherapy in locally advanced non-small-cell lung cancer. *Nat. Cancer* **1**, 176–183 (2020).

## Acknowledgements

This work was supported by the National Cancer Institute (R01CA233975 and R01CA188298 to A.A.A. and M.D.; K08CA241076 to D.M.K.), the Virginia and D. K. Ludwig Fund for Cancer Research (A.A.A. and M.D.), the Bakewell Foundation (A.A.A. and M.D.), the Damon Runyon Cancer Research Foundation (PST no. 09-16 to D.M.K. and DR-CI no. 71-14 to A.A.A.), the American Society of Hematology Scholar Award (A.A.A.), the V Foundation for Cancer Research Abeloff Scholar Award (A.A.A.), the Conquer Cancer Foundation of the American Society of Clinical Oncology (D.M.K.), the Emerson Collective Cancer Research Fund (A.A.A.), the Stinehart/Reed Award (A.A.A.), the CRK Faculty Scholar Fund (M.D.) and the SDW/DT and Shanahan Family Foundations (A.A.A.). A.A.A. is a Scholar of The Leukemia & Lymphoma Society.

## Author contributions

D.M.K., J.S., M.D. and A.A.A. developed the concept, designed the experiments and analyzed the data. D.M.K., J.S., M.D. and A.A.A. wrote the manuscript. D.M.K., J.S. and L.C.T.K. developed the phased variant identification and monitoring workflows with input from M.D. and A.A.A. D.M.K., J.S., S.A., J.J.C., B.J.S., M.C.J., F.S., C.W.M., M.O., A.F.M.C. and E.J.M. performed molecular biology experiments related to the technical performance of PhasED-seq and detection of ctDNA in clinical samples. D.M.K., J.S., L.C.T.K., A.S., A.G., E.G.H., B.C., J.G.S.-M., A.F.M.C., M.S.E. and C.L.L. performed bioinformatic analyses. Participant specimens were provided by D.M.K., E.J.M., U.D., A.H., R.-O.C., J.R.W., M.R., W.H.W., G.G., D.R., M.D. and A.A.A. Participant enrollment, sample collection and data curation were organized and performed by D.M.K., U.D., A.H., R.-O.C., J.R.W., M.R., W.H.W., G.G., D.R., M.D. and A.A.A. All authors reviewed the manuscript.

## Competing interests

D.M.K. reports paid consultancy from Roche and Genentech. A.A.A. reports research funding from Celgene and Pfizer, ownership interests in FortySeven and CiberMed, and paid consultancy from Roche, Genentech, Janssen, Pharmacyclics, Gilead, Celgene and Chugai. M.D. reports research funding from Varian Medical Systems, AstraZeneca and Illumina, ownership interest in CiberMed, and paid consultancy from Roche, AstraZeneca, Novartis, Genentech, Illumina, RefleXion, Gritstone Oncology, Boehringer Ingelheim and BioNTech. M.D., A.A.A., D.M.K., J.C. and M.S.E. report patent filings related to cancer biomarkers. M.D., A.A.A., D.M.K. and J.C. report ownership interest in Foresight Diagnostics. R.O.C. is a member of the Roche advisory board and received research support from Roche. D.R. received research support from Gilead, Janssen, Roche and AbbVie, outside the submitted work. E.J.M. has served as a paid consultant for DeciBio. The remaining authors declare no competing interests.

## Additional information

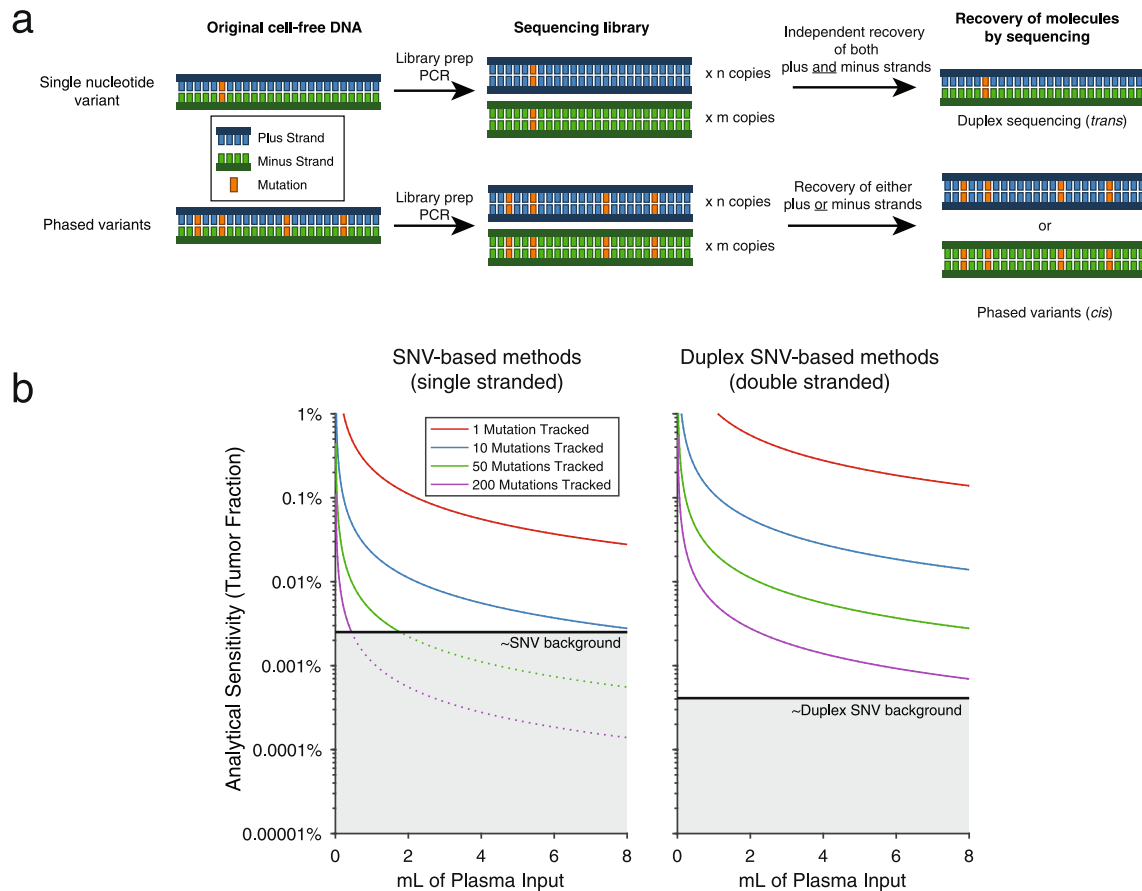
**Extended data** is available for this paper at <https://doi.org/10.1038/s41587-021-00981-w>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41587-021-00981-w>.

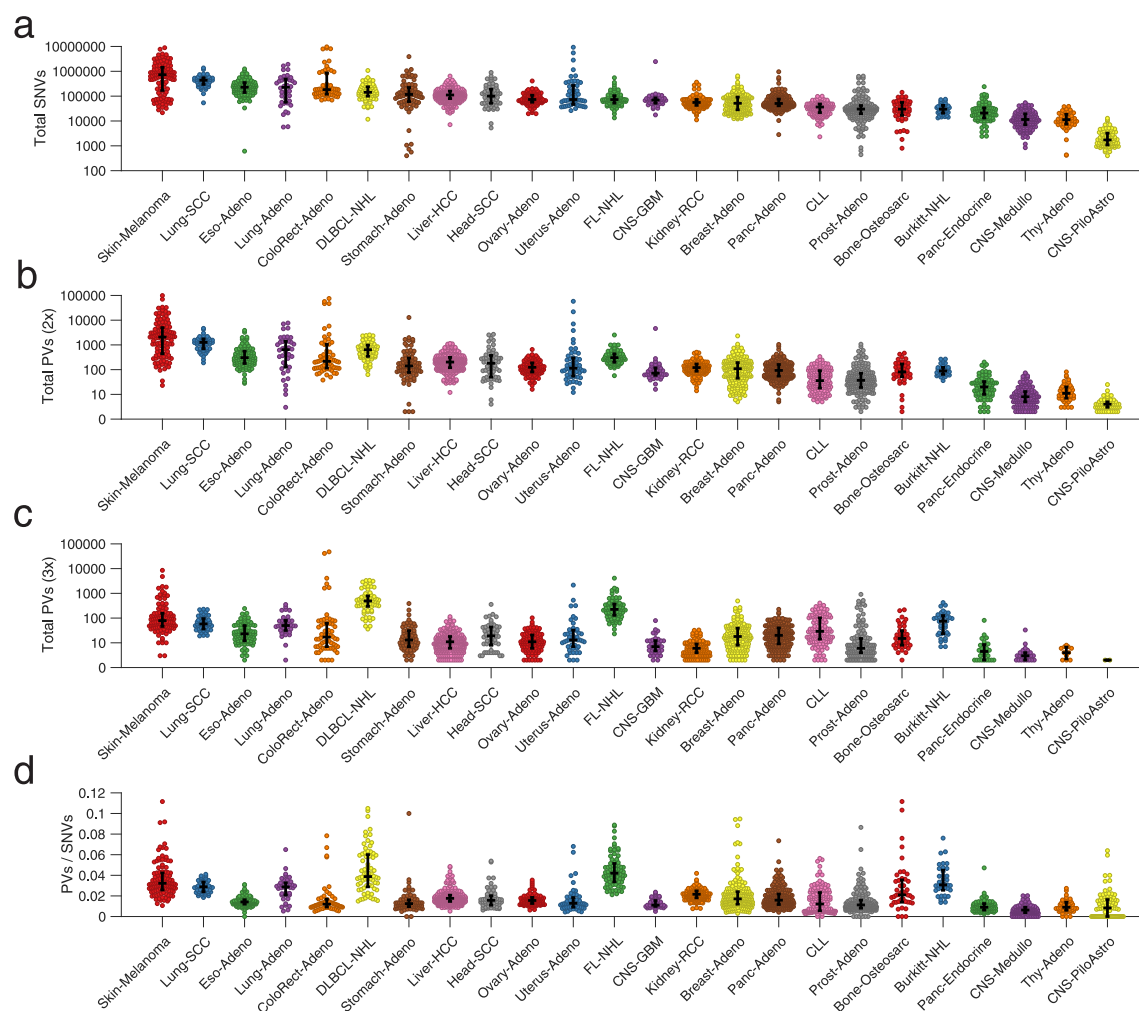
**Correspondence and requests for materials** should be addressed to M.D. or A.A.A.

**Peer review information** *Nature Biotechnology* thanks Paul Spellman, Christian Steidl and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

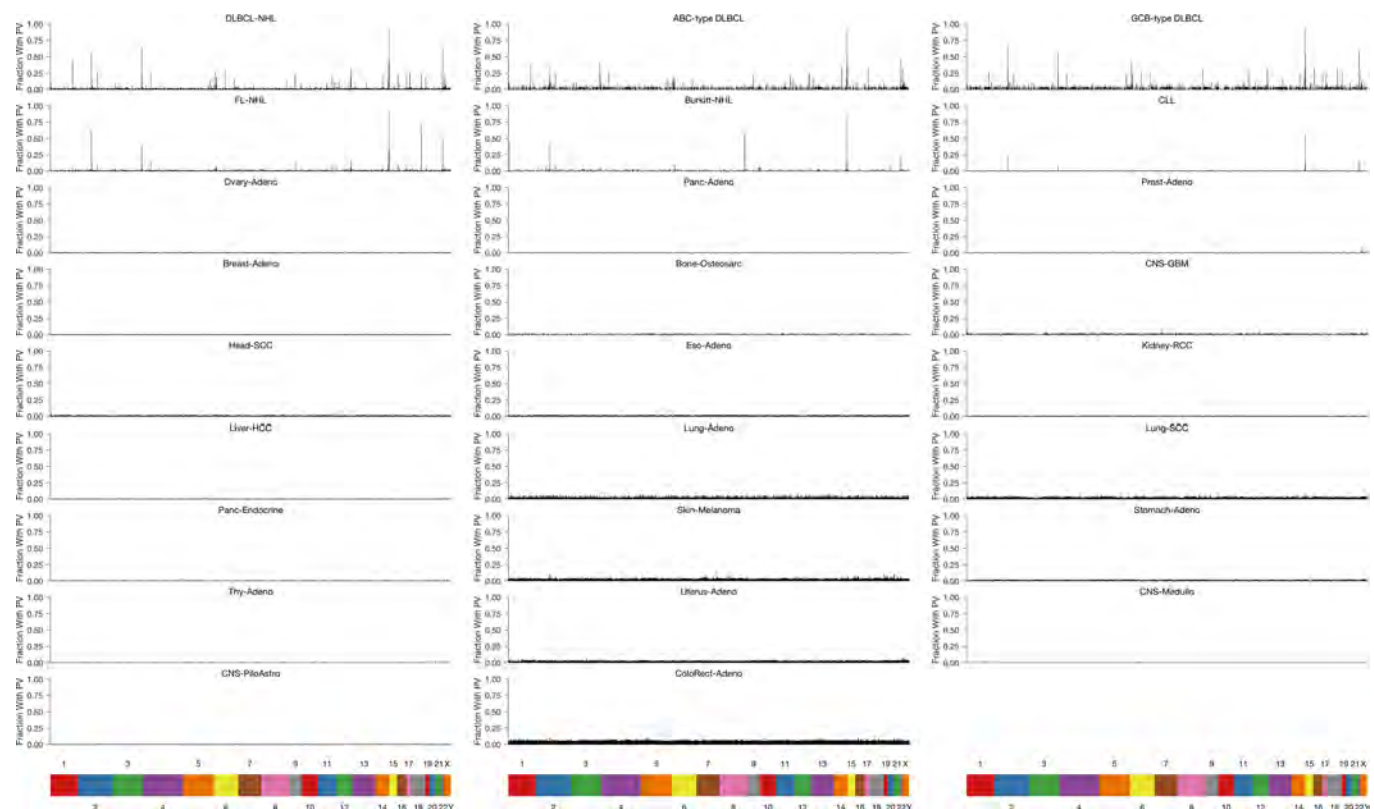
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



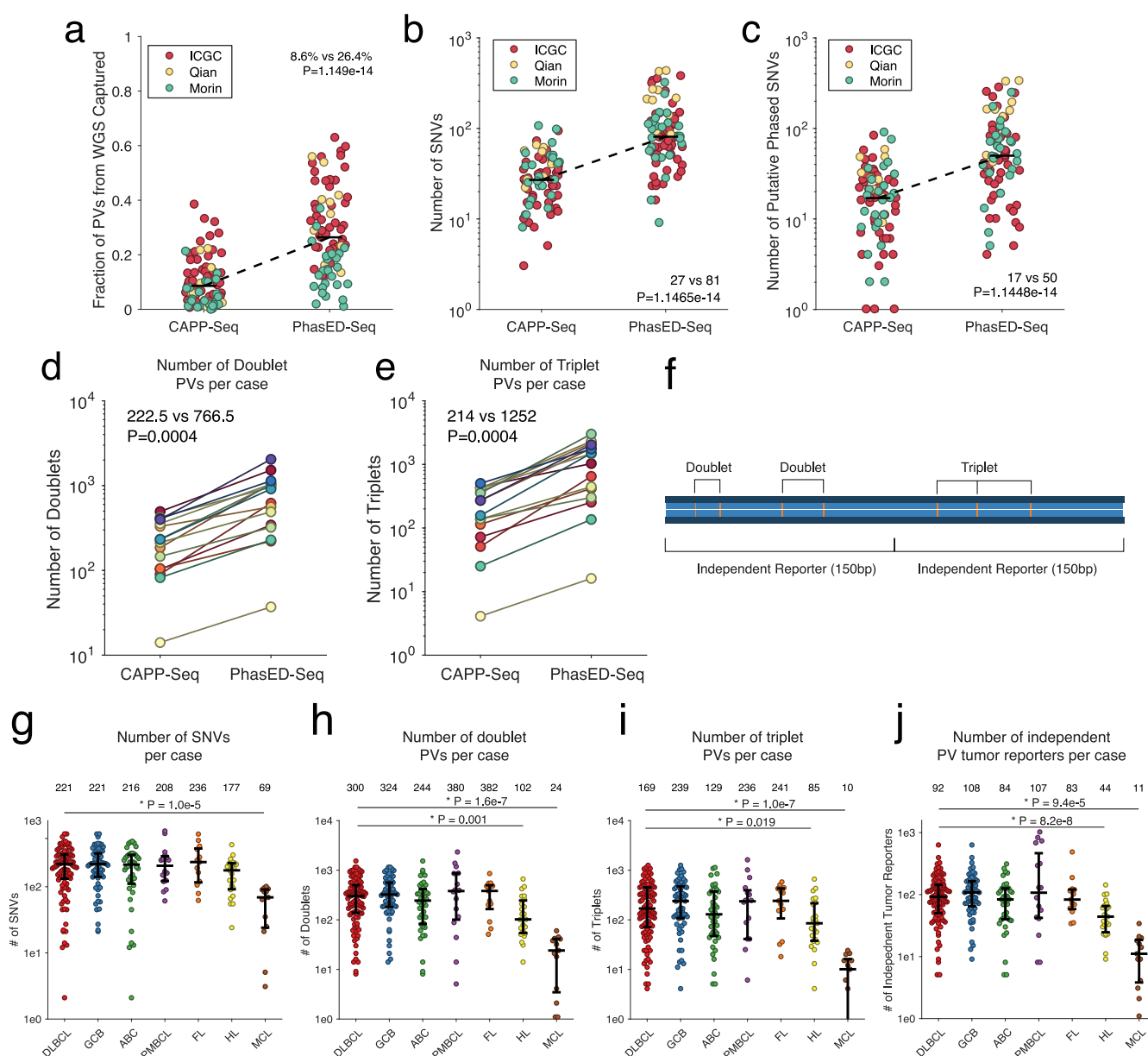
**Extended Data Fig. 1 | Comparison of duplex sequencing to phased variant sequencing. a**, A schema comparing error-suppressed sequencing by duplex sequencing vs. recovery of phased variants. In duplex sequencing, recovery of a single SNV observed on both strands of an original DNA double-helix (that is, in *trans*) is required. This requires independent recovery of two molecules by sequencing as the plus and minus strands of the original DNA molecule go through library preparation and PCR independently. In contrast, recovery of PVs requires multiple SNVs observed on the same single strand of DNA (that is, in *cis*). Thus, recovery of only the plus or the minus strand (rather than both) is sufficient for identification of PVs. **b**, A model showing the two possible reasons for limited sensitivity for ctDNA MRD assays. An assay can be limited by either having i) an insufficient number of cfDNA fragments evaluable for tumor content, or ii) an inadequate error-profile. This plot demonstrates the analytical sensitivity as the number of evaluable cfDNA fragments increase with either the amount of plasma input or the number of mutations tracked, until eventually becoming limited by the background signal (grey). Separate plots shown for single-stranded and double-stranded SNV based methods, assuming 8.92 ng cfDNA/mL plasma; 50% efficiency of library preparation, and 20% efficiency of duplex sequencing.



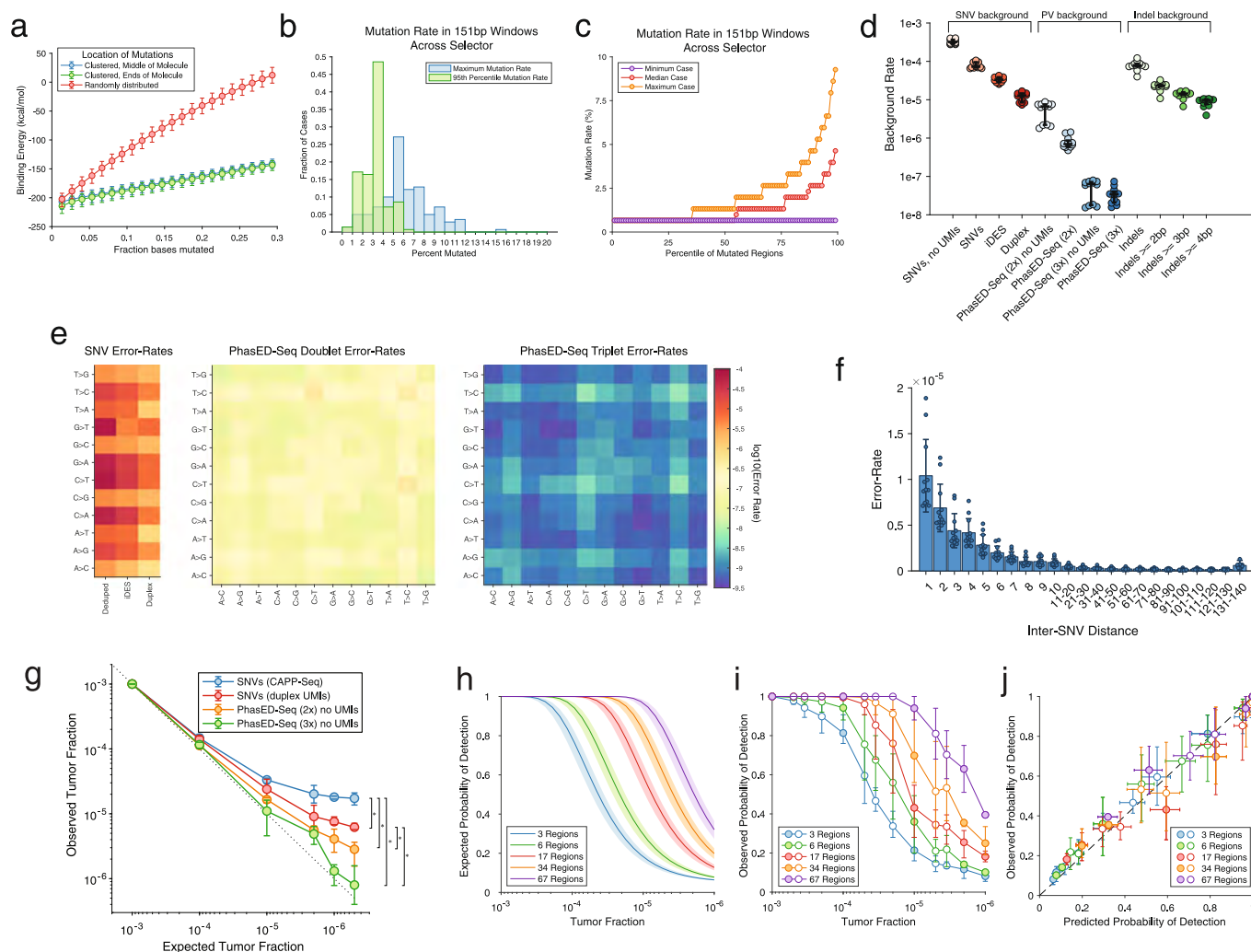
**Extended Data Fig. 2 | Enumeration of SNVs and PVs in diverse cancers from WGS.** **a–d**, Univariate scatter plots showing the number of a) SNVs, b) 2x-PVs (2 SNVs in phase), c) 3x-PVs, and d) total 2x-PVs, controlling for total number of SNVs, from WGS data for 24 different histologies of cancer. Data are presented as median and interquartile range. (FL-NHL, follicular lymphoma; DLBCL-NHL, diffuse large B cell lymphoma; Burkitt-NHL, Burkitt lymphoma; Lung-SCC, squamous cell lung cancer; Lung-Adeno, lung adenocarcinoma; Kidney-RCC, renal cell carcinoma; Bone-Osteosarc, osteosarcoma; Liver-HCC, hepatocellular carcinoma; Breast-Adeno, breast adenocarcinoma; Panc-Adeno, pancreatic adenocarcinoma; Head-SCC, head and neck squamous cell carcinoma; Ovary-Adeno, ovarian adenocarcinoma; Uterus-Adeno, uterine adenocarcinoma; Stomach-Adeno, stomach adenocarcinoma; CLL, chronic lymphocytic leukemia; ColoRect-Adeno, colorectal adenocarcinoma; Prost-Adeno, prostate adenocarcinoma; CNS-GBM, glioblastoma multiforme; Panc-Endocrine, pancreatic neuroendocrine tumor; Thy-Adeno, thyroid adenocarcinoma; CNS-PiloAstro, pilastrocytoma; CNS-Medullo, medulloblastoma.).



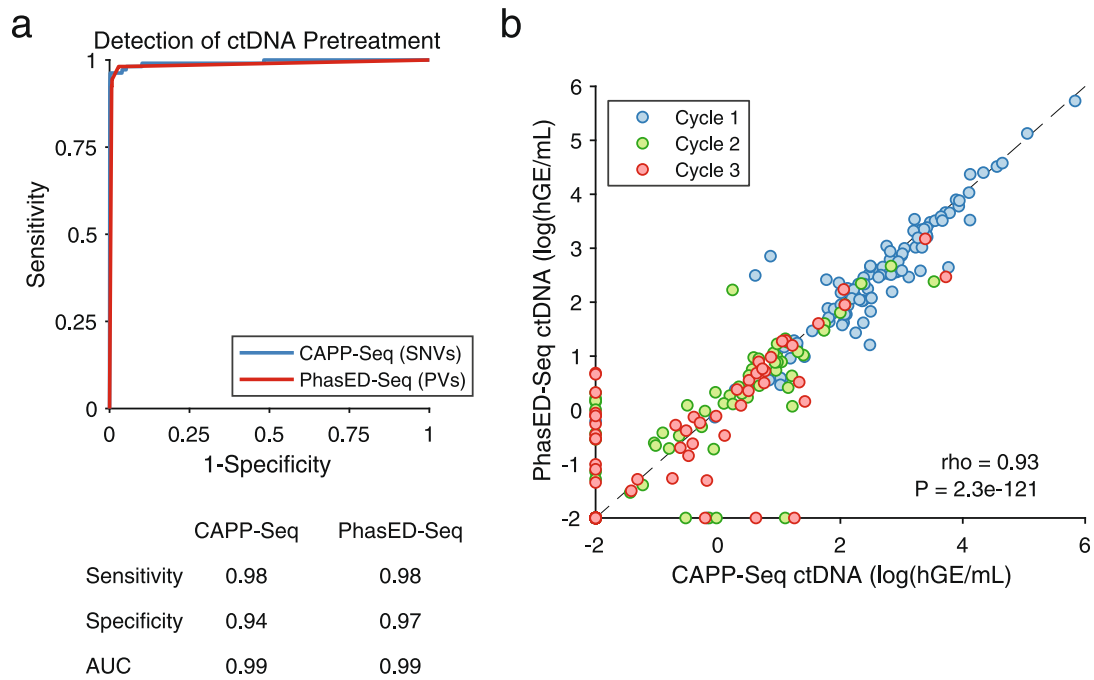
**Extended Data Fig. 3 | Distribution of PVs in stereotyped regions across the genome.** Distribution of PVs occurring in stereotyped regions across the genome of multiple cancer types. In this plot, the genome was divided into 1000 bp bins, and the fraction of samples of a given histology with a PV in each 1000 bp bin was calculated. Only bins that have at least a 2 percent recurrence frequency in any cancer subtype are shown.



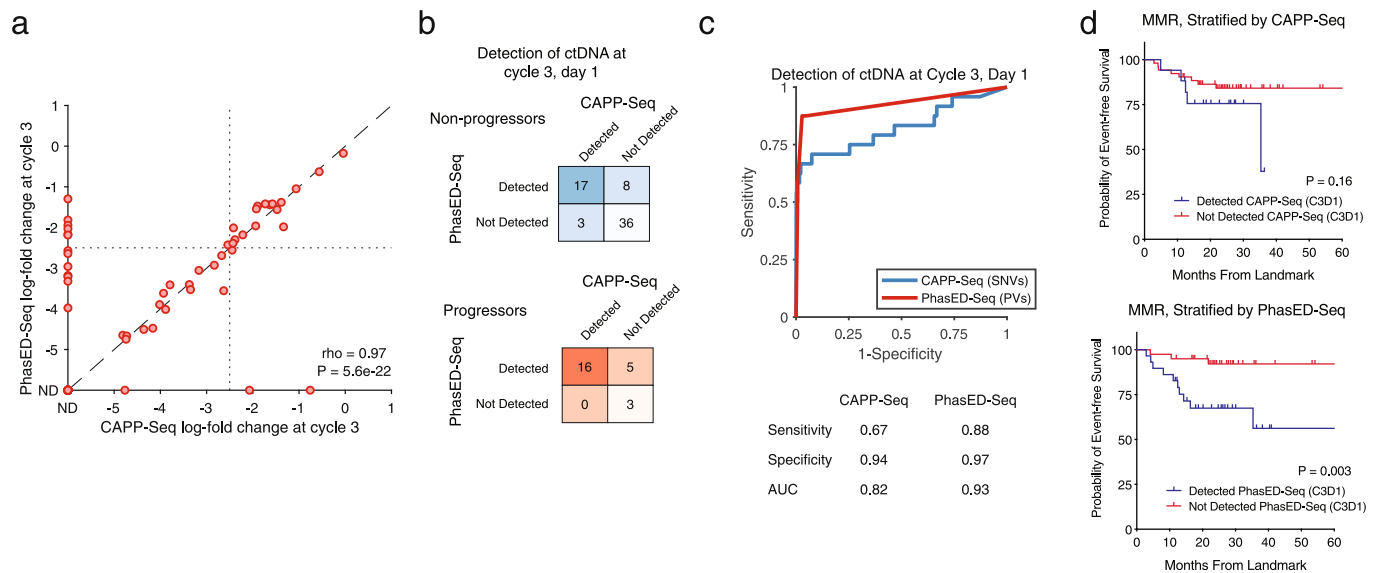
**Extended Data Fig. 4 | Performance of PhasED-Seq for recovery of PVs across lymphomas.** **a**, Univariate scatter plot comparing the fraction of all PVs across the genome identified by WGS ( $n = 79$ ) that were recovered by our previously reported lymphoma CAPP-Seq panel<sup>3</sup> (left) compared to PhasED-Seq (right). **b**, Univariate scatter plot comparing the expected yield of SNVs per case identified from WGS using a previously established lymphoma CAPP-Seq panel or the PhasED-Seq panel. **c**, Univariate scatter plot comparing the expected yield of PVs per case identified from WGS using a previously established lymphoma CAPP-Seq panel or the PhasED-Seq panel. Data from three independent publicly available cohorts are shown in **a-c**. **d,e**, Plots showing the improvement in recovery of PVs by PhasED-Seq compared to CAPP-Seq in 16 patients sequenced by both assays. This includes improvement in **d**) two SNVs in phase (that is, 2x or 'doublet PVs') and **e**) three SNVs in phase (3x or 'triplet PVs'). Statistical testing in panels **a-e**) performed by 2-sided Wilcoxon signed-rank test. **f**, A cartoon describing the terminology for phased variants in this manuscript. The figure shows one region of an individual's cancer genome (300 bp). Phased variants on a single strand of DNA can occur with different numbers of SNVs, including 2 variants in phase (doublets) and 3 in phase (triplets). For the purpose of detecting ctDNA, 'independent reporters' are defined as PVs that will typically co-segregate on separate cfDNA molecules, resulting in independent evaluable fragments. Given the size of cfDNA molecules, these are separated in 150 bp regions. **g-j**, These panels show the number of SNVs and PVs identified for patients with different types of lymphomas. These panels show the number of **g**) SNVs, **h**) doublet PVs, **i**) triplet PVs, and **j**) independent PV reporters; bars represent median and interquartile range. \*,  $P < 0.05$  by two-sided Wilcoxon rank sum test; comparisons only shown for all histologies vs DLBCL. (DLBCL, diffuse large B-cell lymphoma; GCB, germinal center B-cell like DLBCL; ABC, activated B-cell like DLBCL; PMBCL, primary mediastinal B-cell lymphoma; FL, follicular lymphoma; HL, Hodgkin lymphoma; MCL, mantle cell lymphoma).



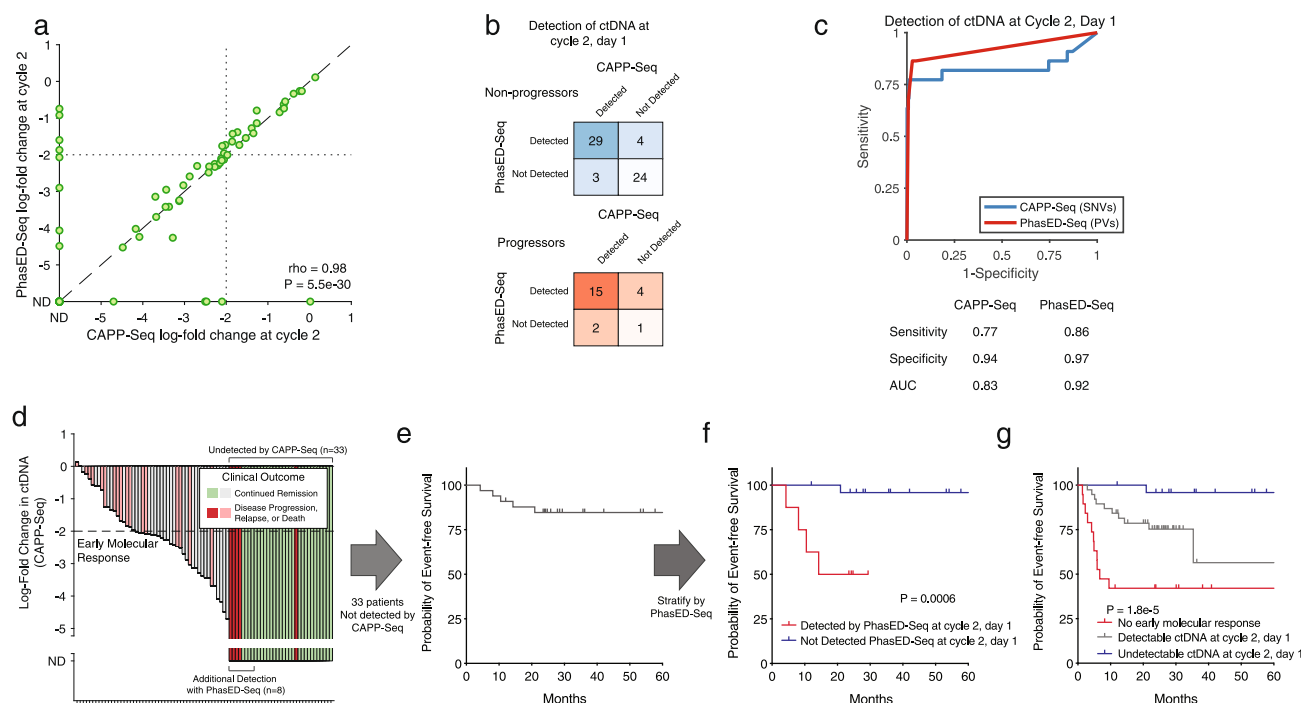
**Extended Data Fig. 5 | Technical aspects of PhasED-Seq by hybrid-capture sequencing.** **a**, Theoretical binding energy for 150-mers across the genome. Mutations were either clustered to one end (green), clustered in the middle (blue), or randomly throughout the sequence (red). Data represent the median and IQR from 10,000 *in silico* simulations. **b**, Histograms of summary metrics of the mutation rate of 151-bp windows from all patients in this study. **c**, The percentile of mutation rate across all mutated 151-bp windows across all patients in this study. **d**, Rate of background-signal in the PhasED-Seq panel for multiple variants, including SNVs (red), PVs (blue), and indels (green). Different methods of error-suppression for each variant type are shown. Bars represent median and IQR. UMIs, unique molecular identifiers; PhasED-Seq 2x, doublet PVs; PhasED-Seq 3x, triplet PVs. **e**, Error-rate for SNVs (left), doublet PVs (middle), and triplet PVs (right) by type of mutation. For triplet PVs, the x and y-axis represent the first and second type of base alteration in the PV. **f**, Error rate for doublet PVs across  $n=12$  healthy cfDNA samples as a function of inter-SNV distance. Data show mean and standard deviation. **g**, Limiting dilution series simulating cfDNA similar to Fig. 5a; cfDNA from 3 independent patient samples were used in each dilution. In this plot, PhasED-Seq is assessed without the use of UMIs. Data are presented as mean and range. \*,  $P < 0.05$ ; CAPP-Seq vs duplex,  $P = 3.2 \times 10^{-5}$ ; CAPP-Seq vs PhasED-Seq (2x),  $P = 1.6 \times 10^{-4}$ ; CAPP-Seq vs PhasED-Seq (3x),  $P = 1.9 \times 10^{-5}$ ; duplex vs PhasED-Seq (2x), 0.017; duplex vs PhasED-Seq (3x), 0.0046. **h**, Theoretical rate of detection for a sample with a given number of PV-containing regions, according to binomial sampling, assuming unique sequencing depth of 4000–6000x (shaded area; 5000x shown as line). **i**, Observed rate of detection given a true tumor fraction, with varying numbers of PV-containing regions. Filled-in points represent 'wet' experiments; open points represent *in silico* dilution experiments. Data represent mean and range. **j**, Predicted vs observed rate of detection for samples from the dilution series shown in panels **h** and **i**. Error-bars are as described in **h** and **i** above (see the Supplementary Methods).



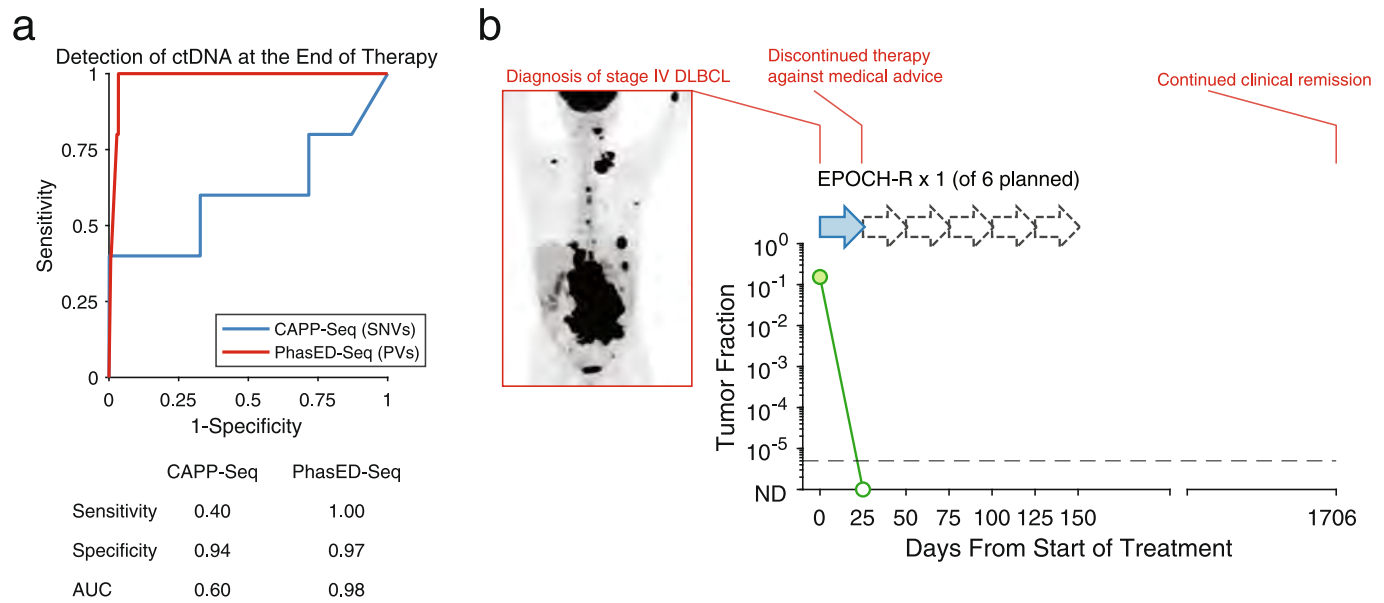
**Extended Data Fig. 6 | Comparison of ctDNA quantitation by PhasED-Seq to CAPP-Seq and clinical applications. a**, ROC curve of the performance for detection of ctDNA from SNVs (that is, CAPP-Seq) and PVs using PhasED-Seq. Positive samples are 107 pretreatment plasmas, negative samples are 40 control plasmas assessed for evidence of ctDNA using 107 personalized mutation lists for 4,280 total samples. Sensitivity and specificity at optimum point and AUC are shown. **b**, Quantity of ctDNA (measured as  $\log_{10}(\text{haploid genome equivalents/mL})$ ) as measured by CAPP-Seq vs. PhasED-Seq in individual samples. Samples taken prior to cycle 1 of RCHOP therapy (that is, pretreatment), prior to cycle 2, and prior to cycle 3, are shown in independent colors (blue, green, and red respectively; 277 total samples). Undetectable levels fall on the axes. Spearman correlation and P-value are shown.



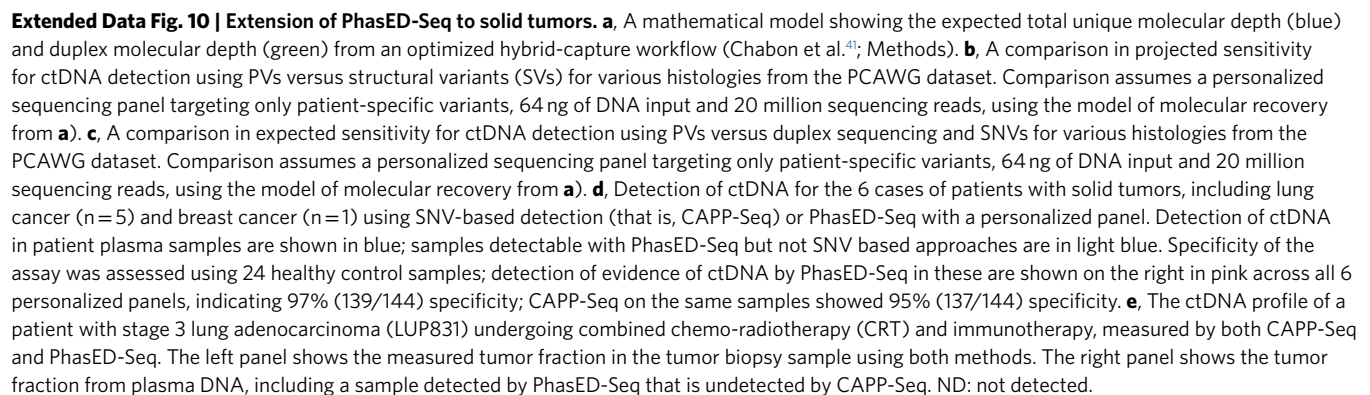
**Extended Data Fig. 7 | Detection of ctDNA after two cycles of systemic therapy.** **a**, Scatterplot shows the log-fold change in ctDNA after 2 cycles of therapy measured by CAPP-Seq or PhasED-Seq for patients receiving RCHOP therapy. Dotted lines show the previously established threshold of a 2.5-log reduction in ctDNA for molecular response. Undetectable samples fall on the axes; the correlation coefficient represents a Spearman rho for the samples detected by both CAPP-Seq and PhasED-Seq. **b**, Detection rate of ctDNA samples after 2 cycles of therapy by PhasED-Seq vs CAPP-Seq. Patients with eventual disease progression are shown in red, while patients without eventual disease progression are shown in blue. **c**, ROC curve for detection of ctDNA after 2 cycles of treatment. Positive samples include 24 samples from patients with eventual disease progression, and therefore are known to have residual disease. Negative samples are from 4,280 tests on healthy controls as described in Extended Data Fig. 6a. **d**, Kaplan-Meier plots and two-sided log-rank test showing the event-free survival of 69 patients achieving an MMR stratified by ctDNA detection with CAPP-Seq (top) or PhasED-Seq (bottom).



**Extended Data Fig. 8 | Detection of ctDNA after one cycle of systemic therapy.** **a**, Scatterplot showing the log-fold change in ctDNA after 1 cycle of therapy measured by CAPP-Seq or PhasED-Seq for patients receiving RCHOP therapy. Dotted lines show the previously established threshold of a 2-log reduction in ctDNA for molecular response. Undetectable samples fall on the axes; the correlation coefficient represents a Spearman rho for the samples detected by both CAPP-Seq and PhasED-Seq. **b**, Detection rate of ctDNA samples after 1 cycle of therapy by PhasED-Seq vs CAPP-Seq. Patients with eventual disease progression are shown in red, while patients without eventual disease progression are shown in blue. **c**, ROC curve for detection of ctDNA after 2 cycles of treatment. Positive samples include 22 samples from patients with eventual disease progression, and therefore are known to have residual disease. Negative samples are from 4,280 tests on healthy controls as described in Extended Data Fig. 6a. **d**, Waterfall plot showing the change in ctDNA levels measured by CAPP-Seq after 1 cycle of first-line therapy in patients with DLBCL. Patients with undetectable ctDNA by CAPP-Seq are shown as 'ND' ('not detected'), in darker colors. The colors of the bars also indicate the eventual clinical outcomes for these patients. **e**, A Kaplan-Meier plot showing the event-free survival for 33 DLBCL patients with undetectable ctDNA measured by CAPP-Seq after 1 cycle of therapy. **f**, A Kaplan-Meier plot and two-sided log-rank test showing the event-free survival of 33 patients shown in **f** (undetectable ctDNA by CAPP-Seq) stratified by ctDNA detection via PhasED-Seq at this same time-point (cycle 2, day 1).  $P = 0.0006$ . **g**, A Kaplan-Meier plot and two-sided log-rank test showing the event-free survival for 82 patients with DLBCL stratified by ctDNA at cycle 2, day 1 separated into 3 strata - patients failing to achieve an early molecular response (red), patients with an early molecular response who still have detectable ctDNA by PhasED-Seq and/or CAPP-Seq (grey), and patients who have a stringent molecular remission (undetectable ctDNA by PhasED-Seq and CAPP-Seq; blue).  $P = 1.8e-5$ .



**Extended Data Fig. 9 | Performance of ctDNA detection at the end of systemic therapy.** **a**, ROC curve for detection of ctDNA after the completion of planned systemic therapy. Positive samples include 5 samples from patients with eventual disease progression, and therefore are known to have residual disease. Negative samples are from 4,280 tests on healthy controls as described in Extended Data Fig. 6a. **b**, The ctDNA profile of a patient with stage 4 DLBCL undergoing systemic chemotherapy, with pretreatment PET scan shown on the left. This patient only received one cycle of EPOCH-R chemotherapy from 6 planned treatments (dashed arrows – planned therapy that was not given). Following this, the patient self-discontinued treatment. This patient was found to have cleared their ctDNA by PhasED-Seq and continues in clinical remission after > 4 years.



## 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 our [Editorial Policies](#) 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.

- |     |           |
|-----|-----------|
| n/a | Confirmed |
|-----|-----------|
- ☐ ☒ 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  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
  - ☐ ☒ 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

Sequencing libraries were prepared as described in the methods and sequenced on the Illumina HiSeq4000 platform using 2x150bp paired end reads. Primary processing of sequencing data was performed with BWA (0.5.9-r16), samtools (versions 0.1.18 and 1.3.1), and GATK (v3.8-1-0-gf15c1c3ef). Variant calling was performed by VarScan2 (v2.3.9), Mutect (v1.1.7), and Strelka2 (v2.9.1) with variant annotation by annovar (version 2018Apr16) and Oncotator (v1.9.8.0).

#### Data analysis

Statistical analyses were performed using MATLAB (R2018a), R (version 3.4.1), and perl (v5.10.1) in a CentOS 6.7 Linux environment or with GraphPad Prism (version 8.0.2). The contribution of known mutational processes to phased and isolated SNVs from WGS data was assessed with the deconstructSigs R package using the COSMIC signature set (v2) as described in the methods. WGS of primary Calculation of AUC accounting for survival and censorship was performed using the R 'survivalROC' package version 1.0.3 with default settings. The PhasED-Seq software was developed using C++17. PhasED-Seq v.1.0 was used to generate the results in this work and is freely available for academic research use at <http://phasedseq.stanford.edu>.

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 and 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 count of SNVs and phased variants from WGS data for all cancer types analyzed, along with the distribution of phased variants from WGS in lymphoid malignancies, is provided in the Supplementary Tables. The coordinates of regions of the genome with enriched phased variants from WGS are also provided in the Supplementary Tables. Anonymized and summarized clinical and demographic data from patients where PhasED-Seq was used for disease detection are also provided. Whole genome sequencing data was obtained from the Pan-Cancer Analysis of Whole Genomes (PCAWG) from <https://dcc.icgc.org/pcawg>. Additional lymphoma WGS was obtained from the International Cancer Genome Consortium website (malignant lymphoma project, MALY-DE), as well as from DOI: 10.1182/blood-2013-02-48372733 and 10.1016/j.cell.2014.11.01334. Raw data for samples used in the dilution series is available at <http://phasedseq.stanford.edu>. This website provides users with software for phased-variant monitoring (PhasED-Seq v.1.0). Included here are the list of SNVs and PVs used to detected tumor DNA from each of the three samples used in this dilution series, as well as the sequencing files for each concentration condition. Owing to restrictions related to dissemination of germline sequence information included in the informed consent forms used to enroll study subjects, we are unable to provide access to other raw sequencing data. Reasonable requests for additional data will be reviewed by the senior authors to determine whether they can be fulfilled in accordance with these privacy restrictions. Requests for additional materials related to this work should be directed to A.A.A.

## 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.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

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

|                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Sample size     | The goal of our study was to determine if a more sensitive ctDNA assay has additional prognostic value in large B-cell lymphoma patients with undetectable ctDNA by current methodologies (i.e., CAPP-Seq) after 1 or 2 cycles of standard immuno-chemotherapy. We powered our study to detect a meaningful difference (35% difference in event-free survival at 24 months) between PhasED-Seq detectable and undetectable patients. A sample size of 34 patients is needed to detect this difference with 80% power and alpha=0.05 (assuming a 1:1 ratio between detectable and undetectable patients by PhasE-Seq). However, these 34 patients must have undetectable ctDNA by CAPP-Seq after 1 or 2 cycles of therapy. Previous literature suggests that ~50% of patients will have undetectable ctDNA by CAPP-Seq at these milestones, therefore ~78 total subjects are required to achieve the necessary sample size. This justifies our sample size of 82 subjects with cfDNA available after 1 cycle of therapy and 88 with cfDNA after 2 cycles of therapy. |
| Data exclusions | Samples deemed to be technical failures during DNA isolation or library preparation were excluded. Phased variants were filtered using the PhasED-Seq algorithm as described in the Methods sections "Identification of phased variants and allelic quantitation" and "Genotyping phased variants from pretreatment samples".                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Replication     | Demonstration of the lower error-profile from phased variants as compared to SNVs was performed in 12 independent samples with consistent results. Hybridization experiments were performed in three replicates. Disease detection in the limiting dilution series was performed using three independent patient-derived cell-free DNA samples. Results describing the clinical significance of ctDNA detection using PhasED-Seq in the setting of undetectable ctDNA by alternative methods (such as CAPP-Seq) were replicated with samples after both one cycle or two cycles of therapy, as well as at the end of therapy (3 sample-sets at different time-points n=82, 88, and 19 samples). The performance of PhasED-Seq in cancers outside of B-cell lymphomas was shown from 6 patients with solid tumors. All attempts at replication were successful.                                                                                                                                                                                                      |
| Randomization   | Patients and controls were enrolled in the study based on a confirmed diagnosis of B-cell lymphoma receiving standard first-line immuno-chemotherapy. Additional subjects were enrolled based on a confirmed diagnosis of lung cancer or breast cancer. Therefore, randomization of individuals to different groups is not applicable as there was no intervention.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| Blinding        | As there was no therapeutic intervention in this observational biomarker study, no blinding was 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 &amp; experimental systems

|                                     |                                                                 |
|-------------------------------------|-----------------------------------------------------------------|
| n/a                                 | Involved in the study                                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies                             |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines                  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology          |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms            |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> Human research participants |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data                          |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern           |

## Methods

|                                     |                                                 |
|-------------------------------------|-------------------------------------------------|
| n/a                                 | Involved in the study                           |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry         |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

## Human research participants

Policy information about [studies involving human research participants](#)

## Population characteristics

Study participants include patients with B-cell lymphomas, including diffuse large B-cell lymphoma, primary mediastinal B-cell lymphoma, Hodgkin lymphoma, follicular lymphoma, and mantle cell lymphoma. Blood samples were collected at Stanford University (USA), MD Anderson Cancer Center (USA), the National Cancer Institute (USA), University of Eastern Piedmont (Italy), Essen University Hospital (Germany), CHU Dijon (France). To extend PhasED-Seq to solid tumors, additional subjects with a diagnosis of lung cancer or breast cancer were collected at Stanford University (USA). The median age was 57 (range 20-84), and 47% of participants were female. Additional clinical and demographic details are provided with the paper.

## Recruitment

All subjects were enrolled with Institutional Review Board-approved protocols and all biospecimens analyzed in this study were collected with informed consent. Patients receiving first-line therapy for large B-cell lymphomas at these centers were enrolled separately at each institution for observational study of blood-based biomarkers with serial blood samples collected and stored locally. Patients were treated with first-line therapy according to local standards. To extend PhasED-Seq to solid tumors, subjects with a diagnosis of limited stage (stage I-III) lung cancer or breast cancer undergoing treatment at Stanford University were enrolled for observational study of blood-based biomarkers with serial blood samples collected and stored locally.

These patients represent a typical population receiving cancer care at tertiary academic medical centers, which could differ from other patient populations.

## Ethics oversight

All samples analyzed in this study were collected with informed consent from subjects enrolled on Institutional Review Board approved protocols that complied with all relevant ethical regulations at their respective centers, including Stanford University, MD Anderson Cancer Center, the National Cancer Institute, University of Eastern Piedmont, Essen University Hospital, and CHU Dijon.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

---

## **Supplementary information**

---

# **Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA**

---

In the format provided by the  
authors and unedited

## **Table of Contents**

### **Supplementary Figures**

|                                                                                                   |          |
|---------------------------------------------------------------------------------------------------|----------|
| <b>Figure S1. Quantity and genomic location of PVs from WGS in lymphoid malignancies.</b>         | <b>2</b> |
| <b>Figure S2. Differences in PVs between lymphoma types in mutations in the <i>IGH</i> locus.</b> | <b>4</b> |

|                                                          |          |
|----------------------------------------------------------|----------|
| <b><u>Supplementary Tables (attached separately)</u></b> | <b>5</b> |
|----------------------------------------------------------|----------|

|                                     |          |
|-------------------------------------|----------|
| <b><u>Supplementary Methods</u></b> | <b>7</b> |
|-------------------------------------|----------|

|                                          |          |
|------------------------------------------|----------|
| <b>Identification of phased variants</b> | <b>7</b> |
|------------------------------------------|----------|

*Identification of phased variants and allelic quantitation*

*Genotyping phased variants from pretreatment samples*

*Determination of tumor fraction in a sample from phased variants*

*Monte Carlo simulation for empirical significance of PV detection within a specimen*

*Assessment of specificity of PhasED-Seq*

*Calculation of error rates*

|                                                                 |           |
|-----------------------------------------------------------------|-----------|
| <b>Differences in phased variants between lymphoma subtypes</b> | <b>11</b> |
|-----------------------------------------------------------------|-----------|

|                                                                                         |           |
|-----------------------------------------------------------------------------------------|-----------|
| <b>Model to predict the probability of detection for a given set of phased variants</b> | <b>13</b> |
|-----------------------------------------------------------------------------------------|-----------|

|                                                                                           |           |
|-------------------------------------------------------------------------------------------|-----------|
| <b>Model to assess theoretical sensitivity of ctDNA by tracking PVs vs other variants</b> | <b>15</b> |
|-------------------------------------------------------------------------------------------|-----------|

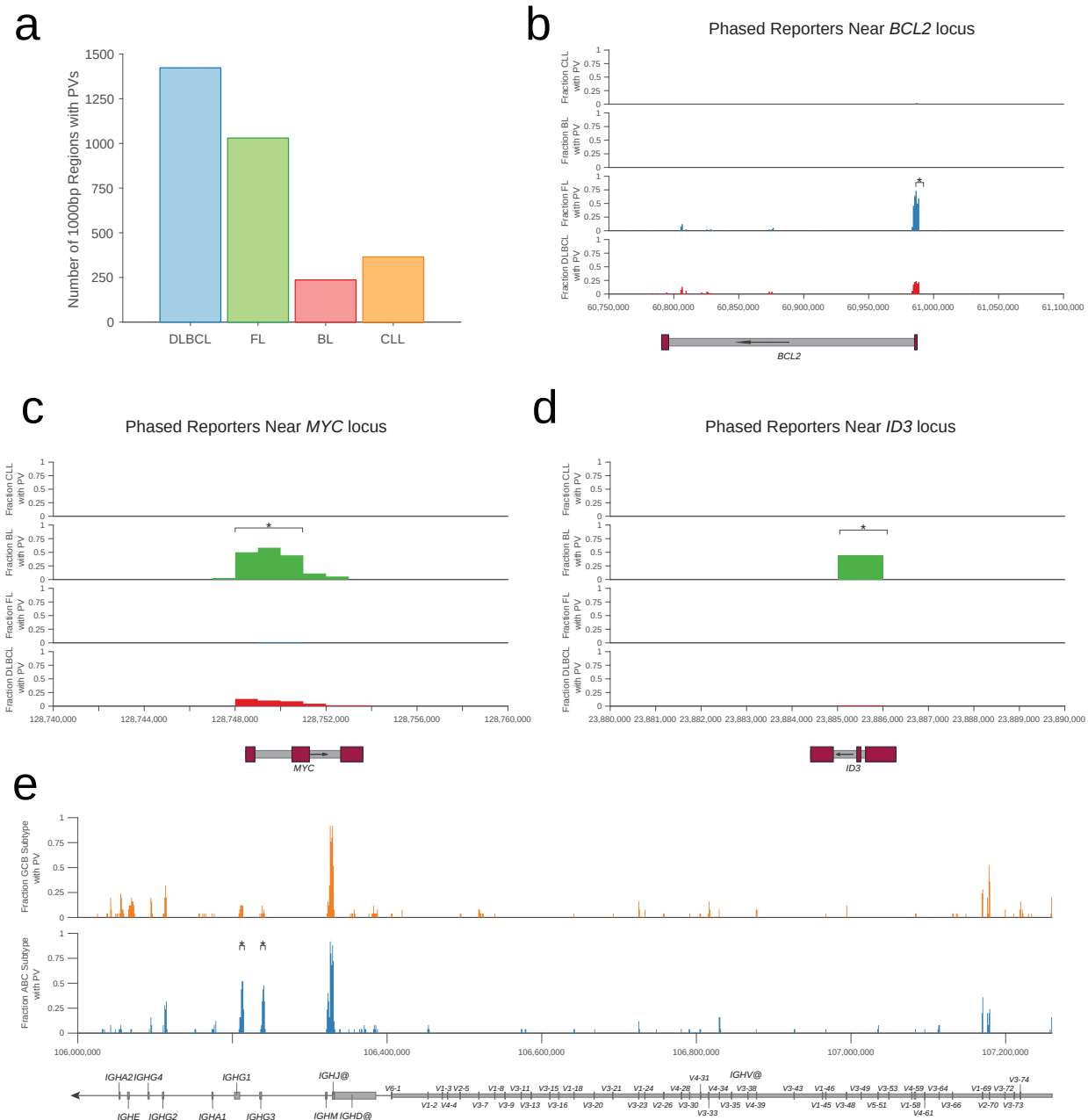


**Figure S1. Quantity and genomic location of PVs from WGS in lymphoid malignancies.**

a) Bar plot showing the number of independent 1000bp regions across the genome that recurrently contain PVs for DLBCL, FL, BL, and CLL (n=68, 74, 36, and 151 respectively).

b-d) Plots showing the frequency of PVs for multiple lymphoid malignancies with relationships to specific genetic loci, including b) *BCL2*, c) *MYC*, and d) *ID3*. The location of the transcript for a given gene is shown below the plot in grey; exons are shown in red. \* indicates regions with

significantly more PVs in a given cancer histology compared to all other histologies by Fisher's Exact Test ( $P < 0.05$ ).

e) Similar to b-d), these plots show the frequency of PVs across lymphoma subtypes. Here, we show the @*IGH* locus, consisting of *IGHV*, *IGHD*, and *IGHJ* parts, for ABC and GCB subtype DLBCLs ( $n=25$  and  $25$ , respectively). Coding regions for Ig parts, including Ig-constant regions and V-genes, are shown. \* indicates regions with significant difference in PVs between ABC and GCB subtypes by Fisher's Exact Test ( $P < 0.05$ ). (DLBCL, diffuse large B-cell lymphoma; FL, follicular lymphoma; BL, Burkitt lymphoma, CLL, chronic lymphocytic leukemia)

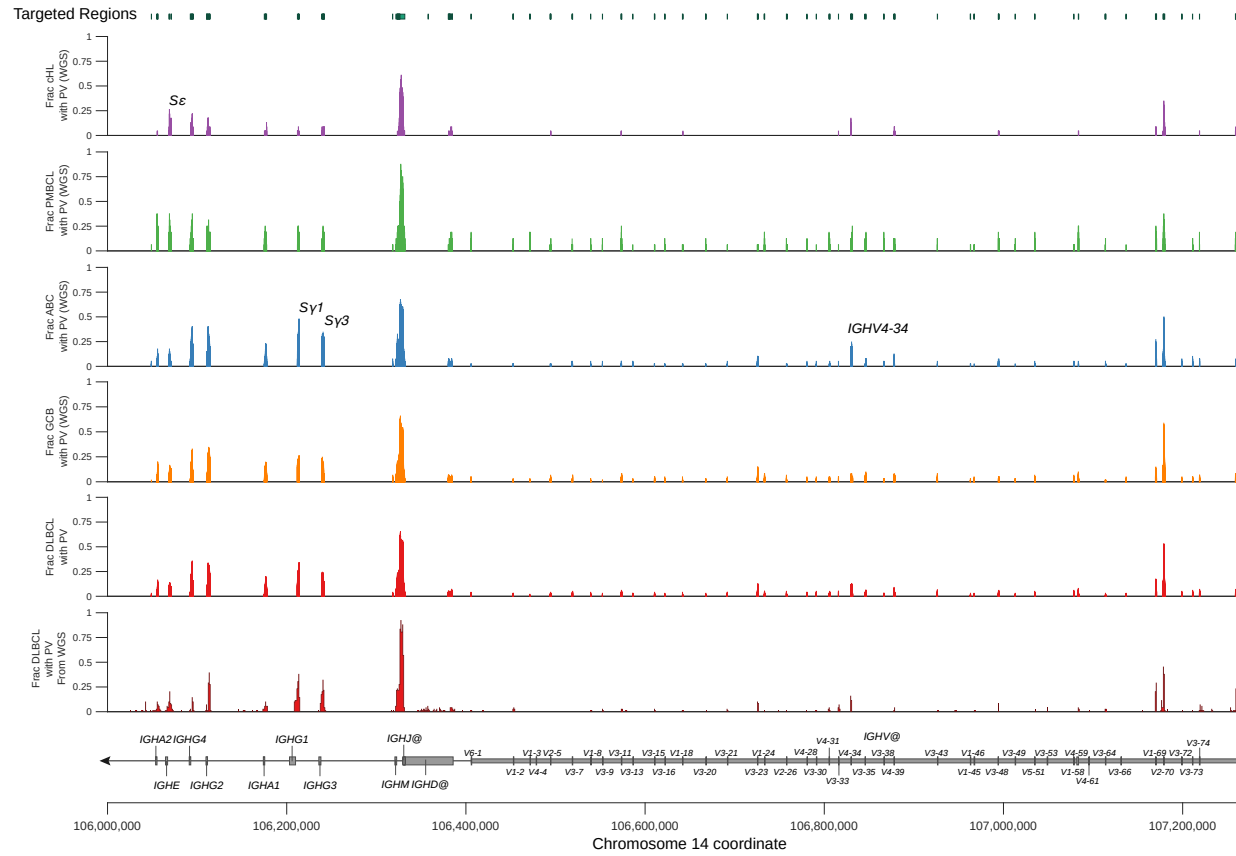


**Figure S2. Differences in PVs between lymphoma types in mutations in the *IGH* locus.**

This figure shows the frequency of PVs from PhasED-Seq across the @*IGH* locus for different types of B-cell lymphomas. The bottom track (grey) shows the structure of the @*IGH* locus and gene-parts, including Ig-constant genes and V-genes. The next (red, outlined) track shows the frequency of PVs in this genomic region from WGS data (ICGC cohort). The remainder of the tracks show the frequency of PVs from PhasED-Seq targeted sequencing data, including 1) DLBCL (red), GCB-DLBCL (orange), ABC-DLBCL (blue), PMBCL (green), and HL (purple). The regions targeted by the PhasED-Seq panel are shown at the top. Selected immunoglobulin parts with PVs enriched in specific histologies are labeled (i.e., *IGHV4-34*, *S $\epsilon$* , *S $\gamma$ 3* and *S $\gamma$ 1*).

## **Supplementary Tables (Attached in separate Excel file)**

### **Table S1. WGS data for discovery of phased variants.**

This table contains the relevant samples and data sources from PCAWG and ICGC used for discovery of putative phased variants (PVs) in WGS from 2538 cases across 24 histologies of cancer. The number of SNVs (numSNVs), putative PVs for both two and three variants in phase (numPVs 2x, numPVs 3x), and enrichment of PVs, defined as number of doublet PVs / number of SNVs (pvEnrichment) are also shown.

### **Table S2. Regions of recurrent PVs in lymphoid malignancies.**

This table shows the frequency of putative PVs in 1-kb regions across the genome in lymphoid malignancies. Only regions with >1 subject containing a putative PV are shown.

### **Table S3. Technical comparison between CAPP-Seq and PhasED-Seq.**

This table shows the sequencing depth, SNV counts, and PV counts for samples sequenced by both CAPP-Seq and PhasED-Seq panels.

### **Table S4. Clinical characteristics of cases used for PV enrichment across PhasED-Seq panel.**

### **Table S5. Enrichment of PVs across regions of PhasED-Seq panel according to lymphoma subtype.**

This table shows the enrichment in PVs at specific genetic loci for lymphoma subtypes throughout the PhasED-Seq panel. Genetic loci were determined by separating the panel into 50bp bins and assigning each bin to a specific gene based on the nearest GENCODE gene annotation.

### **Table S6. Oligonucleotides used for hybridization bias experiment.**

This table provides the sequences of oligonucleotides synthesized to assess hybridization and molecular recovery bias with increasing mutational burden.

### **Table S7. Technical assessment of PhasED-Seq limit of detection through limiting dilution series.**

This table provides the sequencing characteristics including deduplicated read count and depth for samples used in the limiting dilution experiment establishing the limit of detection for PhasED-Seq. Three separate patient cfDNA samples were spiked into healthy control cfDNA and serially diluted to expected tumor content levels of 1 in 1,000 to 1 in 2,000,000. The allele fraction of each sample assessed by standard CAPP-Seq, duplex sequencing, and PhasED-Seq using doublets and triplets, is also provided. Allele fractions were normalized to the tumor fraction of the highest sample concentration (i.e., the 1:1,000 sample). Background detection of patient-specific PV reporters in control samples from 12 standard depth and 1 deep control sample are also shown.

**Table S8. Clinical characteristics of cases used for evaluation of MRD performance by lymphoma PhasED-Seq.**

This table provides the clinical characteristics of lymphoma patient samples used for assessment of MRD detection at cycle 2 day 1, cycle 3 day 1, and/or end of therapy.

**Table S9. Characteristics of cases used for personalized PhasED-Seq in solid tumors.**

Clinical characteristics of the cases used for extension of PhasED-Seq to solid tumors via personalized approach. The stage, relevant treatment, and plasma samples utilized are also shown. All samples were assessed for ctDNA by SNV-based CAPP-Seq for comparison as well.

## Supplementary Methods

### Identification of phased variants

#### *Identification of phased variants and allelic quantitation*

After generating UMID error-suppressed alignment files (i.e., BAM files), we identified PVs from each sample as follows. First, matched germ-line sequencing of uninvolved peripheral blood mononuclear cells (PBMCs) was performed to identify patient-specific constitutional single nucleotide polymorphisms (SNPs). These were defined as non-reference positions with a variant allele fraction (VAF) above 40% with a depth of at least 10, or a VAF of above 0.25% with a depth of at least 100. Next, we identified PVs from read-level data for a sample of interest. Following UMID-mediated error suppression, we analyzed each individual paired-end (PE) read and identified all non-reference positions using 'samtools calmd'. We used PE data rather than single reads to identify variants occurring on the same template DNA molecule, which could subsequently fall into either read 1 or read 2. Any read-pair containing  $\geq 2$  non-reference positions was considered to represent a possible somatic PV. For reads with  $> 2$  non-reference positions, each permutation of size  $\geq 2$  was considered independently: i.e., if 4 non-reference positions were identified in a read-pair, all combinations of 2 SNVs (i.e., 'doublet' phased variants) and all combinations of 3 SNVs (i.e., 'triplet' phased variants) were independently considered. PVs containing putative germ-line SNPs were also removed as follows: if in a given  $n$ -mer (i.e.,  $n$  SNVs in phase on a given molecule)  $\geq n-1$  of the component variants were identified as germ-line SNPs, the PV was redacted. This filtering strategy ensures that for any remaining PV, at least 2 of the component SNVs were not seen in the germ-line, as relevant for both sensitivity and specificity.

We further filtered putative somatic PVs using a heuristic blacklisting approach in considering sequencing data from 170 germ-line DNA samples serving as controls. In each of these samples, we identified PVs on read-pairs as described above, but without filtering for matched germ-line. Any PV that occurred in one or greater paired-end read, in one or more of these control samples, was included in the blacklist and removed from patient-specific somatic PV lists.

To calculate the VAF of each PV, we calculated a numerator representing the number of DNA molecules containing a PV of interest over a denominator representing the total number of DNA molecules that covered the genomic region of interest. That is, the numerator is simply the total

number of deduplicated read-pairs that contain a given PV while the denominator is the number of read-pairs that span the genomic locus of a given PV.

#### *Genotyping phased variants from pretreatment samples*

The above strategy resulted in a list of PVs of  $\geq 1$  read-depth in each sample. To identify PVs serving as tumor-specific somatic reporters for disease monitoring, for each case we identified a 'best genotyping' specimen – either DNA from a tumor tissue biopsy (preferred), or pretreatment cell-free DNA. After identifying all possible PVs in the 'best genotyping sample', we further filtered the list for specificity as follows. For any  $n$ -mer PV set, if  $\geq n-1$  of the constituent SNVs were present as germ-line SNPs in the 170 control samples described above, the PV was removed. Furthermore, only PVs that meet the following criteria were considered: 1) AF > 1%; 2) depth of the PV locus of  $\geq 100$  read-pairs, 3) at least one component SNV in the on-target space, 4) zero read-support in matched germline. Finally, 5) any PV meeting these criteria was assessed for read-support in a cohort of 12 healthy control cfDNA samples. If any read-support was present in >1 of these 12 samples, the PV was removed. For genotyping from cell-free DNA samples identified as low tumor fraction by SNVs (i.e., < 1% mean AF across all SNVs), the AF threshold for determining PVs was relaxed to >0.2%. This filtering resulted in the PV lists used for disease monitoring and MRD detection.

#### *Determination of tumor fraction from phased variants*

When evaluating a sample for minimal residual disease (MRD) detection with prior knowledge of the tumor genotype, we assessed for the presence of any PV identified in the best pretreatment genotyping sample in the MRD sample of interest. Given a list of  $k$  possible tumor-derived PVs observed in the best genotyping sample, we first determined all read-pairs covering at least 1 of the  $k$  possible PVs. This value,  $d$ , can be thought of as the aggregated 'informative depth' across all PVs spanned by cfDNA molecules in a PhasED-Seq experiment<sup>1</sup>. We then assessed how many of these  $d$  read-pairs actually contained 1 or more of the  $k$  possible PVs – this value,  $x$ , represents the number of tumor-derived molecules containing somatic PVs in a given sample. The number of tumor-derived molecules containing PVs divided by the informative depth –  $x/d$  – is therefore the phased-variant tumor fraction (PVAF) in a given sample. For detection of MRD in each sample, we calculated PVAF independently for doublet and triplet PVs.

#### *Monte Carlo simulation for significance of PV detection*

To assess the statistical significance of the detection of tumor-derived PVs in any sample, we implemented an empiric significance testing approach. We first define a test statistic  $f$  as follows – from a given list of  $k$  possible tumor-derived PVs observed in the best genotyping sample, we calculate the arithmetic mean of allele fractions across all  $k$  PVs (allele fraction defined as the number of read-pairs containing an individual PV ( $x_i$ ) over the number of read-pairs spanning the PV positions ( $d_i$ )):

$$f = \frac{\sum_{i=1}^k \frac{x_i}{d_i}}{k} \quad (1)$$

to assess the hypothesis that  $f$  is not significantly different from the background error-rate of similar PVs assessed from the same sample. We used a Monte Carlo approach to develop a null distribution and perform statistical testing as follows:

1. Given a set of  $k$  PVs,  $\{pv_1 \dots pv_i \dots pv_k\}$ , we generated an ‘alternate’ list of PVs,  $\{pv'_1 \dots pv'_i \dots pv'_k\}$ , such that for each alternate PV had the same type of base change and distance between SNVs as the test PV. For example, if a doublet PV, chr14:106329929 C>T and chr14:106329977 G>A, was identified in the genotyping sample, we searched for an alternate two positions at the same genomic distance (here, 48bp) with reference bases C and G, and assessed for read-pairs with the same types of base changes (i.e., C>T and G>A), using the heuristic search scheme below.
2. For each tumor  $pv_i$  in our set of  $k$ , we identified 50 such alternates. This was performed with a random search algorithm to scan the targeted space and identify alternates. To find these 50 alternates, we began by identifying a random position on the same chromosome as the test  $pv_i$  and then searched for the same types of reference bases at the same genomic distance as described above. Synteny of observed/alternate PVs was used to control for regional variation in SHM/aSHM as well as copy number variation, as potential confounders of the null distribution. Alternate positions that were identified as a germ-line SNP, defined as having AF > 5%, were excluded.
3. After identifying 50 such alternates for each  $pv_i$ , we generated 10,000 random permutations of 1 alternate for each of the  $k$  original PVs and calculated the phased-

variant fraction  $f'$  for these alternate lists in the sample of interest being evaluated for presence of MRD, as described above.

4. We then calculated an empiric P-value, defined as the fraction of times the true phased-variant fraction  $f$  is observed to be less than or equal to the alternate  $f'$  across the 10,000 random PV lists as an empirical measure of significance of MRD significance in the blood sample of interest.
5. This empiric P-value is calculated independently for both doublet and triplet PVs (i.e., two and three SNVs in phase). These P-values are then combined using Fisher's method to result in a composite signal for tumor content.

While this resulting comparison is a measure of the significance for PV detection of our tumor-reporter list compared to the empirically defined background PV error-rate *within* the sample of interest, we also evaluated its relationship to specificity of detection *across* cases and control samples, as described below.

#### *Assessment of specificity of PhasED-Seq*

To determine the specificity of disease and MRD detection through PhasED-Seq, we first identified patient-specific PVs from 107 patients with DLBCL using pretreatment tumor or plasma DNA along with paired germ-line samples. We then assessed 40 independent plasma DNA samples from healthy individuals for presence of these patient-specific PVs, using the Monte Carlo approach outlined above. We empirically determined a P-value threshold from our Monte Carlo to ensure high specificity; in the lymphoma cohort, this resulted in a P-value threshold of 0.01 yielding 97% specificity, with 98% sensitivity in pretreatment samples (**Fig Extended Data Fig 6a**).

#### *Calculation of error rates*

To assess the error profile of both isolated SNVs and PVs, we examined the non-reference base observation rate of each type of variant across all reads. For isolated SNVs, we calculated the error-rate for each possible base change  $e_{n1>n1'}$  as the fraction of on-target bases with reference allele  $n1$  that are mutated to alternate allele  $n1'$ , when considering all possible base-changes of the reference allele. Positions with a non-reference allele rate exceeding 5% were classified as probable germ-line events, and excluded from the error-rate analysis. We also

calculated a global error rate, defined as the rate of mutation from the hg19 reference allele to any alternate allele.

For phased variants, we performed a similar calculation. For the error-rate of a given type of phased variant composed of  $k$  constituent base-changes  $\{e_{n1>n1'} \dots e_{nk>nk'}\}$ , we calculated the error-rate by determining both the number of instances of the type of base change (i.e., the numerator), as well as the number of possible instances for the base change (i.e., the denominator). To calculate the numerator,  $N$ , we counted the number of occurrences of the PV of interest over all read-pairs in a given sample. For example, to calculate the error-rate of C>T and G>A phased doublets, we first counted the number of read-pairs that include *both* a reference C mutated to a T as well as a reference G mutated to an A.

To calculate the denominator,  $D$ , we also calculated the number of possible instances of this type of phased variant; this was performed first for each read-pair  $i$ , and then summed over all read pairs. A PV with  $k$  components can be summarized as having certain set of reference bases  $p_A, p_C, p_G, p_T$ , where  $p_N$  is the number of each reference base in the PV. Similarly, a given read pair contains a certain set of reference bases  $b_A, b_C, b_G, b_T$ , where  $b_N$  is the number of each reference base in the read pair. Therefore, for each read pair in a given sample, the number of possible occurrences of our PV type of interest can be calculated combinatorically as:

$$D_i = \binom{b_A}{p_A} \binom{b_C}{p_C} \binom{b_G}{p_G} \binom{b_T}{p_T} \quad (2)$$

For example, consider a read-pair with 40 reference As, 50 reference Cs, 45 reference Gs, and 35 reference Ts. The number of positions for a C>T and G>A PV is:

$$D_i = \binom{40}{0} \binom{50}{1} \binom{45}{1} \binom{35}{0} = 2250 \quad (3)$$

The aggregated denominator,  $D$ , for error rate calculation is then simply the sum of this value over all read pairs. The error rate for this type of PV is then simply  $N/D$ .

### Differences in phased variants between lymphoma subtypes

To compare the distribution of phased variants in different types of lymphomas, we identified tumor-specific PVs in 101 DLBCL, 16 PMBCL, 23 cHL, 13 FL, and 13 MCL patients via sequencing of tumor biopsy specimens and/or pre-treatment cell-free DNA and paired germ-line specimens. The clinical characteristics of these patients are detailed in **Table S4**. After identifying these tumor-specific PVs, we then assessed their distribution across the targeted sequencing panel. We began by dividing our panel into 50bp bins; for each patient, we then determined if each patient had evidence of a PV within the 50bp bin, defined as having at least one component of the PV within the bin. We furthermore determined the nearest gene to each 50bp bin, based on GENCODEv19 annotation of the reference genome.

To assess how the distribution of PVs between subtypes of lymphoma varies at the level of specific genes, we examined the distribution of PVs across the 50bp bins spanning each gene (or nearest gene). For example, consider a given gene with  $n$  such 50bp bins represented in our targeted sequencing panel. For each bin, we first determined the fraction of patients,  $f$ , in each type of lymphoma with a PV falling within the 50bp bin – i.e., we determine  $\{f_{type1,1}, \dots, f_{type1,n}\}$  and  $\{f_{type2,1}, \dots, f_{type2,n}\}$ . We then compared any two histologies for the fraction of cases harboring PVs in the set of 50bp bins assigned to each gene. The number of 50bp bins in each gene as well as summary statistics for each histology are provided in **Table S5**. These comparisons are depicted for individual representative genes in **Fig 3c**, with additional figures available for all significantly differentially effected genes available at <https://phasedseq.stanford.edu> (description of significance testing outlined below).

We statistically compared the enrichment in PVs in a specific lymphoma type or subtype vs. another by calculating the difference in the fraction of patients which contain a PV in each 50bp bin across all bins assigned to a gene (i.e., overlapping a given gene or with a given nearest gene). Specifically, for any comparison between two lymphoma types ( $type_1$  and  $type_2$ ), we first identified this set of differences in PV-rate between histologies  $\{f_{type1,1} - f_{type2,1}, \dots, f_{type1,n} - f_{type2,n}\}$ . We then compared this set of gene-specific differences in frequency of PVs between types of lymphoma against the distribution of all other 50bp bins in the sequencing panel by the Wilcoxon rank sum test. For this test, we compared the set of  $n$  50bp bins assigned to a given gene to all other 50bp bins (i.e.,  $6755 - n$ , since there are 6755 50bp bins in our sequencing panel). This P-value, along with the mean difference in fraction of patients with a PV in each bin

for each gene between histologies, is depicted as a volcano plot in **Fig 3d**. To account for the global difference in rate of PVs between different histologies, the mean difference in fraction of patients with a PV between histologies was centered on 0 by subtracting the mean difference across all genes.

### **Model to predict the probability of detection for a given set of phased variants**

To build a mathematical model to predict the probability of detection for a given sample of interest, we began with the common assumption that cfDNA detection can be considered a random process based on binomial sampling<sup>2</sup>. However, unlike SNVs occurring at large genomic distances apart from one another, detection of PVs can be highly inter-dependent, especially when PVs are degenerate (i.e., when two PVs share component SNVs) or occur in close proximity. To account for this, we only considered PVs occurring >150bp apart from each other as independent ‘tumor reporters’ (**Extended Data Fig 4f**). The number of ‘tumor reporters’ to allow for disease detection in a given sample can thus be determined as follows. The PhasED-Seq panel was broken apart into 150bp bins. Each PV in a given patient’s reporter list was then turned into a BED coordinate, consisting of the start position (defined as the left-most component SNV) and end position (defined as the right-most component SNV). For each PV, the 150bp bin from the PhasED-Seq selector panel containing the PV was determined; if a PV spanned two or more 150bp bins, it was assigned to both bins. The number of independent tumor reporters was then defined as the number of separate 150bp bins containing a tumor-specific PV.

We then developed a mathematical model comparing the expected probability of detection for a given sample at a given tumor fraction with a given number of independent tumor reporters (i.e., 150bp bins). With a given number of tumor reporters  $r$ , at a given tumor fraction  $f$ , with a given sequencing depth  $d$ , the probability of detecting 1 or more cell-free DNA molecule containing a tumor-specific PV containing can be defined as:

$$Pr(detection) = 1 - Pr(nondetection) \quad (4)$$

$$= 1 - \binom{d * r}{0} f^0 (1 - f)^{d * r} \quad (5)$$

based on simple binomial sampling. However, as we trained our ctDNA detection method to have a 5% false positive rate, we add this false positive rate term to our model as well:

$$Pr(detection) = 1 - Pr(nondetection) + 0.05 * Pr(nondetection) \quad (6)$$

$$Pr(detection) = 1 - 0.95 * Pr(nondetection) \quad (7)$$

$$= 1 - 0.95 * \binom{d * r}{0} f^0 (1 - f)^{d * r} \quad (8)$$

**Extended Data Fig 5h** shows the results of this model for a range of tumor reporters  $r$  from 3 to 67 at depth  $d$  of 5000. The confidence envelope on this plot shows solutions for a range of depth  $d$  from 4000 to 6000.

To empirically validate this model assessing the probability of disease detection, we utilized samples from our limiting dilution series. In this dilution series, 3 patient cfDNA samples, each containing patient-specific PVs, were spiked into healthy control cfDNA. For each list of patient specific PVs, we performed 25 random subsamplings of the 150bp bins containing patient-specific PVs to generate reporter lists containing variable numbers of tumor-specific reporters. We selected a maximum bin number of 67 to allow sampling from all 3 patient-specific PV lists, followed by scaling down the number of bins by 2x or 3x per step. This resulted in reporter lists containing patient-specific PVs from 3, 6, 17, 34, or 67 independent 150bp bins. We then assessed for disease detection using each of these patient-specific PV lists of increasing size in each of our 'wet' limiting dilution samples from 1:1,000 to 1:1,000,000 (**Extended Data Fig 5i**, closed circles). We furthermore created *in silico* mixtures using sequencing reads from our limiting dilution samples with varying expected tumor-content, and again assessed for the probability of disease detection using patient-specific subsampled PV reporter lists of varying lengths (open circles). For this experiment, we down-sampled both the 'wet' and '*in-silico*' dilution bam files to achieve a depth of ~4000-6000x to correspond with our modeled depth. The final mean and standard deviation of depth across all down-sampled bam files was 4214x  $\pm$  789. The probability of detection was summarized across all tests at a given expected tumor fraction, for a given patient-specific PV list. For each given dilution, we considered multiple independently sampled sets of reads to allow superior estimation of the true probability of

detection. Specifically, we considered the following number of replicates at each dilution indicated:

| Dilution        | Replicates | Number of Tests<br>(Replicates * 25) | Wet or <i>In silico</i> |
|-----------------|------------|--------------------------------------|-------------------------|
| 1 : 1,000       | 1          | 25                                   | Wet                     |
| 5 : 10,000      | 3          | 75                                   | <i>In silico</i>        |
| 3.5 : 10,000    | 3          | 75                                   | <i>In silico</i>        |
| 2 : 10,000      | 3          | 75                                   | <i>In silico</i>        |
| 1 : 10,000      | 3          | 75                                   | Wet                     |
| 5 : 100,000     | 3          | 75                                   | <i>In silico</i>        |
| 3.5 : 100,000   | 3          | 75                                   | <i>In silico</i>        |
| 2 : 100,000     | 3          | 75                                   | <i>In silico</i>        |
| 1 : 100,000     | 3          | 75                                   | Wet                     |
| 5 : 1,000,000   | 8          | 200                                  | <i>In silico</i>        |
| 3.5 : 1,000,000 | 8          | 200                                  | <i>In silico</i>        |
| 2 : 1,000,000   | 8          | 200                                  | Wet                     |
| 1 : 1,000,000   | 8          | 200                                  | Wet                     |

The total number of tests, for each patient-specific PV list, is therefore the number of randomly subsampled PV lists (e.g., 25) times the number of independently downsampled bam files; this number is provided in the table above. In **Extended Data Fig 5i**, the points and error-bars represent the mean, minimum, and maximum across all three patients. The concordance between the predicted probability of disease detection from our theoretical mathematical model and our wet and *in silico* samples validating this model, is shown in **Extended Data Fig 5j**.

### Model to assess theoretical sensitivity of ctDNA by tracking PVs vs other variants

To assess the utility of tracking phased variants for ctDNA in diverse cancers, we assess the theoretical sensitivity that would be afforded by developing a personalized sequencing panel for each patient to track a specific set of variants. To do this, we need to know the expected size of a personalized sequencing panel, which depends on the number of variants targeted, as well as the expected depth after UMID-mediated deduplication for a given number of sequencing reads.

We utilized the data from a recently published optimized hybrid capture workflow<sup>3</sup> to infer the expected deduplicated or duplex depth from a given amount of sequencing, for a given panel size, from a given amount of starting cfDNA. Specifically, the above optimized workflow describes a given amount of depth (UMID deduped or duplex) per ng of DNA input. We fit an exponential-asymptotic model to this data, of the form:

$$depth = N * \left(1 - e^{-k \frac{r}{1,000,000}}\right) \quad (9)$$

Where  $N$  is the carrying capacity (i.e., the maximum depth that can be recovered) in depth/ng DNA input, and  $r$  is the number of reads given to a sample per ng DNA input per bp of selector. The result of this model-fitting is shown in **Extended Data Fig 10a**, and resulted in the following coefficients:

|                                      | <b>N</b><br>(maximum depth; x / ng DNA<br>input) | <b>k</b><br>(exponential<br>coefficient) |
|--------------------------------------|--------------------------------------------------|------------------------------------------|
| <b>Total De-duplicated<br/>Depth</b> | 212.7                                            | 2.7e5                                    |
| <b>Duplex Depth</b>                  | 102.8                                            | 4.7e4                                    |

Using this model, we then assessed the theoretical sensitivity of possible personalized cfDNA assays considering capturing: 1) phased variants, 2) SNVs utilizing duplex sequencing, or 3) structural variants. In each case from PCAWG, we considered a hypothetical panel covering the maximum number of the variant type of interest (up to a maximum of 10,000), requiring a hybrid capture tile of 120bp (e.g., a panel tracking 100 variants would be 12,000bp, etc). We then assessed the theoretical number of recovered evaluable fragments from this panel using 64ng of DNA input and 20,000,000 sequencing reads, as inferred by the above model. For each case from the PCAWG set, we then assessed which methodology would result in the greatest number of recovered cfDNA fragments evaluable for tumor content<sup>1</sup>. The inferred detection-limit was then considered to be the lowest allele fraction predicted to be detectable with 95% probability based on binomial sampling. The inferred detection-limit for personalized PhasED-

Seq is shown in **Fig 7a**, while the comparison to duplex sequencing or tracking structural variants is shown in **Extended Data Fig S10b-c**.

## References

- 1 Wan, J. C. M. *et al.* ctDNA monitoring using patient-specific sequencing and integration of variant reads. *Science translational medicine* **12**, doi:10.1126/scitranslmed.aaz8084 (2020).
- 2 Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine* **20**, 548-554, doi:10.1038/nm.3519 (2014).
- 3 Chabon, J. J. *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245-251, doi:10.1038/s41586-020-2140-0 (2020).

**Supplementary Table 9**

Clinical characteristics of patient samples used for extension of PhasED-Seq to solid tumors. The for ctDNA detection by both CAPP-Seq and PhasED-Seq.

| <b>SubjectID</b> | <b>Histology</b>        | <b>Clinical Stage</b> | <b>Treatment</b>                      |
|------------------|-------------------------|-----------------------|---------------------------------------|
| LUP502           | Lung Adenocarcinoma     | IB                    | NA                                    |
| LUP503           | Lung Adenocarcinoma     | IA                    | NA                                    |
| LUP649           | Lung Adenocarcinoma     | IB                    | NA                                    |
| LUP814           | Lung Adenocarcinoma     | IIIB                  | Combined Chemoradition, immunotherapy |
| LUP831           | Lung Adenocarcinoma     | IIIB                  | Combined Chemoradition, immunotherapy |
| BRCA001          | denocarcinoma (Triple N | IIB                   | Neoadjuvant chemotherapy, lumpectomy  |

number of plasma samples is also shown, as well as the characteristics

| Number of Samples Pre-Diagnosis | Number of Samples At Diagnosis |
|---------------------------------|--------------------------------|
| 0                               | 1                              |
| 0                               | 1                              |
| 0                               | 1                              |
| 0                               | 1                              |
| 0                               | 1                              |
| 5                               | 1                              |

| Number of Samples After Diagnosis | Number CAPP-Seq Variants |
|-----------------------------------|--------------------------|
| 0                                 | 103                      |
| 0                                 | 120                      |
| 0                                 | 45                       |
| 4                                 | 9                        |
| 5                                 | 3                        |
| 4                                 | 28                       |

| <b>CAPP-Seq Method</b> | <b>WGS Depth, Tumor</b> | <b>WGS Depth, Normal</b> |
|------------------------|-------------------------|--------------------------|
| Personalized           | 44                      | 73                       |
| Personalized           | 44                      | 61                       |
| Personalized           | 21                      | 35                       |
| Population-based       | 37                      | 38                       |
| Population-based       | 26                      | 36                       |
| Personalized           | 24                      | 28                       |

**Number of Validated PVs**

116

223

120

622

82

14