

Error-corrected flow-based sequencing at whole-genome scale and its application to circulating cell-free DNA profiling

Received: 21 November 2022

Accepted: 4 March 2025

Published online: 11 April 2025

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Differentiating sequencing errors from true variants is a central genomics challenge, calling for error suppression strategies that balance costs and sensitivity. For example, circulating cell-free DNA (ccfDNA) sequencing for cancer monitoring is limited by sparsity of circulating tumor DNA, abundance of genomic material in samples and preanalytical error rates. Whole-genome sequencing (WGS) can overcome the low abundance of ccfDNA by integrating signals across the mutation landscape, but higher costs limit its wide adoption. Here, we applied deep (~120×) lower-cost WGS (Ultima Genomics) for tumor-informed circulating tumor DNA detection within the part-per-million range. We further leveraged lower-cost sequencing by developing duplex error-corrected WGS of ccfDNA, achieving 7.7×10^{-7} error rates, allowing us to assess disease burden in individuals with melanoma and urothelial cancer without matched tumor sequencing. This error-corrected WGS approach will have broad applicability across genomics, allowing for accurate calling of low-abundance variants at efficient cost and enabling deeper mapping of somatic mosaicism as an emerging central aspect of aging and disease.

Distinguishing sequencing errors from true variants continues to challenge the genomics field, with tradeoffs between cost and error suppression strategies. These challenges are magnified when identifying ultrarare variants. For example, profiling circulating cell-free DNA (ccfDNA) is a promising clinical tool for noninvasive cancer detection^{1–8}. Sequencing of somatic variants in circulating tumor DNA (ctDNA)^{9–12} can aid in detection of low-burden disease, such as cancer screening, minimal residual disease (MRD)^{11,13–16} and relapse monitoring^{17,18}. However, when disease burden is low, robust detection requires methods with exquisite sensitivity to detect ctDNA signal over the background rate of sequencing or library preparation errors. This technical challenge is typically overcome by increasing sequencing depth at select genomic locations, accompanied by approaches that decrease sequencing error rate, including unique molecular identifier (UMI) error suppression techniques^{10,19} or duplex sequencing^{12,20–22}, which enable increased accuracy in differentiating true somatic variants from sequencing artifacts to optimize detection of low-burden disease.

Prevailing methods of ctDNA detection use targeted sequencing, which increases the number of genomes sequenced at a targeted location. However, high-throughput targeted sequencing rapidly exhausts available genomes (1,000–10,000 genome equivalents (GEs) per ml of plasma²³), setting a ceiling on ctDNA detection, where further increases in sequencing depth at targeted sites afford no advantage after the limited number of GEs has already been sequenced. Alternatively, whole-genome sequencing (WGS) approaches exploit breadth of coverage to supplant depth, eliminating the reliance on detecting few sites to increase ctDNA characterization in low tumor fraction settings. For example, our recent methods MRDetect¹⁴ and MRD-EDGE²⁴ use matched primary tumor mutational profiles to inform genome-wide tumor single-nucleotide variant (SNV) detection in ccfDNA, such that the available number of GEs is no longer the limiting factor for ctDNA detection.

The detection challenges presented by sparsity call for broad, accurate and deep ccfDNA sequencing. Thus, whole-genome, low-error,

✉ e-mail: alexandre.cheng@etsmtl.ca; dlandau@nygenome.org

high-coverage methods are necessary for robust ctDNA analysis. However, the costs associated with these approaches are often prohibitive. Although genome sequencing costs have rapidly dropped, more recently this decrease has stagnated²⁵, rendering sequencing cost a substantial barrier for implementing high-depth WGS for liquid biopsies, where shallow WGS is insufficient for ctDNA detection when tumor fractions are low (for example, $\sim 10^{-5}$).

Recently, low-cost mostly natural sequencing by synthesis^{26,27} has been developed by Ultima Genomics, where the flow-based platform produces single-end reads at ~ 10 billion reads per run for \$1 per gigabase, lowering costs compared to current platforms. To date, mostly natural sequencing by synthesis/Ultima sequencing has not been harnessed for clinical ctDNA sequencing in ccfDNA samples, and error rate profiles have not been fully characterized or rigorously compared to competing technologies. For potential application to clinical ctDNA monitoring, accurate error rate estimates are critical due to the required high sensitivity of ctDNA detection²⁸.

To investigate deep WGS for ctDNA detection, we used the Ultima platform to sequence ccfDNA from plasma samples from healthy individuals, individuals with cancer and patient-derived xenograft (PDX) mouse models. We show that deep plasma WGS ($\sim 120\times$) allows tumor-informed ctDNA detection within the part-per-million range. We further leveraged the cost-effective and high-throughput nature of mostly natural sequencing by synthesis to develop high-coverage WGS duplex error-corrected libraries of ccfDNA, achieving error rates as low as 7.7×10^{-8} . This allowed us to combine the advantages of genome-wide mutational integration on the one hand and molecular error correction on the other to assess disease burden in individuals with melanoma and urothelial cancer without relying on matched tumor sequencing. Together, our results demonstrate the feasibility and utility of deep WGS for ctDNA detection and duplex sequencing at the whole-genome scale.

Results

Flow-based sequencing enables highly accurate SNV discovery

Different sequencing methods have advantages and drawbacks. In flow-based Ultima sequencing, sequencing signal intensity translates to the number of consecutive bases of a given nucleotide, increasing susceptibility to homopolymer size estimation errors. However, as each sequencing cycle encompasses a single base, flow-based sequencing systems could be particularly robust to SNV errors.

To investigate error rates of the flow-based Ultima Genomics platform, we ligated molecular barcodes to the plasma of mouse PDX samples ($n = 4$; $n = 1$ lung cancer; $n = 3$ diffuse large B cell lymphoma human-mapped fractions of 0.4, 40, 73 and 96%). We computationally tracked PCR duplicates in the three PDX samples with high human-mapped read fractions to identify sequencing artifacts and compare error profiles in matched Ultima–Illumina datasets. First, we analyzed homopolymer length discordance rate between PCR duplicates and between a read and the reference genome. We randomly sampled up to 33 million unique ccfDNA molecules (1.5 million molecules for each autosome) containing at least two PCR duplicates from the $n = 3$ PDX datasets (human-mapped reads only). Each read was aligned to its duplicate and to the reference genome, and matching homopolymers were compared by their sequenced size. In Ultima datasets, we observed a strong concordance between PCR duplicates (99.34%; Q score = 21.8) and between reads to the reference genome (99.58%; Q score = 23.8) for homopolymers of size 1 to 12 (Extended Data Fig. 1a,b). We also observed artifactual homopolymer sizes near lengths of 12, as Ultima sequencing reports a maximum homopolymer size of 12 bases (ref. 26). However, homopolymer size estimations were more accurate in Illumina datasets (99.99% and Q score = 39.9 between PCR duplicates; 99.98% and Q score = 36.3 between a read and the reference genome; Extended Data Fig. 1c,d). We also found that accuracy decreased as a function of homopolymer length in both

technologies (Spearman's $\rho = -0.925$, $P = 6.5 \times 10^{-16}$ (Ultima) and Spearman's $\rho = -0.996$, $P < 2.2 \times 10^{-16}$ (Illumina); Extended Data Fig. 1e,f). For example, the homopolymer size estimation error rate was 1.57×10^{-3} , 4.46×10^{-3} and 8.51×10^{-3} for homopolymers of sizes 1, 2 and 3, respectively, in Ultima datasets and 1.02×10^{-4} , 1.66×10^{-4} and 2.63×10^{-4} in Illumina datasets. As expected, concordance improved in Ultima datasets after PCR duplicate consensus calling (error rates of 4.08×10^{-4} , 5.24×10^{-4} and 8.06×10^{-4} for homopolymers of sizes 1, 2 and 3, respectively; family size = 2; Extended Data Fig. 1e). For single-read accuracy, we observed that Ultima homopolymer accuracy falls below 99% at homopolymer lengths of 4 and greater (3.17% of the human genome; Extended Data Fig. 1e,g), whereas matched Illumina datasets fall below 99% at homopolymer lengths of 8 and greater (0.09% of the genome; Extended Data Fig. 1f,g).

To further compare performance in terms of SNV errors, we first identified putative sequencing errors, defined as mismatched PCR duplicates, where only one of the reads contains a mismatch with the reference genome. Overall, single-nucleotide differences between PCR duplicates occurred at a rate of $1.58 \times 10^{-4} \pm 2.65 \times 10^{-4}$ in Ultima datasets and $9.85 \times 10^{-4} \pm 10.80 \times 10^{-4}$ in Illumina datasets (Extended Data Fig. 2a). Interestingly, we found wide variation in sequencing error rates that depended on the trinucleotide context of the variant. In Ultima datasets, errors were most likely to occur in specific motifs: trinucleotides containing a 2-mer (for example, C[C>A]T) or where a 2-mer would form following mutation (for example, G[A>G]T), likely reflecting homopolymer size estimation errors that manifest as SNV errors. Conversely, we found that certain trinucleotide contexts were robust to sequencing error, namely when a reported trinucleotide mutation would cause a shift in the number of sequencing cycles compared to the reference (termed cycle shift motifs; Extended Data Fig. 2b). For example, assuming a sequencing cycle of T>G>C>A, T[G>A]C would be considered a cycle shift, whereas T[G>C]C would not. In Ultima datasets, cycle shift motifs had significantly lower error rates than noncycle shift motifs (mean of 5.23×10^{-5} versus 2.32×10^{-4} in cycle shifts and noncycle shifts, respectively; $P < 2.2 \times 10^{-16}$; Extended Data Fig. 2c). Cycle shift motifs with the highest error rates were exclusively in C>T mutations, the most common somatic mutation. These may present rare cases where the PCR duplicate containing the mismatch is correct and the PCR duplicate that matches the reference contains the error. As expected, there were no significant differences in error rates between cycle shift and noncycle shift motifs in Illumina datasets (mean error rate of 9.3×10^{-4} versus 1.0×10^{-3} in cycle shifts and noncycle shifts, respectively, $P = 0.16$; Extended Data Fig. 2c).

Plasma WGS can detect low tumor burden

Having benchmarked Ultima versus Illumina error rates, we sought to test Ultima sequencing for ctDNA detection in clinical samples. Lower limits of ctDNA mutation detection by plasma WGS are dictated by tumor mutational burden, depth of sequencing and error rates from library preparation and sequencing. To explore these dependencies, we modeled variable tumor fractions, depths of coverage and error rates for a cancer with 10,000 SNVs (~ 3.7 mutations per megabase; Methods). Tumors with $>10,000$ SNVs are seen across $\sim 30\%$ of cancers in the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium dataset²⁹ and are enriched in lung (85%), skin (79%), bladder (83%) and other cancers. This analysis suggests that detection of tumor fractions below 10^{-5} requires a sequencing depth of $\sim 100\times$ with error rates below 10^{-4} (Fig. 1a). In these high-mutational-burden cancers, WGS can provide more opportunities for ctDNA detection than targeted approaches by sequencing a greater number of unique ccfDNA molecules. However, detecting low-mutational-burden tumors or specific driver variants is better served with targeted approaches (Extended Data Fig. 3).

Given the need for deeper plasma WGS, sequencing costs impede broad application. We therefore hypothesized that lower-cost Ultima sequencing can help overcome this barrier, provided that sequencing

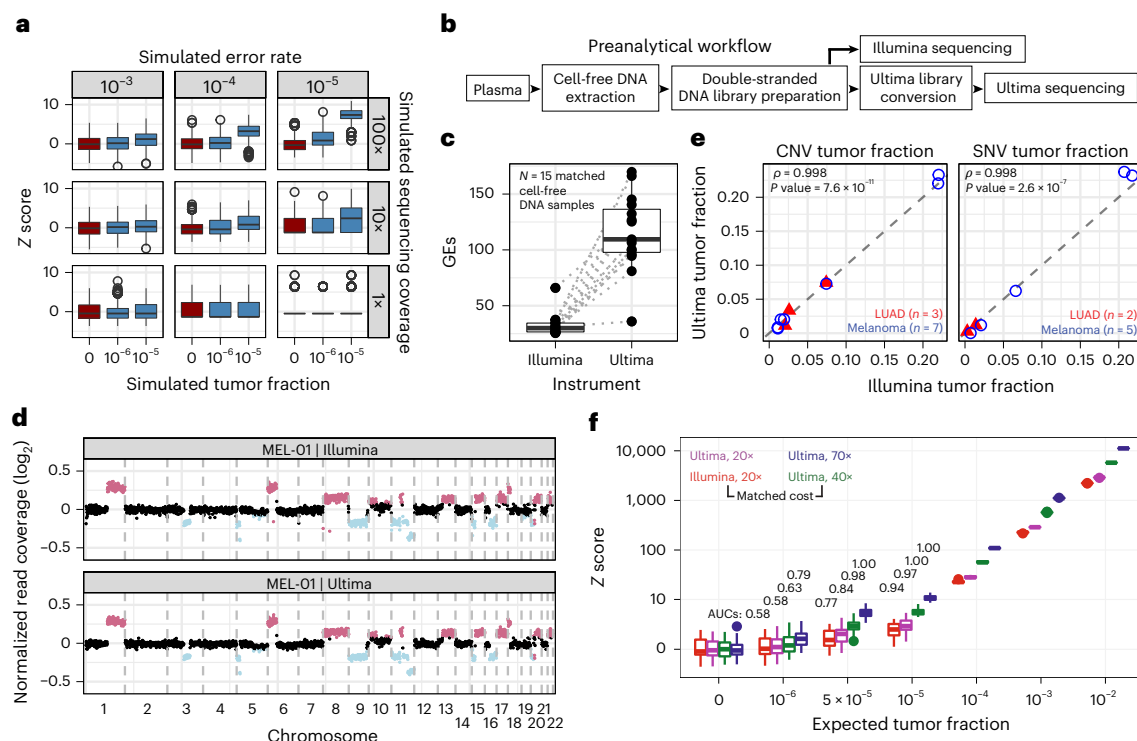


Fig. 1 | Ultralow ctDNA detection requires deep sequencing coverage and low error rates. **a**, Simulation of ctDNA detection given different error rates (columns), whole-genome coverages (rows) and tumor fractions (x axis); $n = 1,000$ replicates per set of conditions. **b**, Cell-free DNA library preparation preanalytical workflow. **c**, Sequencing depth of matched Illumina and Ultima datasets. **d**, Normalized read coverage for Illumina- (top) and Ultima-sequenced (bottom) matched cell-free DNA samples. **e**, Left, copy-number-based tumor fraction estimation measured with Illumina or Ultima sequencing in matched samples using ichorCNA. Matched cancer-free samples were used to create a panel of normal samples before tumor fraction estimation. Right, SNV-based tumor fraction estimation measured with Illumina or Ultima sequencing. Somatic SNVs were identified through matched tumor–normal sequencing. Two samples without tumor sequencing and with low ctDNA fraction ($<5\%$ measured through CNV analysis) were omitted from this analysis. Spearman's ρ coefficient

and corresponding two-sided P values were calculated using the stats package (v.3.6) function in R (v.3.6); LUAD, lung adenocarcinoma. **f**, ctDNA cost and coverage analysis between Illumina and Ultima sequencing in a matched sample. Area under the curve (AUC) values are measured by calculating the area under a receiver operating characteristic curve comparing a given group (for example, Illumina $20\times$ at 10^{-6} expected tumor fraction) to its platform and coverage-matched healthy control (for example, Illumina $20\times$, expected tumor fraction of 0); $n = 20$ replicates per set of conditions. All AUC values at expected tumor fractions of 10^{-4} and greater were 1.00. Z scores of a given sample are calculated against their coverage and platform-matched healthy control (expected tumor fraction of 0). For all box plots, the bottom and top ends of the boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the interquartile range (IQR).

error profiles are sufficiently low. As such, we performed Ultima WGS on 15 ccfDNA libraries ($n = 10$ samples from individuals with cancer ($n = 7$ stage IV melanoma; $n = 3$ III–IV lung adenocarcinoma); $n = 5$ control samples; sequencing depth $115 \times \pm 34 \times$ (including duplicates; mean \pm s.d.)) with matching Illumina sequencing of the same libraries (including duplicates; $33 \times \pm 10 \times$; Fig. 1b,c and Supplementary Tables 1–3). We first measured ctDNA burden by estimating the relative abundance of large-copy-number alterations using ichorCNA³⁰ (Methods). Tumor fractions varied ($<3\%$, $n = 3$; $3–10\%$, $n = 5$; $>10\%$, $n = 2$), and measurements were strongly correlated between matched Illumina and Ultima datasets (Spearman's $\rho = 0.998$, $P = 7.6 \times 10^{-11}$; Fig. 1d,e). Next, we examined estimated ctDNA burden using a tumor-informed SNV approach. We performed WGS of tumor-derived DNA (using standard mutation calling on plasma DNA if tumor DNA was unavailable and ctDNA burden was $>5\%$ by ichorCNA³⁰; Methods and Supplementary Table 4) and matched normal DNA from peripheral blood mononuclear cells to identify tumor-specific mutations. To remove sequencing errors, we developed a quality-filtering pipeline informed by Ultima-specific feature cutoffs and blacklisted regions (Supplementary Fig. 1, Methods and Supplementary Note). We mined the denoised cell-free DNA reads for somatic variants to estimate ctDNA fractions (Methods), finding strong agreement between Illumina and Ultima sequencing sets (Spearman's $\rho = 0.998$, $P = 2.6 \times 10^{-7}$; Fig. 1e),

with comparable tumor-specific error rates (Supplementary Fig. 2 and Supplementary Note) and fragment lengths (Supplementary Fig. 3 and Supplementary Note). Together, these results support the utility of Ultima sequencing for detection of low-burden ctDNA.

To further benchmark the two sequencing technologies, we performed *in silico* mixing studies¹⁴ by computationally mixing sequencing reads from a sample with detectable tumor burden with sequencing reads from a healthy control sample at known ratios. This method is comparable to molecular mixing studies²⁴ where a known amount of ccfDNA from a sample with detectable tumor burden is spiked into healthy ccfDNA at known concentrations. We generated admixtures from sample MEL-05 (stage IV melanoma, tumor fraction = 7.3%) and CTRL-05 (no known cancer) at different ratios to create admixtures of tumor fractions ranging from 10^{-6} to 10^{-2} ($n = 50$ technical replicates per admixture) at $70\times$, $40\times$ and $20\times$ sequencing depth for Ultima-sequenced datasets and $20\times$ for Illumina. Performance was evaluated at different simulated tumor fractions using a receiver operating curve analysis.

We determined analytical limits of detection of a plasma sample with matched Illumina and Ultima sequencing in two different contexts: (1) at matched coverage ($20\times$ each), which shows comparable performance between the two methods (Fig. 1f and Extended Data Fig. 4), and (2) at matched cost ($20\times$ Illumina versus $40\times$ Ultima,

assuming \$1 per gigabase (Ultima) and \$2 per gigabase (Illumina)). As expected, given the similar error rates of the two sequencers without denoising, and slight improvements in Ultima datasets after denoising (Supplementary Fig. 2), the deeper Ultima datasets (40×) demonstrate better limits of detection ($AUC = 0.98$ at tumor fractions of 5×10^{-5}) than price-matched Illumina datasets (20×; $AUC = 0.77$ at tumor fractions of 5×10^{-5}). Finally, we were able to robustly detect ctDNA in the parts-per-million range in deeply sequenced Ultima datasets ($AUC = 0.79$; 70× coverage), indicating that our deep-sequencing framework for ctDNA detection is sensitive enough at low tumor fractions for use in challenging clinical applications such as MRD monitoring.

Ultralow error duplex WGS of ccfDNA

Advances in molecular error correction have radically enhanced deep targeted sequencing approaches, for example using UMIs that are incorporated during library preparation for sequencing error correction^{10,31}. Although strand-agnostic UMIs can correct sequencing and PCR errors, UMIs that link forward and reverse DNA strands (duplex sequencing) can correct errors that arise on only one strand (such as G>T transversions from oxidative DNA damage³²) during library preparation²². At the whole-genome scale, duplex sequencing has been cost prohibitive due to the need for a high rate of duplicate reads. Nonetheless, studies applying duplex sequencing at the genome scale have shown promise for genome-wide rare variant identification^{20,21}.

We reasoned that the lower sequencing cost afforded by mostly natural sequencing by synthesis could open the way for affordable genome-scale duplex sequencing in clinical settings, and decreasing sequencing and library preparation errors could enable tumor-agnostic (de novo) ctDNA detection, where matched tumor tissue, often unavailable for patients^{33,34}, cannot be used to reduce background noise. For this important clinical context, we developed duplex WGS for single-end Ultima reads. Here, we created duplex libraries by replacing standard sequencing adapters with adapters containing three random nucleotides, thus creating a 6-bp duplex UMI (using the random nucleotides from both ends of the DNA). Libraries were then sequenced using the Ultima sequencer. Although duplex sequencing was developed for paired-end sequencing technologies, we recovered the ends of most ccfDNA molecules (80% of all sequenced molecules) due to their modal length of ~170 bp (Supplementary Fig. 3) to create 6-bp UMIs. Next, we developed a decision tree classifier that determines the duplex variant based on the variant pileups of individual sequencing reads contributing to the duplex. This allowed us to use sequencing reads of the same duplex family that might differ in size due to homopolymer size estimation differences, which would be discarded by current duplex collapsing tools and maximize duplex yields (Methods, Supplementary Note and Supplementary Fig. 4), additionally lowering error rates through read end trimming (Methods, Supplementary Note and Supplementary Fig. 5).

To test the accuracy of duplex error correction, we used the ccfDNA duplex libraries from PDX plasma. Tumor fractions (the fraction of reads uniquely mapping to the human genome) were 0.4, 40, 73 and 96% (Supplementary Table 2). For each sample, we denoised sequencing reads in three different ways: (1) UMI agnostic, where PCR duplicates (identified by their mapping positions) are removed from analysis and reads are denoised based on their sequencing and mapping qualities (Methods), (2) using UMIs to identify PCR duplicates for single-strand error correction (reads used to create a consensus are profiled base by base and a consensus base pair is determined by computing the likelihood of that base being an A, T, C or G, using the sequenced nucleotides and their qualities as priors) and (3) using PCR duplicates and forward and reverse strands of a same double-stranded DNA template for error correction (duplex error correction). Similar to recent studies describing novel error correction methods^{35,36}, we defined a residual SNV rate as the number of base pairs in denoised

reads that were discordant with the reference genome and occurred once divided by the total number of interrogated (that is, denoised) bases in a sample (Methods). Here, we limited our analysis to normal ccfDNA (reads mapping to the mouse genome) in the three PDX samples with appreciable mouse-derived ccfDNA. Overall, we obtained residual SNV rates of $3.8 \times 10^{-4} \pm 3.2 \times 10^{-5}$, $5.8 \times 10^{-5} \pm 2.5 \times 10^{-5}$ and $7.7 \times 10^{-7} \pm 5.8 \times 10^{-7}$ for UMI-agnostic, single-strand-corrected and duplex-corrected reads, respectively (Fig. 2a and Supplementary Table 5), achieving a three-orders-of-magnitude reduction in error rate with duplex sequencing. When compared to uncorrected reads (without any quality filtering), we observed a ~3,300× improvement in error rates using duplex WGS ($2.5 \times 10^{-3} \pm 1.6 \times 10^{-3}$ error rates in uncorrected reads without any denoising), consistent with previous reports using whole-genome duplex sequencing (2×10^{-7} in Abascal et al.²¹; 4.5×10^{-7} in Bae et al.³⁶), suggesting that ccfDNA errors are driven more by library preparation and DNA degradation than by sequencing errors.

To further analyze duplex error correction for variant calling in ccfDNA, we assessed the variant allele frequency (VAF) distribution in noncancer control ccfDNA (in humans). Although UMI-agnostic WGS showed only $78.8\% \pm 2.0\%$ of base substitutions identified at the expected VAF range for germline events (Methods and Extended Data Fig. 5), this increased to $97.6\% \pm 0.5\%$ after duplex error correction. In high-burden metastatic melanoma ccfDNA samples, duplex error correction and removal of germline mutations allowed for the detection of somatic mutation VAF modes (Supplementary Table 6), consistent with tumor fraction estimates using ichorCNA (Fig. 2b,c), rendering this approach feasible for ctDNA mutation detection in clinical samples without a matched tumor (nontumor-informed).

Nontumor-informed monitoring of low-burden melanoma

Current methods for targeted detection of ctDNA include de novo detection of somatic mutations (nontumor-informed) from ‘off-the-shelf’ sequencing panels that target recurrent mutations of a given cancer type or tumor-informed deep targeted sequencing, where a tumor is sequenced a priori, and a personalized panel is designed to target cancer-derived mutations. However, these methods have inherent limitations. For example, panel sensitivity is limited by the scarcity of ccfDNA in plasma (only up to 1,000–10,000 GE per ml of plasma²³), incomplete driver detection (~49% driver mutation detected in individuals with stage IV non-small cell lung cancer³⁷) and the inability to distinguish cancer-derived mutations from those arising from alternative biological processes, such as clonal hematopoiesis^{38,39}. Likewise, tumor-informed targeted approaches are limited by ccfDNA scarcity, with intratumoral heterogeneity and primary tumor sequencing failure further restricting the targeted approach by impacting the accuracy of the mutation landscape used to build the panels^{33,34,40}. Furthermore, the reliance of tumor-informed methods on only mutations from primary tumors results in a lack of sensitivity for detecting phylogenetically distant metastasis. For phylogenetically distant metastasis, tumor-informed approaches will offer reduced benefit, whereas a tumor-free approach considers any cancer mutation and thus may overcome this challenge.

WGS-based nontumor-informed mutation detection does not sensitively detect cancer hotspot mutations due to relatively lower coverage per genomic position. Thus, SNV-based WGS methods of ctDNA detection must rely on genome-wide mutational integration, where identifying multiple mutations improves detection power^{14,24}. We hypothesized that signatures of somatic mutation accumulation^{41–43} could provide a framework for genome-wide mutational integration for nontumor-informed WGS ctDNA detection.

Specifically, we reasoned that genome-wide mutations could be integrated and summarized as a weighted sum of single-base substitution (SBS) reference mutational signatures originating from (i) the tumor in individuals with cancer or (ii) clonal hematopoiesis⁴². We explored the trinucleotide contexts of ccfDNA variants through

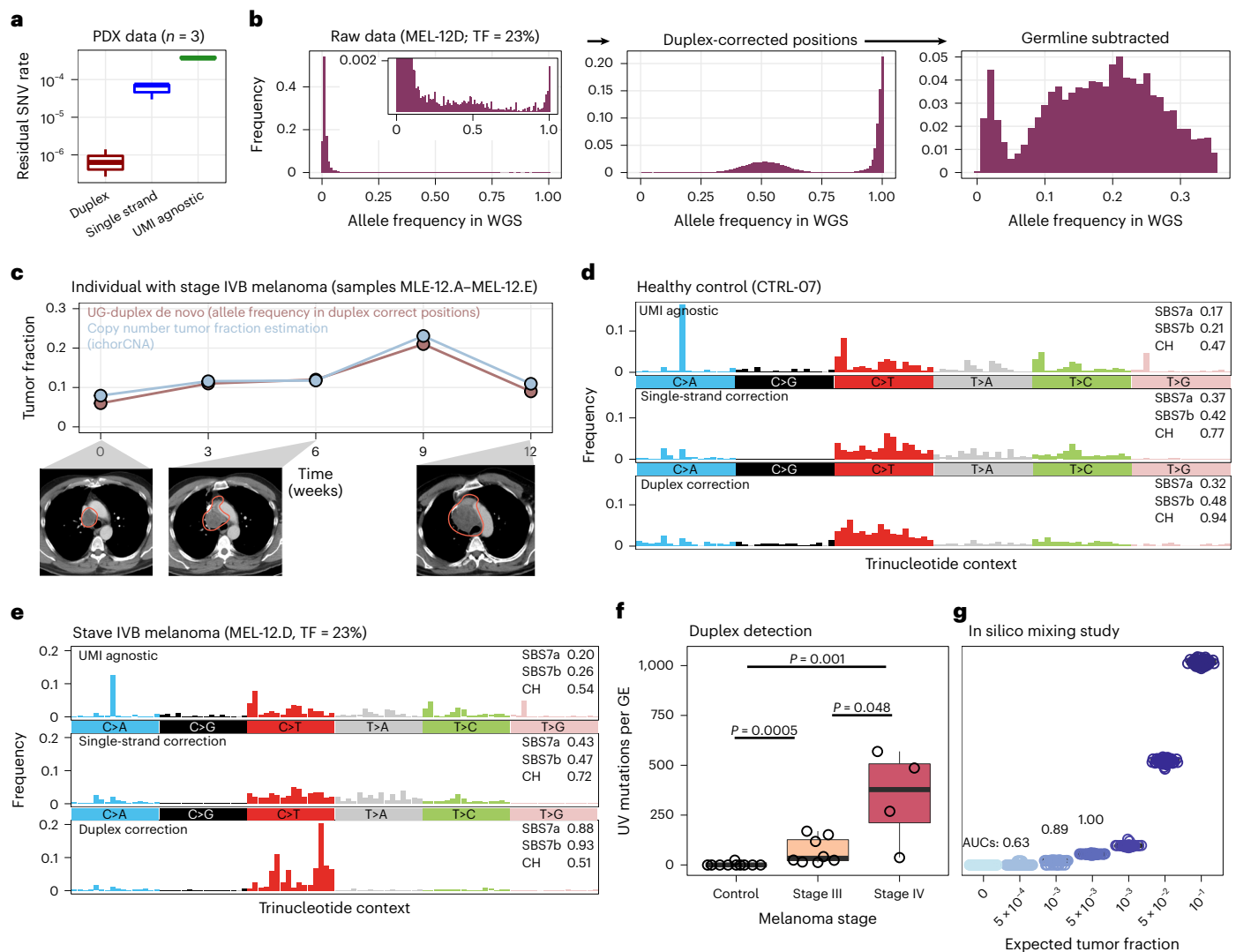


Fig. 2 | Duplex correction allows ctDNA identification without tumor sequencing. **a**, Error rates in WGS on mouse PDX samples ($n = 3$). **b**, VAF, calculated using unfiltered sequencing reads, in positions where a variant was found using uncorrected reads (left; inset highlights higher allele frequencies by enforcing a y axis cutoff of 0.002) and in duplex-corrected reads (middle). Removing germline reads reveals somatic mutations with a modal allele frequency of 0.21 (right); TF, tumor fraction. **c**, Comparison between the modal allele frequency of an individual with progressive disease (samples MEL-12.A–MEL-12.E) in duplex-corrected positions (allele frequencies between 5% and 30% only) and copy-number-based tumor fraction estimations. UG, Ultima Genomics. **d,e**, Trinucleotide frequencies of a healthy plasma sample (CTRL-07; **d**) and a stage IVB cancer plasma sample (MEL-12.D; tumor fraction of 23%; **e**) in UMI-agnostic corrected WGS (top row), single-stranded correction (middle row) and duplex correction (bottom row). Cosine similarity with SBS7a/SBS7b (UV damage; Cosmic v3.3) and clonal hematopoiesis⁴² (CH) is compared across

conditions. **f**, Plasma mutational scores due to UV damage in individuals with melanoma (stage IV ($n = 4$) and stage III ($n = 8$)) and healthy individuals ($n = 10$) at baseline (before treatment or surgery). Plasma signatures were fit to a custom reference of signatures comprising SBS7a, SBS7a (Cosmic v3.3) and clonal hematopoiesis⁴². **g**, In silico mixing study of metastatic melanoma sample MEL-12.B with control sample CTRL-06 (40 replicates per tumor fraction, 6.5× coverage per replicate). Tumor scores were estimated by fitting the sample's trinucleotide frequencies to those of signatures SBS7a, SBS7b (Cosmic v3.3) and clonal hematopoiesis⁴². AUC values were measured by comparing replicates of a given tumor fraction to tumor fraction = 0 replicates. For box plots in **a** and **f**, the bottom and top ends of the boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the IQR. P values were calculated using a two-sided Wilcoxon test. Scans from **c** are adapted from Widman et al.²⁴.

mutation signature analysis at denoising levels (UMI agnostic, UMI single stranded and duplex) to investigate the potential mutation etiologies. Cosine similarities (measures similarities between two mutational signatures^{21,43}) between the UV-associated SBS7a and SBS7b signature (COSMIC⁴⁴ v3.3) and high-burden samples (MEL-12A–E, stage IVB melanoma) were highest after duplex correction (across samples mean cosine similarities to SBS7a of 0.21 ± 0.011 (range 0.19–0.22), 0.35 ± 0.051 (range 0.29–0.43) and 0.94 ± 0.023 (range 0.89–0.96) between UMI-agnostic denoising and single-strand- and duplex-corrected datasets, respectively; Fig. 2d,e and Supplementary Fig. 6). We found similar improvements when measuring cosine

similarities between the clonal hematopoiesis signature and healthy controls, highlighting the importance of duplex correction for accurate signature analysis and demonstrating that clonal hematopoiesis is an abundant source of mutations in ccfDNA WGS (Fig. 2d,e, Supplementary Fig. 7 and Extended Data Fig. 6a).

As de novo mutation identification in error-corrected ccfDNA WGS delivers profiles matching SBS7a, SBS7b and clonal hematopoiesis signatures for identifying melanoma and age-associated circulating DNA fragments, respectively, we developed a tumor-agnostic approach for ctDNA detection based on mutational patterns. First, we tabulated the trinucleotide frequencies of plasma ccfDNA mutations,

fitted to reference mutational signatures (SBS7a, SBS7b and clonal hematopoiesis) using a non-negative likelihood model⁴⁵. We obtained relative contributions of the reference mutational signatures and estimated a tumor score by taking the weight of the tumor-associated SBS7a and SBS7b signatures and multiplying by the number of variants per duplex GEs sequenced (Supplementary Table 6).

We applied our signature-based ctDNA detection platform for preoperative ctDNA detection (tumor-agnostic ctDNA detection), sequencing plasma samples from eight individuals with resectable locoregional stage III melanoma (without tumor or normal DNA) and ten healthy individuals. In our cohort, tumor fractions measured by ichorCNA were undetectable in samples from individuals with stage III disease, suggesting tumor fractions below 3% (ichorCNA³⁰ limit of detection; Extended Data Fig. 7). Tumor scores were separable between control, stage III preoperative melanoma and pretreatment stage IV melanoma samples (2.4 ± 7.6 , 69.9 ± 65.72 and 340.7 ± 238.6 , respectively; Fig. 2f; trinucleotide frequencies are shown in Supplementary Fig. 8). When applying the same signature-based methodology for single-read variant calling in single-strand correction and UMI-agnostic correction, we obtained 0 detections in our melanoma cohort (Extended Data Fig. 6b), owing to the high number of preanalytical errors in single-strand-corrected and UMI-agnostic datasets. Previous studies showed that undetectable ctDNA using targeted panels is associated with favorable prognosis in stage III melanomas^{46,47}. However, individuals with undetectable ctDNA often still experience disease recurrence, highlighting the need for more sensitive tools for improved stratification. Here, tumor scores for ctDNA detection showed strong separation between healthy individuals and individuals with melanoma (Fig. 2f), suggesting that deep error-corrected sequencing can identify ctDNA in a nontumor-informed approach using mutational signatures, even when ctDNA burden in the plasma is low, such as in resectable melanoma.

To analytically validate our approach, we performed an *in silico* mixing study combining duplex-denoised reads from a high-burden ctDNA sample (MEL-12.B, 11% ichorCNA tumor fraction estimate) and a healthy control sample (CTRL-06) at $6.5\times$ duplex sequencing depth at expected tumor fractions of 0 and 5×10^{-4} to 10^{-1} (Fig. 2g). Tumor scores were detectable (AUC of 0.89) at tumor fractions greater than or equal to 10^{-3} . Signature scores dropped at tumor fractions at or below 5×10^{-4} (AUC of 0.63), given that the number of mutations originating from ctDNA was significantly below the number of mutations arising from healthy ccfDNA (mean of 42 and 1,930 variants at 5×10^{-4} from MEL-12.B versus from healthy individuals, respectively). Tumor scores correlated strongly with expected tumor fractions (Spearman's $\rho = 0.97$, $P < 2.2 \times 10^{-16}$). Importantly, these results highlight that we can identify ctDNA contributions when the number of tumor-derived variants is below the number of variants originating from background biological processes, such as clonal hematopoiesis (~85 cancer-derived mutations in ~1,920 background mutations at a tumor fraction of 10^{-3}), which are expected to dominate rare variant signal in early-stage and MRD contexts.

We further tested the ability of duplex WGS to track therapeutic response longitudinally. Here, we performed duplex WGS on the plasma of individuals with melanoma (stage III and stage IV) before and after treatment (Extended Data Fig. 8), including individuals receiving immunotherapy treatment with plasma collected serially ($n = 5$ individuals, three to five samples per individual collected at a pretreatment time point and up to 6 months after treatment), and individuals receiving surgery, with plasma collected before and after surgery (four individuals, two to three samples per individual, post-operative collection up to 6 months). ctDNA levels in individuals who had a partial response or who were recurrence free decreased after treatment decreased, whereas the ctDNA in individuals who showed a recurrence or who had progressive disease increased from their baseline time point (Extended Data Fig. 8a). Specifically, individuals

with progressive disease or recurrence saw a mean increase of 500 melanoma variants per GE after treatment, whereas individuals with a partial response or who were recurrence free from disease exhibited an overall decrease in ctDNA of 196 melanoma variants per GE ($P = 0.0005$ between groups; Extended Data Fig. 8b). Although clinical studies are necessary to validate these findings, our results suggest a potential for duplex-corrected WGS in tumor-agnostic, plasma-only serial profiling of individuals with melanoma to assess therapeutic response.

Duplex WGS resolves complex mutational signatures in ccfDNA

Next, we sought to determine whether duplex WGS could deconvolute complex signatures arising from multiple biological processes, such as APOBEC3-derived processes in urothelial cancer. Many urothelial cancers carry mutations indicating APOBEC cytidine deaminase activity, where C>T transitions can occur as a function of uracil generation by cytidine deaminase activity (SBS2), as well as C>G and C>A mutations that arise from polymerase errors following uracil excision (SBS13)⁴⁸. Thus, measuring the contributions of SBS2 and SBS13 as an APOBEC score could be potentially used to detect ctDNA in individuals with urothelial cancer. Moreover, in neoadjuvant-treated and unresectable or metastatic urothelial carcinomas, platinum-based chemotherapies are often used as first-line treatment, and their use can induce mutagenesis in tumor tissue, which can result in subclonal mutation propagation^{49–51}. Platinum-based chemotherapies mainly affect C[C>T]C and C[C>T]T trinucleotides (SBS31 and SBS35)⁴⁴, thereby adding an additional layer of complexity to an individual's cancer mutational compendium. To extend our mutation signature analysis to urothelial cancer, we sequenced $n = 20$ plasma samples from $n = 20$ individuals with urothelial cancer (stage II–IV). Of these 20 individuals, 11 had previously received neoadjuvant platinum chemotherapy. We hypothesized that this complex mutational compendium, with mutations from platinum chemotherapy and varying APOBEC mutagenesis, would benefit from a ccfDNA whole-genome approach, where distinct mutational signatures can be detected. We measured and summed the contributions of SBS2 and SBS13 as an APOBEC score for individuals with urothelial cancer, finding that urothelial cancer ccfDNA is enriched for mutations containing APOBEC signal compared to ccfDNA from healthy individuals ($P = 0.0031$; Fig. 3a), and the number of SBS2-derived mutations correlated positively with the detection of SBS13-derived mutations (Pearson's $R = 0.94$, $P = 3.2 \times 10^{-14}$).

In a subset of individuals with urothelial cancer ($n = 13$), tumor tissue was available for sequencing ($n = 29$ multiregion samples across 13 individuals), where we could compare the mutational profiles of plasma to the individual-matched tumors. First, all tumors showed evidence of APOBEC mutagenesis (8–86%; Extended Data Fig. 9 and Supplementary Table 7). Second, mutational profiles obtained through *de novo* (without using a matched tumor; Supplementary Fig. 9) detection of ctDNA strongly resembled those of the matching tumor (cosine similarity of >0.83 at plasma tumor fractions of >0.005), and *de novo* identification of APOBEC mutations was possible in plasma samples with tumor fractions as low as 1.1×10^{-4} (Fig. 3a,b and Supplementary Fig. 10). We measured tumor-informed tumor fractions (Methods and Supplementary Table 8) and calculated the expected contribution of APOBEC-derived mutations in the plasma by multiplying the tumor-informed tumor fraction by the fraction of tumor variants attributable to APOBEC mutational signatures. We found a strong correlation between the expected, tumor-informed APOBEC contributions in the plasma and our *de novo* APOBEC score (Pearson's $R = 0.95$, $P = 1.13 \times 10^{-6}$; Fig. 3c).

In terms of driver detection, as expected from the lower sequencing depth, duplex WGS detected only 6 of 59 putative driver mutations (Supplementary Table 9) across the cohort of individuals with matched plasma and tumor sequencing ($n = 29$ multiregion samples across the 13 individuals), preferentially in ccfDNA samples with high tumor burden (Supplementary Tables 8 and 9; Pearson's $R = 0.57$ and $P = 0.04$).

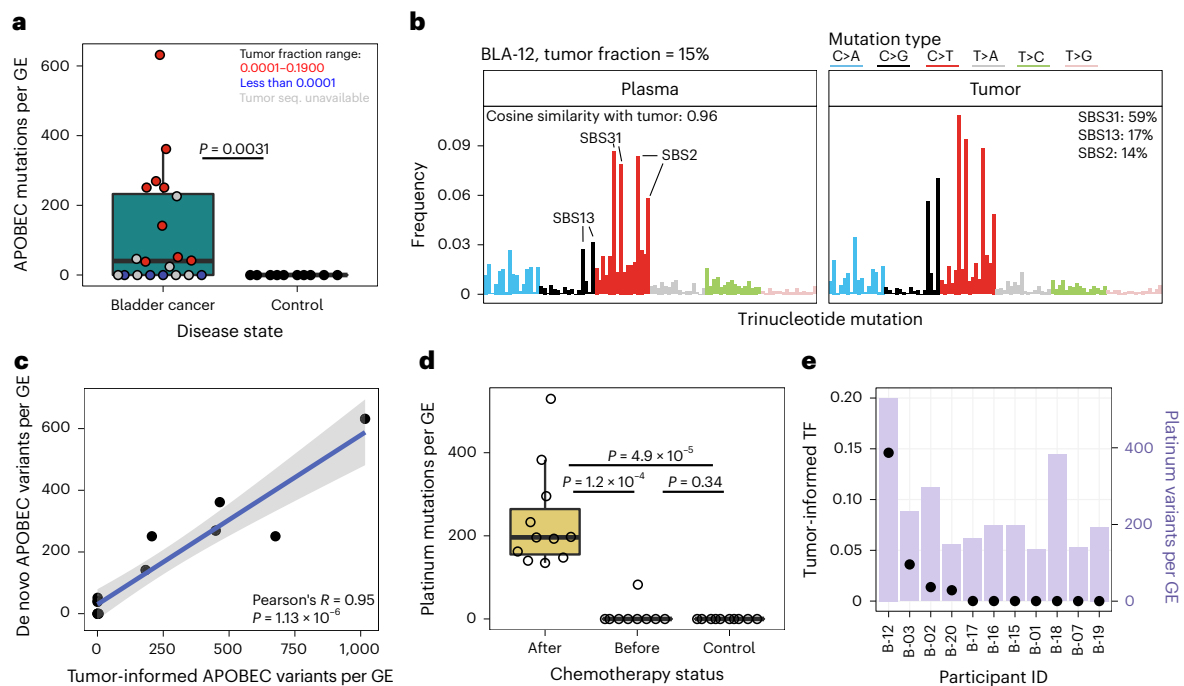


Fig. 3 | Mutational signature analysis of cell-free DNA from individuals with urothelial cancer. **a**, Plasma mutational scores from APOBEC mutagenesis (SBS2 and SBS13) in individuals with urothelial cancer (stage II–IV; $n = 20$) and healthy individuals ($n = 10$). Plasma signatures were fit to a custom reference of signatures comprising SBS2, SBS13, SBS31 and SBS35 (Cosmic v3.3) and clonal hematopoiesis⁴². Tumor fractions were measured in a tumor-informed manner (see Methods). **b**, Trinucleotide frequency of ccfDNA (‘plasma’; left) and tumor (right) for BLA-12. For the tumor, the three highest SBS contributions are highlighted (SBS31, SBS13 and SBS2). Contributions were fit to the entire Cosmic v3.3 catalog. **c**, Tumor-informed measurement of APOBEC-derived mutations (SBS2 and SBS13, x axis) versus tumor-agnostic APOBEC-derived mutations (y axis) in $N = 13$ samples for which tumor tissue sequencing was available. The tumor-agnostic APOBEC mutations were measured as in **a**. Plasma signatures were fit to a custom reference of signatures comprising SBS2, SBS13, SBS31 and

SBS35 (Cosmic v3.3) and clonal hematopoiesis⁴². Tumor signatures were fit to the entire Cosmic v3.3 catalog. The shaded area represents the 95% confidence interval of the data distribution. **d**, Plasma mutational scores from platinum therapy mutagenesis (SBS31 + SBS35) in individuals with urothelial cancer (stage II–IV; $n = 11$ after, $n = 9$ before) and healthy individuals ($n = 10$). Plasma signatures were fit to a custom reference of signatures comprising SBS2, SBS13, SBS31 and SBS35 (Cosmic v3.3) and clonal hematopoiesis⁴². **e**, Bar plot of tumor-informed tumor fractions (black dots) and platinum scores (purple; as calculated in **d**) for samples with available tumors for sequencing. For box plots in **a** and **d**, the bottom and top ends of the boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the IQR. P values were calculated using a two-sided Wilcoxon test.

Although these results demonstrate that our duplex WGS approach can resolve putative driver mutations, deep targeted sequencing could offer more benefit for driver mutation detection.

Interestingly, we noted instances of high cosine similarities between plasma and matched tumors despite low-burden disease (at or below 10^{-4} ; Supplementary Fig. 10b; for example, BLA-18 (cosine 0.92)). Although stereotypical APOBEC-derived peaks (namely T[C>T]A and T[C>T]T for SBS2 and T[C>G]A and T[C>G]T for SBS13) were not visually obvious in the plasma and were undetectable in $n = 8$ individuals (Fig. 3a), we noted strong contributions of platinum-derived chemotherapy mutations (for example, C[C>T]C and C[C>T]T for SBS31) in individuals who were treated with platinum chemotherapy. By fitting the plasma mutational signatures to a custom catalog of APOBEC- and platinum chemotherapy-derived mutational signatures (SBS2/SBS13 and SBS31/SBS35, respectively), we found enrichment of platinum chemotherapy mutations only in the plasma of individuals who had received treatment (Fig. 3d). Interestingly, platinum mutation signatures were detected in the ccfDNA through de novo analysis in some individuals where no ctDNA was detected using a tumor-informed approach (Fig. 3e and Supplementary Fig. 10), potentially reflecting signal originating from platinum mutagenesis in nonmalignant cells. These results suggest that tumor-agnostic de novo profiling of plasma may offer a more comprehensive overview of plasma from individuals with cancer than tumor-informed approaches that solely focus on the detection of a limited number of clonal mutations.

Discussion

Radical decreases in sequencing costs open new opportunities in both clinical genomics and basic biology research. Here, we harnessed low-cost Ultima sequencing to demonstrate the impact of deeper WGS ($\sim 100\times$) on tumor-informed ctDNA monitoring. Moreover, we also developed methodology to integrate duplex sequencing with single-end Ultima sequencing, showcasing the feasibility of cost-efficient, highly accurate WGS that can be broadly applied across genomics research. This advance allowed us to apply duplex error correction to clinical plasma samples at the level of the entire genome. We leveraged these highly error-corrected data for nontumor-informed ctDNA detection based on the similarity of mutational profiles to known cancer mutational signatures. These findings have important clinical implications, as uncoupling ctDNA detection from a tumor-informed mutation profile radically increases the potential use of ctDNA monitoring in common clinical scenarios where tumor samples cannot be obtained (Supplementary Note). Excitingly, this demonstration opens up the possibility for the use of WGS in ctDNA cancer screening. In particular, such an approach might be beneficial for specific settings where there is high genetic (for example, *BRCA* mutation carriers and Lynch syndrome) or environmental (for example, tobacco smoke exposure) cancer risk together with distinct mutational signatures. A limitation of whole-genome duplex sequencing is the fact that only a fraction of reads result in a duplex fragment, resulting in one-order-of-magnitude fewer unique molecules captured than standard WGS. However, the

~500-fold decrease in error rates in duplex sequencing compared to bioinformatically denoised standard WGS may offset this decrease in coverage, depending on the specific application or context. In the tumor-informed setting, where prior knowledge of the somatic mutations can lower the effective error rate of standard WGS, higher inputs are expected to afford lower limits of detection than duplex WGS. In the tumor-agnostic setting, a markedly reduced error rate allows for detection of genome-wide signatures and plasma-only MRD even in lower-burden disease compared to standard WGS, which only allowed detection in high-burden metastatic disease. Finally, we envisage that methods developed herein can provide a general cost-efficient approach for duplex WGS that can be leveraged across genomics research that requires highly accurate detection of rare variants in somatic tissues.

Online content

Any methods, additional references, Nature Portfolio 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/s41592-025-02648-9>.

References

- Cohen, J. D. et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* **359**, 926–930 (2018).
- Sanz-Garcia, E., Zhao, E., Bratman, S. V. & Siu, L. L. Monitoring and adapting cancer treatment using circulating tumor DNA kinetics: current research, opportunities, and challenges. *Sci. Adv.* **8**, eabi8618 (2022).
- Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
- Wan, J. C. M. et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).
- Wang, S. et al. Potential clinical significance of a plasma-based KRAS mutation analysis in patients with advanced non-small cell lung cancer. *Clin. Cancer Res.* **16**, 1324–1330 (2010).
- Murtaza, M. et al. Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).
- Diehl, F. et al. Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* **14**, 985–990 (2008).
- Agarwal, R. et al. Dynamic molecular monitoring reveals that SWI-SNF mutations mediate resistance to ibrutinib plus venetoclax in mantle cell lymphoma. *Nat. Med.* **25**, 119–129 (2019).
- Newman, A. M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
- Newman, A. M. et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **34**, 547–555 (2016).
- Kurtz, D. M. et al. Enhanced detection of minimal residual disease by targeted sequencing of phased variants in circulating tumor DNA. *Nat. Biotechnol.* **39**, 1537–1547 (2021).
- Cohen, J. D. et al. Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands. *Nat. Biotechnol.* **39**, 1220–1227 (2021).
- Chaudhuri, A. A. et al. Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. *Cancer Discov.* **7**, 1394–1403 (2017).
- Zviran, A. et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat. Med.* **26**, 1114–1124 (2020).
- Abbosh, C. et al. Phylogenetic ctDNA analysis depicts early stage lung cancer evolution. *Nature* **545**, 446–451 (2017).
- Bettegowda, C. et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra24 (2014).
- Gale, D. et al. Residual ctDNA after treatment predicts early relapse in patients with early-stage non-small cell lung cancer. *Ann. Oncol.* **33**, 500–510 (2022).
- Tie, J. et al. Circulating tumor DNA analysis guiding adjuvant therapy in stage II colon cancer. *N. Engl. J. Med.* **386**, 2261–2272 (2022).
- Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
- Hoang, M. L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **113**, 9846–9851 (2016).
- Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
- Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).
- Meddeb, R. et al. Quantifying circulating cell-free DNA in humans. *Sci. Rep.* **9**, 5220 (2019).
- Widman, A. J. et al. Ultrasensitive plasma-based monitoring of tumor burden using machine-learning-guided signal enrichment. *Nat. Med.* **30**, 1655–1666 (2024).
- National Human Genome Research Institute. DNA sequencing costs: data. <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data> (2023).
- Almog, G. et al. Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.05.29.493900> (2022).
- Simmons, S. K. et al. Mostly natural sequencing-by-synthesis for scRNA-seq using Ultima sequencing. *Nat. Biotechnol.* **41**, 204–211 (2023).
- Hasenleithner, S. O. & Speicher, M. R. A clinician's handbook for using ctDNA throughout the patient journey. *Mol. Cancer* **21**, 81 (2022).
- Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
- Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
- Rose Brannon, A. et al. Enhanced specificity of clinical high-sensitivity tumor mutation profiling in cell-free DNA via paired normal sequencing using MSK-ACCESS. *Nat. Commun.* **12**, 3770 (2021).
- Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
- 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).
- Cindy Yang, S. Y. et al. Pan-cancer analysis of longitudinal metastatic tumors reveals genomic alterations and immune landscape dynamics associated with pembrolizumab sensitivity. *Nat. Commun.* **12**, 5137 (2021).
- Liu, M. H. et al. DNA mismatch and damage patterns revealed by single-molecule sequencing. *Nature* **630**, 752–761 (2024).
- Bae, J. H. et al. Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. *Nat. Genet.* **55**, 871–879 (2023).


37. Thompson, J. C. et al. Detection of therapeutically targetable driver and resistance mutations in lung cancer patients by next generation sequencing of cell-free circulating tumor DNA. *Clin. Cancer Res.* **22**, 5772–5782 (2016).
38. Hu, Y. et al. False-positive plasma genotyping due to clonal hematopoiesis. *Clin. Cancer Res.* **24**, 4437–4443 (2018).
39. Abbosh, C., Swanton, C. & Birkbak, N. J. Clonal haematopoiesis: a source of biological noise in cell-free DNA analyses. *Ann. Oncol.* **30**, 358–359 (2019).
40. Shaw, J. A. et al. Serial postoperative circulating tumor DNA assessment has strong prognostic value during long-term follow-up in patients with breast cancer. *JCO Precis. Oncol.* **8**, e2300456 (2024).
41. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
42. Osorio, F. G. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep.* **25**, 2308–2316 (2018).
43. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
44. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
45. Jin, H. et al. Accurate and sensitive mutational signature analysis with MuSiCal. *Nat. Genet.* **56**, 541–552 (2024).
46. Tan, L. et al. Prediction and monitoring of relapse in stage III melanoma using circulating tumor DNA. *Ann. Oncol.* **30**, 804–814 (2019).
47. Lee, J. H. et al. Pre-operative ctDNA predicts survival in high-risk stage III cutaneous melanoma patients. *Ann. Oncol.* **30**, 815–822 (2019).
48. Petljak, M. et al. Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature* **607**, 799–807 (2022).
49. Findlay, J. M. et al. Differential clonal evolution in oesophageal cancers in response to neo-adjuvant chemotherapy. *Nat. Commun.* **7**, 11111 (2016).
50. Boot, A. et al. In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res.* **28**, 654–665 (2018).
51. Nguyen, D. D. et al. The interplay of mutagenesis and ecDNA shapes urothelial cancer evolution. *Nature* **635**, 219–228 (2024).

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2025

Alexandre Pellan Cheng ^{1,2,3,4,16} , **Adam J. Widman**^{1,5,16}, **Anushri Arora**^{1,2,16}, **Itai Rusinek** ⁶, **Aaron Sossin**^{1,2}, **Srinivas Rajagopalan**^{1,2}, **Nicholas Midler**^{1,2}, **William F. Hooper** ¹, **Rebecca M. Murray** ^{1,2}, **Daniel Halmos**^{1,2}, **Theophile Langanay**^{1,2}, **Hoyin Chu**², **Giorgio Inghirami** ², **Catherine Potenski**^{1,2}, **Soren Germer** ¹, **Melissa Marton**¹, **Dina Manaa**¹, **Adrienne Helland**¹, **Rob Furatero**¹, **Jaime McClintock**¹, **Lara Winterkorn**¹, **Zoe Steinsnyder**¹, **Yohyoh Wang**^{1,2}, **Asrar I. Alimohamed**⁷, **Murtaza S. Malbari**², **Ashish Saxena**², **Margaret K. Callahan**⁵, **Dennie T. Frederick**⁷, **Lavinia Spain**^{8,9}, **Michael Sigouros**¹⁰, **Jyothi Manohar**¹⁰, **Abigail King** ¹⁰, **David Wilkes** ¹⁰, **John Otilano**¹⁰, **Olivier Elemento** ^{10,11}, **Juan Miguel Mosquera** ^{10,11,12}, **Ariel Jaimovich**⁶, **Doron Lipson**⁶, **Samra Turajlic**^{8,9}, **Michael C. Zody** ¹, **Nasser K. Altorki** ², **Jedd D. Wolchok**^{2,13,14}, **Michael A. Postow**^{2,5}, **Nicolas Robine** ¹, **Bishoy M. Faltas**^{2,10,15,17}, **Genevieve Boland** ^{7,17} & **Dan A. Landau** ^{1,2,17} 

¹New York Genome Center, New York, NY, USA. ²Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, Weill Cornell Medical College, New York, NY, USA. ³Département de Génie des Systèmes, École de Technologie Supérieure, Montréal, Québec, Canada. ⁴Axe Cancer, Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal, Québec, Canada. ⁵Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁶Ultima Genomics, Fremont, CA, USA. ⁷Mass General Cancer Center, Massachusetts General Hospital, Boston, MA, USA. ⁸Cancer Dynamics Laboratory, The Francis Crick Institute, London, UK. ⁹Renal and Skin Unit, The Royal Marsden NHS Foundation Trust, London, UK. ¹⁰Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY, USA. ¹¹Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. ¹²Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. ¹³Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA. ¹⁴Ludwig Institute for Cancer Research, New York, NY, USA. ¹⁵Department of Cell and Developmental Biology, Weill Cornell Medicine, New York, NY, USA. ¹⁶These authors contributed equally: Alexandre Pellan Cheng, Adam J. Widman, Anushri Arora. ¹⁷These authors jointly supervised this work: Bishoy M. Faltas, Genevieve Boland, Dan A. Landau,  e-mail: alexandre.cheng@etsmtl.ca; dlандаu@nygenome.org

Methods

Simulation analysis

Simulations for ctDNA detection scores (Fig. 1a) were performed assuming a tumor mutational compendium of 10,000 SNVs with different error rates (10^{-3} , 10^{-4} and 10^{-5}), coverages (1, 10 and 100) and tumor fractions (0 , 10^{-6} – 10^{-5}). For each of the 10,000 SNV mutations, coverage was simulated using a Poisson distribution. Each simulated sequenced base pair was classified as either ctDNA or ccfDNA according to the tumor fraction, and errors misclassified as ctDNA were determined according to the error rate. Estimated tumor fractions were calculated by summing the ctDNA molecules and the errors and dividing by the total base pairs simulated. Z scores were calculated using the following equation, where *TF* indicates tumor fraction, and *s.d.* indicates standard deviation:

$$Z \text{ score} = \frac{TF - \text{mean}(TF_{TF=0})}{s.d._{TF=0}}$$

PCAWG datasets

We obtained somatic trinucleotide counts of $n = 2,780$ tumors from the PCAWG database⁴³. Total somatic mutation counts were used to generate Extended Data Fig. 3. To simulate and compare the performance of targeted panels and WGS to detect different cancer types, we defined ctDNA detection opportunities to be a function of panel size, GEs available to be sequenced in a sample and sequencing coverage. Specifically,

$$\begin{aligned} \text{ctDNA detection opportunities} \\ = (\text{targeted SNVs} \times \text{GEs} \mid \text{targeted SNVs} \times \text{coverage}) \end{aligned}$$

For WGS simulations, the ‘panel size’ was set to the median number of somatic mutations of a given cancer type. We assumed a 20-ng cell-free DNA input (corresponding to ~6,700 GEs, assuming the mass of a genome to be 3 pg).

Human sample processing

Blood and tissue samples were obtained from individuals after obtaining informed consent and following protocols approved by institutional review boards (IRBs) and in accordance with the Declaration of Helsinki protocol. Samples were obtained from New York Presbyterian/Weill Cornell Medical Center (IRB numbers 0201005295 (Tumor Biobanking), 1008011210 (GU Tumor Biobanking), 1011011386 (Urothelial Cancer Sequencing), 100701157 (Genomic and Transcriptomic Profiling), 1305013903 (Precision Medicine), 1708018519 (Cardiac Surgery Biobank), 2014-0024 (approved by the Institutional Animal Care and Use Committee at Weill Cornell Medicine) and 1610017682 (ctDNA for Early Detection and Management of Non-Small Cell Lung Cancer)), Memorial Sloan Kettering Cancer Center (IRB number 12-245 (Genomic Profiling in Cancer Patients)), Massachusetts General Hospital (IRB number 11-181 (Collection of Tissue and Blood Specimens and Clinical Data from Patients with Melanoma and Other Cutaneous Malignancies)) or the Royal Marsden NHS Foundation Trust in the United Kingdom (Supplementary Table 1). Tumor, normal and plasma samples from the Royal Marsden NHS Foundation Trust were obtained under an approved ethical protocol (Melanoma TRACERx, Research Ethics Committee Reference 11/LO/0003). Cancer diagnosis was established according to World Health Organization criteria and confirmed in all cases by an independent pathology review. Participants did not receive any compensation.

Tumor and germline DNA extraction, library preparation and sequencing

For melanoma and lung cancer samples, genomic DNA was extracted using a QiAamp DNA Mini kit (Qiagen, 56304) and QiAamp DNA Blood kit (Qiagen, 51104) for tissue and blood samples, respectively, and

sheared to 450 bp (Covaris, 500569). Sequencing libraries were prepared from 1 µg of DNA using a TruSeq DNA PCR-Free Library Preparation kit (Illumina, 20015963), with one additional bead cleanup performed after end repair and after adapter ligation. DNA was quantified using a Qubit 3.0 fluorometer, and length analysis was performed using an Agilent Bioanalyzer or High Sensitivity Fragment Analyzer. Paired-end sequencing (2×150 bp) was performed on either a HiSeq X or NovaSeq v1.0 Illumina machine. Urothelial tumor/normal sequencing data were obtained from Nguyen et al.⁵¹.

Cell-free DNA extraction

Cell-free DNA was extracted from plasma using a Magbind ccfDNA extraction kit (Omega Biotek, M3298). Manufacturer recommendations for extraction were followed, but elution volume was increased to 35 µl, and elution time was increased to 20 min on a thermomixer at 1,600 rpm (room temperature). Extracted ccfDNA was quantified using a Qubit 3.0 fluorometer, and length analysis was performed using an Agilent Bioanalyzer or High Sensitivity Fragment Analyzer.

Cell-free DNA library preparation

Whole-genome library preparation (without duplex). Next-generation sequencing libraries were generated using a double-stranded preparation kit (Kapa Hyper Prep kit, Roche, KK8502). Full-length adapters (IDT TruSeq UDI plate, Illumina, 20023784) were used for adapter ligation. Six PCR cycles were performed when input DNA was above 5 ng, and eight cycles were performed when the input was below 5 ng. Libraries were quantified using a Qubit 3.0 fluorometer, and length analysis was performed using an Agilent Bioanalyzer or High Sensitivity Fragment Analyzer. Illumina sequencing libraries were sequenced on a HiSeq X or NovaSeq1.0 using 2×150 bp paired-end sequencing. Library input amounts can be found in Supplementary Table 3.

Whole-genome library preparation (with duplex). *Version 1.* ccfDNA libraries were generated in a similar fashion as described above, although the full-length adapters were replaced with stubby Y adapters containing three UMI bases (IDT Duplex Seq adapters, 1080799), and sample indexing was performed during PCR amplification. To enhance duplicate recovery in human samples, a maximum of 10 ng was used as input, and 4 ng of prepared libraries was subjected to six additional PCR cycles before Ultima library conversion. Mouse PDX samples did not undergo additional PCR cycles before Ultima library conversion.

Version 2. Version 2 of duplex library preparation differs from version 1 as follows: (1) 1 ng of ccfDNA was used as input, (2) eight cycles of PCR were performed, and (3) the step of taking 4 ng of prepared libraries and performing six additional PCR cycles was omitted. Library input amounts can be found in Supplementary Table 3.

Ultima sequencing. Illumina sequencing libraries underwent Ultima library conversion. Briefly, Illumina libraries were converted to Ultima libraries by PCR using primers matching Illumina read 1 and read 2 sequences and containing Ultima-specific barcodes (R1 conversion adapter: 5' TCC ATC TCA TCC CTG CGT GTC TCC TGC ACA ATG TGT GCT AGA TCT ACA CGA CGC TCT TCC GAT CT 3'; R2 conversion adapter: 5' CTG TGT GCC TTG GCA GTC TCA GCT CAG ACG TGT GCT CTT CCG ATC T 3'). Samples were then pooled and sequenced on an Ultima sequencer prototype to a target depth of $120\times$. One sample achieved lower effective coverage (MEL-03.A, $36\times$), likely due to a sample pooling error.

WGS (without duplex) adapter trimming and alignment

Illumina fastQ reads were adapter trimmed using skewer⁵² (version 0.2.2). Trimmed reads were then aligned to the human genome (version hg38) using bwa mem⁵³. Duplicate reads were marked in a UMI-unaware

fashion using novosort⁵⁴. Depth of coverage was estimated using mosdepth⁵⁵ (version 0.2.9), and duplicate reads were considered. For Ultima reads, adapter trimming was performed to remove the Illumina conversion adapters. Cutadapt⁵⁶ (version 2.10; cutadapt-mask-adapt-a-CTA-CACGACGCTCTCCGATCT; max_error_rate = 0.15; min_overlap = 10; required...AGATCGGAAGAGCACACGTCTGCTG; max_error_rate = 0.2; min_overlap = 6) was used to mask adapter sequences, and adapter trimming was then performed using GATK⁵⁷ (private Ultima fork, since merged to the latest 4.3.0.0 GATK release; ClipReads function). Alignment was performed using bwa mem⁵³ (version 0.7.15-r1140), and coverage was estimated using mosdepth counting duplicate reads. Alignment statistics can be found in Supplementary Table 3. We note that coverage was calculated as the total number of bases mapping to the human genome divided by the size of the genome. In cell-free DNA, coverage has been shown to be inversely correlated with DNA accessibility, as the open regions of the genome are more susceptible to degradation in the blood. For example, highly expressed genes can see a loss of coverage of nearly 50%. Thus, somatic variants in open chromatin regions may be more difficult to detect via the plasma.

Copy-number-based tumor fraction estimation

Genome-wide coverage was calculated over a 1-Mbp window and normalized for mappability and GC content biases (using hmmscopy⁵⁸ version 0.99). Tumor fractions were estimated using ichorCNA³⁰ (version 0.3.2) after correcting for library and sequencing artifacts via a panel of normals from healthy individuals (CTRL-01 to CTRL-05). A separate panel of normals was created for Illumina- and Ultima-sequenced samples using libraries sequenced on the respective machines. For plotting purposes (Fig. 1d), corrected log₂ (read counts) outputted by ichorCNA were used. Bins marked by ichorCNA as copy gains, amplifications and high-level amplifications were marked and colored as chromosome gains (pink). Bins marked as homozygous deletion states and hemizygous deletions were marked and colored as chromosome losses (blue). Copy neutral regions were marked as neutral (black). Bins with corrected log₂ (read counts) between -0.05 and 0.05 were also marked as neutral (black). Given that ichorCNA provides multiple solutions ordered by log likelihood, the CNV-based tumor fraction reported in the manuscript was manually selected among all solutions according to ichorCNA guidelines³⁰. Deviation from the most likely solution is justified in Supplementary Table 10.

WGS (without duplex) SNV-based tumor fraction estimation

SNV-based tumor fraction estimation was performed by counting cell-free DNA reads with matching tumor-specific somatic mutations. To limit the effect of problematic regions of the genome, a platform-specific blacklist was built. For Illumina sequencing, regions identified in the ENCODE blacklist⁵⁹, centromeres⁶⁰, simple repeat regions⁶⁰ and positions with high mutation rates (GNOMAD⁶¹, allele frequency > 0.001) were not considered. For Ultima sequencing, Ultima-specific low-confidence regions composed of homopolymers, AT-rich regions, tandem repeats and regions with poor mappability and high coverage variability were additionally excluded (Extended Data Fig. 3). To limit the effect of sequencing errors, custom scripts were used for platform-specific denoising (Supplementary Note).

WGS (without duplex) tumor-informed error rate estimation

Tumor-informed error rates were computed by intersecting a given healthy individual's cell-free DNA sequencing reads with the somatic mutations from an individual with cancer. Reads were then denoised in a platform-specific manner as described above (except for data in Supplementary Fig. 2, which compares denoised data to nondenoised data). An error rate was defined as the total number of single-occurring variants divided by the total number of denoised bases overlapping the somatic mutations.

PCR duplicates analysis for indel and SNV error rates

To measure indel and SNV error rates, alignment files were split by chromosome. For each autosomal chromosome, up to 1,500,000 unique DNA molecules were collected and scanned for PCR duplicates (here defined as two sequencing reads mapping to the same strand of the reference genome, each having a mapping quality of 60 and both containing the same UMI). For each unique molecule, two PCR duplicates were randomly selected. The PCR duplicates were then aligned to each other and the reference genome (obtained via the alignment file) using global pairwise sequencing alignment (using the pairwise2 module from biopython version 1.79 in Python version 3.6). Alignment points and penalties were set to 1, -1.5, -1 and -1 for identical bases, nonidentical bases, opening gaps and extending gaps, respectively. Gaps in the alignment were considered to be indels, and differing bases were considered to be SNV errors. For gaps, the size differences between homopolymers of the PCR duplicates were tabulated to create Extended Data Fig. 1. For differing bases, the reference trinucleotide, reference base and the PCR duplicate read bases were collected for Extended Data Fig. 2. For the SNV analysis, the reference base was assumed to be correct when PCR duplicates differed.

In silico mixing study for analytical lower limit of detection estimation (standard WGS)

We created in silico mixes at various tumor fractions by computationally combining aligned reads from a high-tumor-burden plasma sample (MEL-07, estimated tumor fraction of 7%) with aligned reads from a healthy individual (CTRL-05). Reads were mixed to create 20× (Illumina and Ultima), 40× (Ultima) and 70× (Ultima) bam files harboring 10⁻⁶, 5 × 10⁻⁵, 10⁻⁵, 10⁻⁴, 10⁻³ or 10⁻² tumor fractions. The coverage of the high-burden sample necessary to obtain a given expected tumor fraction at an expected coverage was defined as

$$\text{Coverage needed}_{\text{high-burden sample}} = \frac{\text{Expected tumor fraction}}{\text{high-burden tumor fraction}} \times \text{expected coverage}$$

Coverage of the healthy individual was subsequently defined as

$$\text{Coverage needed}_{\text{healthy control}} = \text{expected coverage} - \text{coverage needed}_{\text{high-burden sample}}$$

An in silico mixed replicate was then obtained by randomly down-sampling MEL-07 and CTRL-05 to obtain the respective coverages (using samtools view -s). Downsampled files were merged and denoised as described above (WGS (without duplex) SNV-based tumor fraction estimation). Tumor fractions were estimated with platform-specific denoising. Z scores were calculated as

$$Z \text{ score} = \frac{TF - \text{mean}(TF_{TF=0})}{\text{s.d.}_{TF=0}}$$

Specifically, the Z score of a given replicate was calculated against the mean and standard deviation of coverage and platform-matched healthy control replicates.

UMI WGS data processing

FastQ reads were adapter and UMI trimmed using cutadapt67 (version 2.10). Trimmed reads were then aligned to the human genome (version hg38) using bwa mem64 (with parameters -K 100000000 -p -v3 -t 16 -Y). Trimmed UMIs were added to the alignment files as an additional RX tag. Unique molecules were identified by in either single-stranded mode (that is, collecting PCR duplicates as unique molecules) or in duplex mode (that is, collecting PCR duplicates and UMI-concordant Watson and Crick strands as unique molecules) using the fgbio suite of tools (version 2.0). Because duplex correction via

fgbio requires paired-end reads, we created a synthetic R2 read directly from single-end bam alignment files. R2 reads were built using the same mapping information (such as CIGAR string and mapping quality) and read information (such as sequence and qualities) as the R1 read. The subsequent paired-end alignment file was grouped by UMI (fgbio GroupReadsByUMI with parameters -m0 -s paired -e 1).

Single-stranded and duplex consensus reads and UMI-agnostic denoising. Single-stranded consensus reads were created by following the fgbio workflow. Molecular consensus reads were generated using the CallMolecularConsensusReads command (with options -M1; -consensus-call-overlapping-bases false). Duplex analysis included the paired option. Next, FilterConsensusReads was applied (with options -M2; -N0; -E 1; -e 1; -n 1). Reads passing the filter were remapped to the human genome (version hg38) for analysis (or a concatenated human and mouse genome for mouse PDX samples; hg38 and mm39, respectively). Single-strand and/or duplex metrics (such as consensus read depth, consensus error rate, number of Ns on the consensus molecule and number of reads with matching UMIs) and mapping information were integrated as additional read tags to the original single-end alignment file. Variant frequencies in the original alignment files (without denoising) were calculated using lofreq (version 2.1.3a, with all filtering modes disabled). The original single-end bam, with the additional single-strand or duplex tags, was processed through the FlowFeatureMapper tool (described above), which allows for processing of the additional UMI tags, to obtain putative variants. For single-stranded consensus reads, only consensus variants obtained by the fgbio pipeline were considered. For duplex variants, a decision-tree-based classifier was used to create duplex consensus variants (Supplementary Note). For mouse PDX samples, the following filters were then applied: (1) all reads contributing to a consensus read must have a mapping quality of 60, (2) all reads contributing to a consensus must have the five flanking bases of a variant match the reference genome, and (3) the variant must not be within 25 bases of either end of the cell-free DNA molecule. Finally, UMI-agnostic denoising was performed by filtering by (1) variant position in read (the variant cannot be within 25 bp of an end of the read), (2) template length (must be lower than 200 bp), (3) mapping quality (cannot be below 60), (4) edit distance (must be below 4) and (5) total variants on the read (must be below 11). Duplicate reads were not considered. The same filters were applied for human samples, with the addition of an edit distance filter to further decrease errors (at least one duplex strand (or read for single-strand consensus) must have an edit distance below 2).

Residual SNV rate estimation

The residual SNV rate was defined as

$$\text{Residual SNV rate} = \frac{\text{Number of single – occurring mutations}}{\text{Interrogated bases}}$$

Here, the number of single-occurring mutations refers to the number of variants that only occur in a single denoised duplex, single strand or read (depending on the analysis performed), and the number of interrogated bases refers to the total number of bases that pass denoising filters. Given that the tools we developed in this manuscript are designed to detect variants and do not report reference bases, the number of interrogated bases was estimated according to

$$\text{Interrogated bases} = \text{total observed molecules} \times \text{molecule length} \\ \times \text{homozygous variant filtering ratio}$$

Here, total observed molecules refers to the total number of duplexes sequenced in a given region of the genome, and the molecule length was set to 170 bp. The product of these two variables estimates the total number of duplex (or single-strand) bases

sequenced. To account for the effect of filtering, we multiply the total number of bases by the fraction of homozygous variants that pass all filtering criteria.

Duplex depth estimation

The duplex coverage for each sample was estimated by multiplying the number of double-stranded DNA molecules recovered (that is, the number of unique molecules with at least one top and one bottom strand) by 170 (roughly the size of a cell-free DNA molecule) and dividing by the size of the genome (2,875,001,522).

Trinucleotide frequency tabulation and signature contribution estimation

After denoising, the trinucleotide frequencies of duplex-corrected SNVs (or single-stranded or UMI-agnostic-corrected SNVs) were tabulated using deconstructSigs⁶². Next, signature contributions were estimated using MuSiCal⁴⁵. Here, the previously tabulated trinucleotide frequencies and cancer-specific reference catalogs and specificity thresholds were used. In melanoma studies, MuSiCal was used to refit the sample trinucleotide frequencies to a catalog containing SBS7a, SBS7b (from Cosmic v3.3) and clonal hematopoiesis⁴². MuSiCal was run in likelihood-bidirectional mode using an empirically defined threshold of 0.007. This mode performs signature refitting using a likelihood-based sparse non-negative least squares algorithm. Urothelial cancer samples were run similarly, although the catalog was pre-defined to contain signatures from APOBEC mutagenesis (SBS2 and SBS13, Cosmic v3.3), platinum chemotherapy (SBS31 and SBS35, Cosmic v3.3) and clonal hematopoiesis⁴² (with a MuSiCal threshold of 0.025).

In silico mixing (duplex WGS)

The in silico mixing study was performed by computationally mixing variants from high-burden (HB) sample MEL-12.B (tumor fraction of 11.59%) with variants from cancer-free control (CFC) CTRL-06. First, the duplex coverage for each sample was estimated by multiplying the number of double-stranded DNA molecules recovered (that is, the number of unique molecules with at least one top and one bottom strand) by 170 (roughly the size of a cell-free DNA molecule) and dividing by the size of the genome (2,875,001,522). Here, MEL-12.B and CTRL-06 had coverages of 16× and 8.8×, respectively. A file containing all sequencing reads contributing to an SNV-containing duplex molecule was generated for the HB and CFC sample.

Next, expected tumor fractions and coverages were set for the in silico mixes. Here, expected tumor fractions ranged from 0.1 to x , and the expected coverage was set to 6.5×. A downsampling ratio of reads from the original HB sample was set as

$$\text{HB downsampling ratio} = \frac{\text{final coverage}}{\text{HB coverage}} \times \frac{\text{expected tumor fraction}}{\text{HB tumor fraction}}$$

The CFC downsampling ratio was set as

$$\text{CFC downsampling ratio} \\ = \frac{\text{final coverage}}{\text{CFC coverage}} \times \frac{1 - \text{expected tumor fraction}}{\text{HB tumor fraction}}$$

Therefore, the total number of duplexes to sample from an HB or CFC file was equivalent to

$$\text{Reads to sample} = \text{downsampling ratio} \times \text{number of SNV} \\ \text{—containing duplexes}$$

The exact number of reads per seed was sampled from a normal distribution with a standard deviation of 2.5 and a mean equivalent to ‘reads to sample’. Next, the randomly sampled SNV-containing duplexes (and the sequencing reads that contribute to this duplex) of the HB

and CFC sample were merged and denoised as described above (UMI WGS data processing). Trinucleotide frequencies were tabulated, and melanoma signature scores were measured as described above (Trinucleotide frequency tabulation and signature contribution estimation). Germline variants were removed after in silico mixing using allele frequencies of the original samples.

Duplex, single-strand and UMI-agnostic error rates in mouse PDX plasma samples

Denoising was performed as described above. Variants at a given genomic position, for each correction method, were compared to the frequency of that variant in uncorrected datasets. If the variant occurred two or fewer times in an uncorrected dataset, the variant was considered an error. The error rate was defined as the sum of the errors divided by the total number of base pairs for that correction method. For example, the error rate for duplex datasets corresponded to the number of errors divided by the total number of mapped base pairs from consensus duplex reads.

Statistical analysis

Statistical analysis was performed in R (version 3.6). Box plots were generated using the ggplot2 (version 3.3.5) R package. The bottom and top ends of the boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal line represents the median. The whiskers represent at most 1.5 times the IQR.

Reporting summary

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

Data availability

The raw genomic sequencing data generated are available from the European Genome–Phenome Archive under dataset accession code [EGAD50000001234](https://www.ebi.ac.uk/ena/browser/view/EGAD50000001234). Datasets obtained from the PCAWGC (Supplementary Table 11) are available at <https://www.icgc-argo.org/>. Urothelial cancer tumor/normal alignment files were obtained from Nguyen et al.⁵¹ and were deposited to dbGap under accession number [phs001087.v4.p1](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=phs001087.v4.p1).

Code availability

Code and custom scripts are available at <https://github.com/alexpcheng/WGSDuplex>.

References

- Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 182 (2014).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Novocraft. NovoSort. A multi-threaded sort/merge for BAM files. <https://www.novocraft.com/documentation/novosort-2/>
- Bs, P. & Ar, Q. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
- Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
- DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
- Lai, D., Ha, G. & Shah, S. HMMcopy: copy number prediction with correction for GC and mappability bias for HTS data. Bioconductor version: release (3.15). <https://doi.org/10.18129/B9.bioc.HMMcopy> (2022).
- Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
- Kent, W. J. et al. The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
- Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (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).

Acknowledgements

We thank the participants and their families for contributing plasma and tissue for this study. We also thank H. R. He at Weill Cornell, J. Park and all members of the laboratory of D.A.L., the New York Genome Center computational biology team, especially M. Shah, and the New York Genome Center research sequencing laboratory for thoughtful discussions throughout this work. This work was supported by the Mark Foundation Emerging Leader Award, the Vallee Scholar Award, the Burroughs Wellcome Fund Career Award for Medical Scientists, a National Cancer Institute R01 grant (R01-CA266619-01) and the Melanoma Research Alliance Established Investigator Award (D.A.L.). A.P.C. received support from the American Cancer Society Postdoctoral Fellowship program. Memorial Sloan Kettering Cancer Center investigators are supported by Cancer Center Support Grant P30 CA08748 from the National Institutes of Health/National Cancer Institute. A.J.W. received support from the Conquer Cancer Foundation Young Investigator Award, the Melanoma Research Alliance Young Investigator Award and the NCI K08 Mentored Career Scientist Award (K08 CA263301-03). D.A.L. is a Scholar of the Leukemia and Lymphoma Society. This work was made possible by the MacMillan Family Foundation and the MacMillan Center for the Study of the Non-Coding Cancer Genome at the New York Genome Center. The opinions, results and conclusions reported in this paper are those of the authors and are independent from these funding sources.

Author contributions

D.A.L., A.P.C., A.J.W., G.B. and B.M.F. conceived and designed the project. D.A.L., G.B. and B.M.F. served as lead principal investigators. A.J.W., A.I.A., M.S.M., A. Saxena, M.K.C., D.T.F., L.S., M.S., J.M., A.K., S.T., D.W., J.O., O.E., J.M.M., N.K.A., J.D.W., M.A.P. G.B. and B.M.F. performed participant selection, curated participant data and prepared samples for sequencing. G.I. provided mouse PDX samples. M.M., D.M., A.H., R.F., J.M., Z.S. and L.W. performed library preparation and sequencing. A.P.C., A.A., I.R., A. Sossin, S.R., N.M., W.F.H., R.M.M., D.H., T.L., H.C., S.G., M.C.Z., N.R., Y.W., A.J. and D.L. performed data analysis. A.P.C., C.P. and D.A.L. wrote the manuscript with comments and contributions from all authors.

Competing interests

A.P.C. and D.A.L. have filed a provisional patent regarding certain aspects of this manuscript. D.A.L. and A.J.W. have also filed two additional patent applications regarding work presented in this manuscript. A.P.C. is listed as an inventor on submitted patents pertaining to cell-free DNA (US patent applications 63/237,367, 63/056,249, 63/015,095 and 16/500,929) and receives consulting fees from Eurofins Viracor and has received conference travel support from Ultima Genomics. I.R. and A.J. are employees and shareholders of Ultima Genomics. D.L. is a shareholder of Ultima Genomics. G.I. has received consulting fees from Daiichi Sankyo. J.D.W. is a consultant for Apricity, Ascentage Pharma, Bicara Therapeutics, Bristol Myers Squibb, Daiichi Sankyo, Dragonfly, Imvq, Larkspur, Psioxus, Takeda, Tizona, Trishula Therapeutics, Immunocore – Data Safety board and Scancell; reports grant and research support from Bristol Myers Squibb and Enterome; has equity in Apricity, Arsenal IO/Cell Carta, Ascentage, Imvq, Linneaus, Georgiamune, Takeda, Tizona

Pharmaceuticals and Xenimmune; and is an inventor on the following patents: Xenogeneic DNA Vaccines; Newcastle Disease viruses for Cancer Therapy; Myeloid-derived suppressor cell (MDSC) assay; Prediction of Responsiveness to Treatment with Immunomodulatory Therapeutics and Method of Monitoring Abscopal Effects during such Treatment; Anti-PD1 Antibody; Anti-CTLA4 antibodies; Anti-GITR antibodies and methods of use thereof; CD40 binding molecules and uses thereof. A. Saxena receives research funding from AstraZeneca, has served on Advisory Boards for G1 Therapeutics, Boehringer Ingelheim, Novocure, InxMed, Bristol Myers Squibb and Galvanize Therapeutics, and as a consultant for Galvanize Therapeutics. M.A.P. has received consulting fees from Bristol Myers Squibb, Merck, Novartis, Eisai, Pfizer, Lyvgen and Chugai and has received institutional support from RGenix, Merck Infinity, Bristol Myers Squibb, Merck and Novartis. M.K.C. has received consulting fees from Bristol Myers Squibb, Merck, InCyte, Moderna, ImmunoCore and AstraZeneca and receives institutional support from Bristol Myers Squibb. S.T. is funded by Cancer Research UK (grant reference number A29911); the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC10988), the UK Medical Research Council (FC10988) and the Wellcome Trust (FC10988); the National Institute for Health Research Biomedical Research Centre at the Royal Marsden Hospital and Institute of Cancer Research (grant reference number A109), the Royal Marsden Cancer Charity, The Rosetrees Trust (grant reference number A2204), Ventana Medical Systems (grant reference numbers 10467 and 10530), the National Institute of Health (U01 CA247439) and Melanoma Research Alliance (686061). S.T. has received speaking fees from Roche, AstraZeneca, Novartis and Ipsen. S.T. has the following patents filed: Indel mutations as a therapeutic target and predictive biomarker PCTGB2018/051892 and PCTGB2018/051893. G.B. has sponsored research agreements through her institution with Olink Proteomics, Teiko Bio, InterVenn Biosciences and Palleon Pharmaceuticals; served on advisory boards for Iovance, Merck,

Nektar Therapeutics, Novartis and Ankyra Therapeutics; consulted for Merck, InterVenn Biosciences and Ankyra Therapeutics and holds equity in Ankyra Therapeutics. B.M.F. is on the advisory boards for Astrin Bioscience, Natera, Guardant, Janssen, Gilead, Merck, Immunomedics and QED Therapeutics, is a consultant for QED Therapeutics, Astra Biosciences and BostonGene and obtains patent royalties from Immunomedics and Gilead, honoraria from UroToday and Axiom Healthcare Strategies and research support from Eli Lilly. B.M.F. reports support from the NIH, DoD-CDMRP, Starr Cancer Consortium and the P-1000 Consortium. D.A.L. is on the Scientific Advisory Board of Mission Bio, Pangea, Alethiomics and Veracyte, and has received prior research funding support from Illumina, Ultima Genomics, Celgene, 10x Genomics and Oxford Nanopore Technologies. The remaining authors declare no competing interests.

Additional information

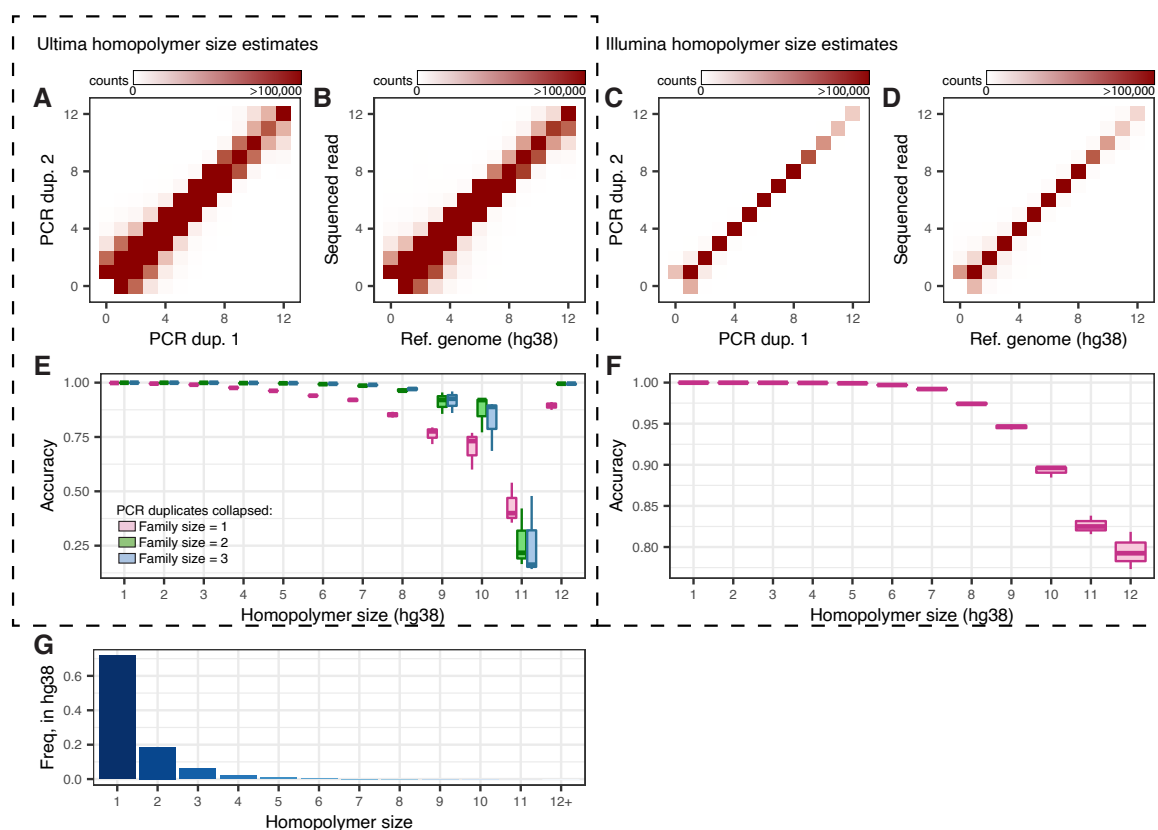
Extended data is available for this paper at <https://doi.org/10.1038/s41592-025-02648-9>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-025-02648-9>.

Correspondence and requests for materials should be addressed to Alexandre Pellán Cheng or Dan A. Landau.

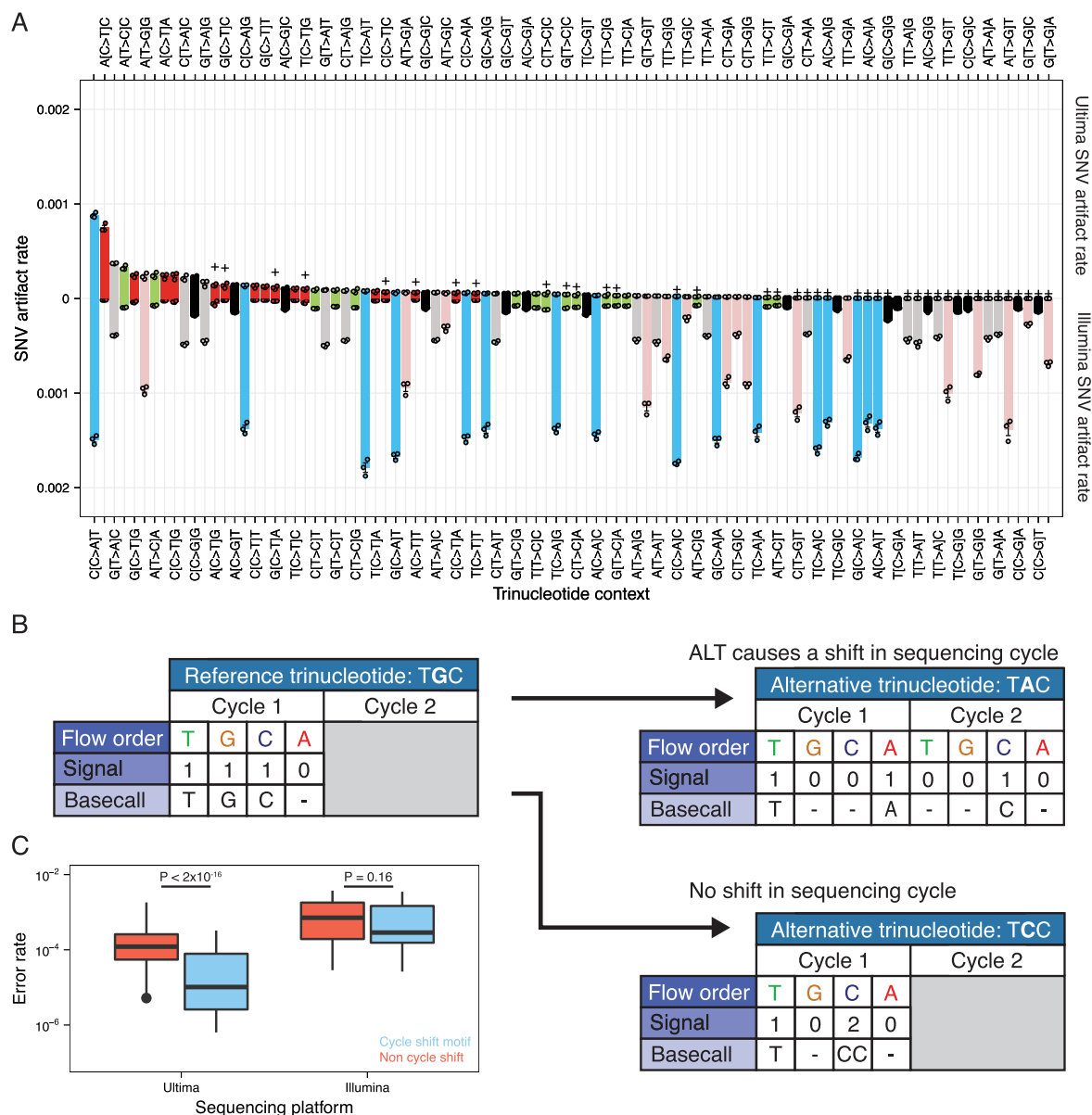
Peer review information *Nature Methods* thanks Andrew Lawson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Lei Tang and Hui Hua, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.



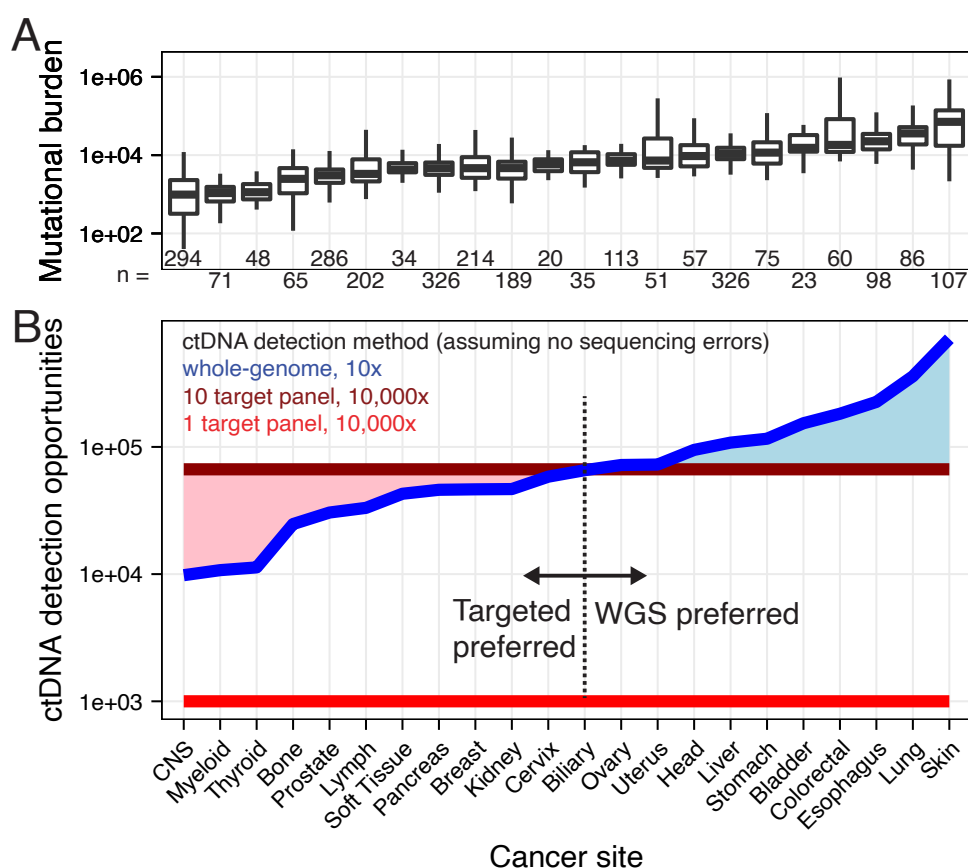
Extended Data Fig. 1 | Ultima and Illumina sequencing datasets of human-mapped reads in mouse PDX datasets ($n = 3$). **A** Homopolymer size estimation of bases between two PCR duplicates (all samples combined) in Ultima datasets. **B** Homopolymer size estimation of bases between a read and the aligned reference (all samples combined) in Ultima datasets. **C** Homopolymer size estimation of bases between two PCR duplicates (all samples combined) in Illumina datasets. **D** Homopolymer size estimation of bases between a read and the aligned reference (all samples combined) in Illumina datasets. **E** Indel calling accuracy by PCR duplicate family sizes in Ultima datasets ($n = 3$ in each boxplot).

F Indel calling accuracy of Illumina sequencing reads (for single family reads, $n = 3$ in each boxplot). **G** Frequency of homopolymer sizes across the human genome. For boxplots in (E) and (F), the lower and upper ends of boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the IQR. Accuracy in (E) and (F) is defined as the number of correct homopolymer assignments in individual sequencing reads divided by the occurrences of that homopolymer size in the human genome in all sequenced reads.



Extended Data Fig. 2 | Flow-based sequencing provides predictable error-robust motifs. **A** Single-nucleotide variant analysis of matched Ultima and Illumina sequencing datasets across 96 trinucleotide contexts. Cycle shift motifs (described in B) are indicated by plus signs. **B** Left: Example sequencing of a TGC trinucleotide in flow space. Given a flow order of T > G > C > A, one full flow cycle of each nucleotide should provide a 1 > 1 > 1 > 0 signal. **Top, right:** Example of how a T[G > A]C alt disrupts the cycles in flow space basecalling. Two sequencing cycles are required to fully resolve a TAC sequencing motif. We refer to these types of motifs as cycle shift motifs. **Bottom, right:** Example of how a T[G > C]C variant does not affect the cycles of flow space basecalling. **C** Error rates in Ultima and Illumina sequencing datasets for trinucleotide variants that

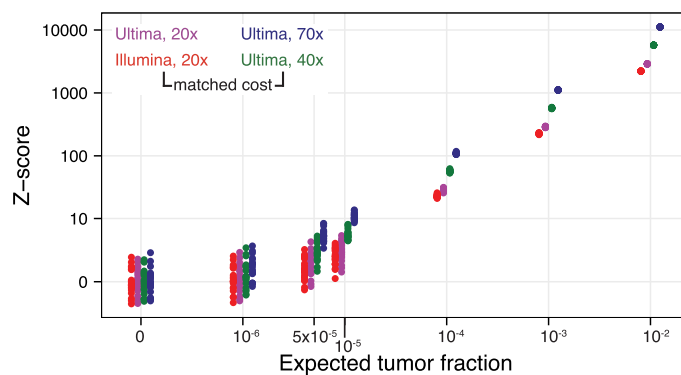
alter the flow space sequencing cycle ($n = 120$ in the cycle shift motif boxplots (blue), corresponding to the 40 trinucleotide variants that are classified as cycle shift motifs across 3 mouse PDX plasma samples. $n = 168$ in the non cycle shift motif boxplots (red), corresponding to the 52 trinucleotide variants that are not classified as cycle shift motifs across 3 mouse PDX plasma samples). P -values were measured using a two-sided Wilcoxon test. Error bars in (A) represent the standard error of the mean. For boxplots in (C), the lower and upper ends of boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the IQR.



Extended Data Fig. 3 | Tradeoffs between deep-targeted sequencing and modest whole-genome sequencing for ctDNA detection. **A** Mutational burden (number of SNVs) of 22 cancer types retrieved from the Pan Cancer Analysis of Whole Genomes consortium. The numbers along the x-axis represent the number of tumors analyzed per cancer type. **B** Median ctDNA detection opportunities using a whole-genome approach with 10x sequencing coverage, a 10-target panel at 10,000x coverage and a 1-target panel at 10,000x coverage.

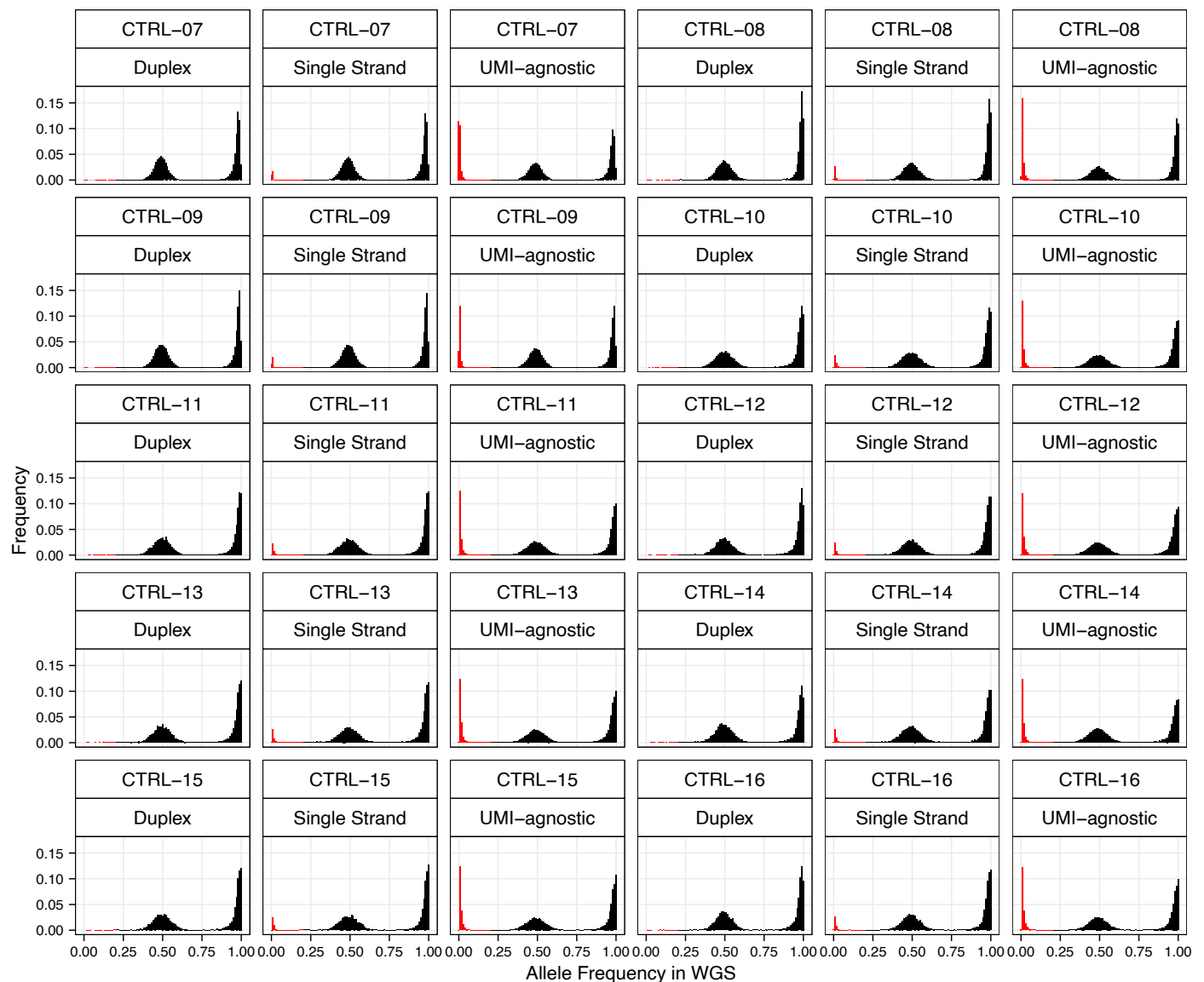
Cancer site

The pink shaded area represents tumor types for which targeting only a few sites may offer benefit over whole-genome sequencing. The blue shaded area represents tumor types for which a whole-genome approach will offer more opportunities to detect ctDNA over targeted panels. The lower and upper ends of the boxplots in (A) represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the IQR.

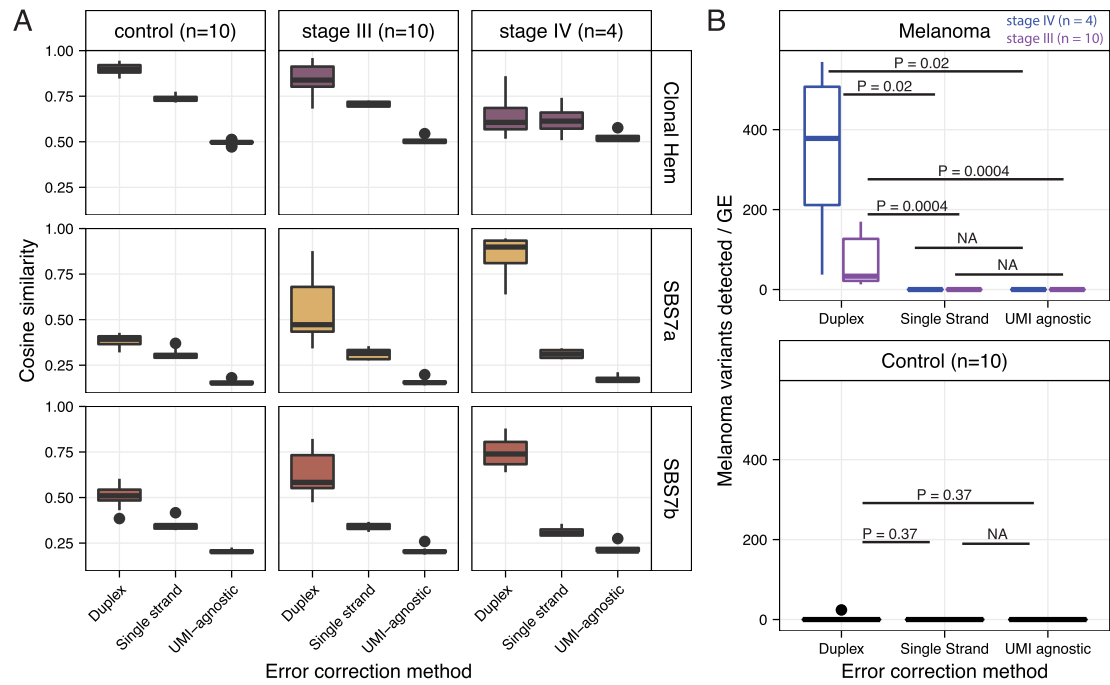


Extended Data Fig. 4 | Circulating tumor DNA cost and coverage analysis between Illumina and Ultima sequencing in a matched sample. Areas under the curve (AUCs) are measured by calculating the area under a receiver operating characteristic curve comparing a given group (for example, Illumina 20x at 10^{-6} expected tumor fraction) to its platform and coverage-matched cancer-free

control (for example, Illumina 20x, expected tumor fraction of 0). All AUCs at expected tumor fractions of 10^{-4} and greater were 1.00. Z-scores of a given sample are calculated against their coverage and platform matched cancer-free control (expected tumor fraction of 0).

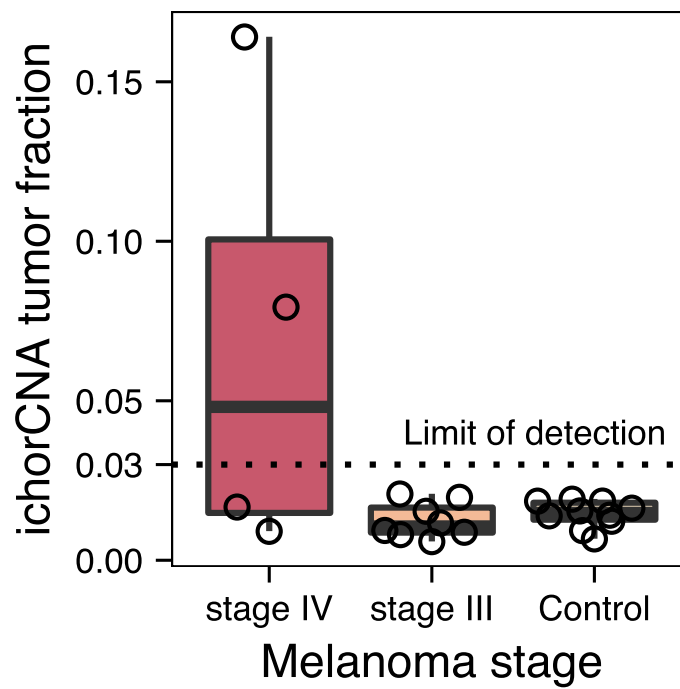


Extended Data Fig. 5 | Variant allele frequencies for variants across denoising approaches. Variant allele frequencies (calculated using unfiltered sequencing reads) in positions where a variant was found using UMI-agnostic denoised reads, Single strand corrected reads and in duplex corrected reads. Allele frequencies of 0.2 and below are colored in red.



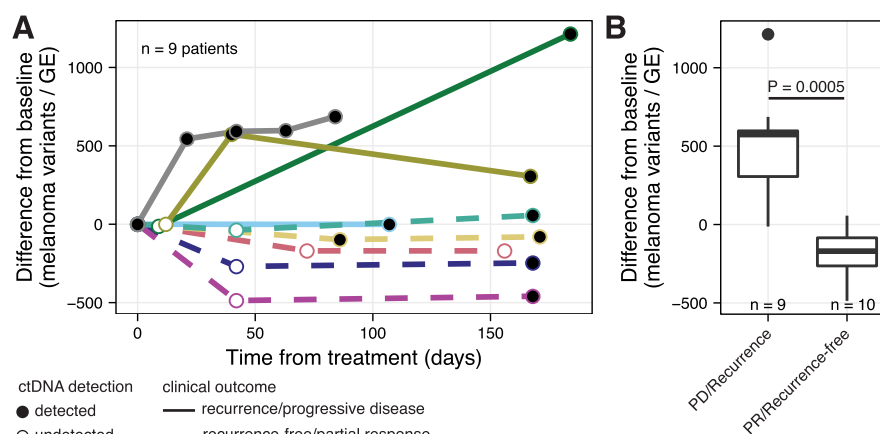
Extended Data Fig. 6 | Comparison of detected UV-derived mutations using duplex, single-strand and UMI-agnostic denoising methods. A Cosine similarities by cancer stage at baseline timepoints (pre-treatment or pre-surgery) for UV and CH-associated signatures. **B** Comparison of duplex, single-strand and UMI-agnostic denoising methods to detect melanoma-associated variants using a single-read variant calling pipeline for pre-treatment plasma samples

from melanoma patients (top) and cancer-free controls (bottom). P-values were measured using a two-sided Wilcoxon test. For all boxplots, the lower and upper ends of boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the IQR.



Extended Data Fig. 7 | Tumor-agnostic copy-number based tumor fraction estimation in stage III and IV melanoma and cancer-free control samples. Samples include cancer-free controls ($n = 10$); stage III melanoma (pre-surgery; $n = 10$) and stage IV melanoma (pre-treatment; $n = 4$). Dotted line at 0.03

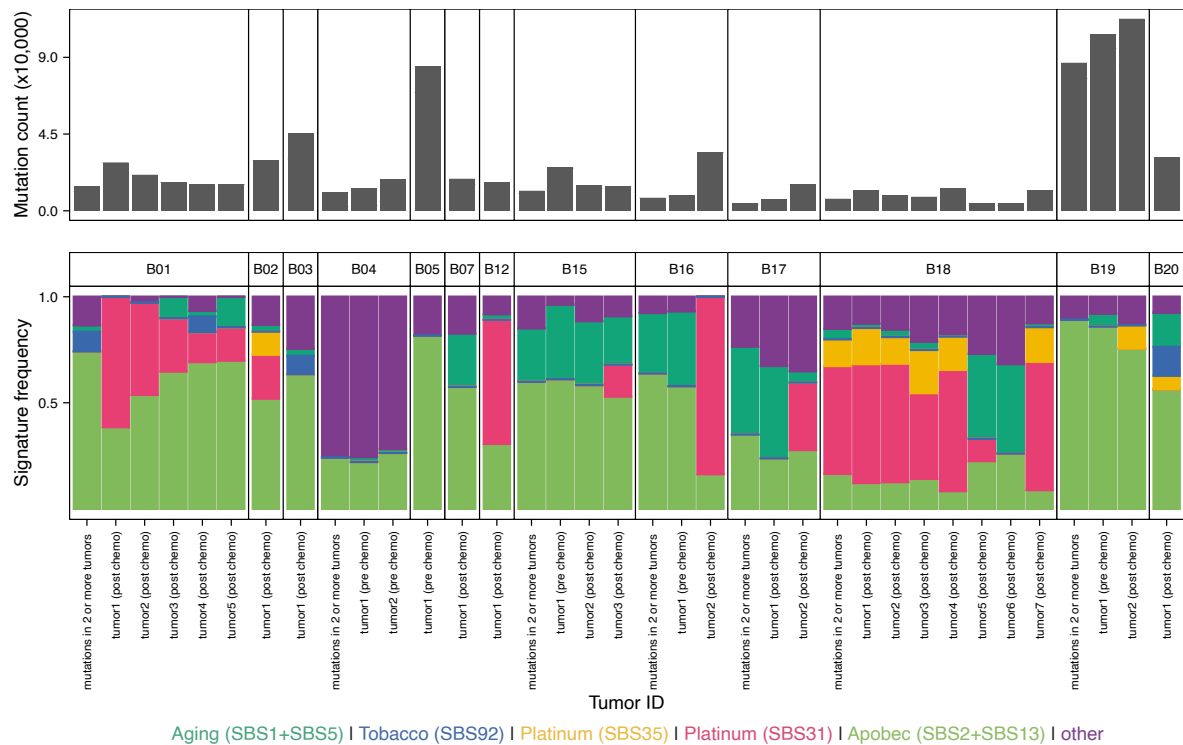
represents the limit of detection of ichorCNA. For boxplots, the lower and upper ends of boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median.



Extended Data Fig. 8 | ctDNA dynamics throughout treatment in melanoma patients. **A** Changes in circulating tumor DNA (increase or decrease) relative to the earliest sampled timepoint. Solid lines represent patients with recurrence or progressive disease, and dashed lines represent patients with either partial response or who are recurrence-free following treatment. Closed and open circles represent samples with and without detected ctDNA, respectively.

B Difference in ctDNA relative to the pre-treatment timepoint stratified by

clinical outcome. One sample did not have a pre-treatment timepoint available (MEL-15; progressive disease) and so a day 9 post-treatment time point was used as baseline. For boxplots in **(B)**, the lower and upper ends of boxes represent the 25th and 75th percentiles of the data, respectively, and the horizontal lines represent the median. The whiskers represent at most 1.5 times the IQR. *P*-values were calculated using a two-sided Wilcoxon test.



Extended Data Fig. 9 | Major signature contributions from urothelial cancer patients' tumors measured through whole-genome sequencing. Top: total mutation counts per sequenced tumor. Bottom: signature contributions. Trinucleotide frequencies were fit to the entire COSMIC database (version v.3.3).

When a patient had two or more tumors (B01, B04, B15, B16, B17, B18, B19), we measured signature contributions of mutations that were present in two or more tumors and thereby likely reflect mutations that arise earlier in tumor evolution.

Reporting Summary

Nature Portfolio 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 Portfolio 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 |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> 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) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> 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	NewYork-Presbyterian/Weill Cornell Medical Center, Memorial Sloan Kettering Cancer Center, Massachusetts General Hospital, or the Royal Marsden NHS Foundation Trust.
Data analysis	Data was analyzed using BWA-MEM (v.0.7.15-r1140), GATK (v4.0), Samtools (v1.19), MuTect (v1.1.7), LoFreq (v2.1.3a), Ichor-CNA (v0.3.2), skewer (v.0.2.2), mosdepth (v.0.2.9), cutadapt (v. 2.10), hmmcopy (v. 0.99), bedtools (v. 2.29), fgbio (v.2), Novosort MarkDuplicates (v1.03.01) Custom python scripts (python version 3.6) and R scripts (version 3.6). Code and custom scripts are available at https://github.com/alexpcheng/WGSDuplex .

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 Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The raw genomic sequencing data generated is available from the European Genome-Phenome Archive (EGA) under dataset accession code EGAD50000001234. . Datasets obtained from the Pan Cancer Analysis of Whole Genomes Consortium (Supplementary Table 11) are available at <https://www.icgc-argo.org/>. Urothelial cancer tumor/normal alignment files were obtained from Nguyen et al.⁵¹ and were deposited to dbGap under accession number phs001087.v4.p1.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Patient-level information was collected on biological sex (see Supplementary Table 1). Sex and gender were not considered as covariates and this study was not designed to capture sex or gender-based effects. Sex of participants was determined by self-report.

Population characteristics

WGS without duplex and WGS with duplex control cohorts (n = 5 and n = 3, respectively) had median ages of 74 and 75, and were 40% and 33% female, respectively. LUAD cohort (n = 3) had a median age of 79 and were 100% female. Melanoma cohorts (WGS without duplex and WGS with duplex) had median ages of 64.5 and 56 and were 25% and 20% female, respectively).

Recruitment

This manuscript is a methods-focused study and samples were selected retrospectively to test the detection power of the technology (example, cancer-free controls, stage IV disease, stage III disease, melanoma, lung cancer)

Ethics oversight

Blood and tissue samples were obtained from patients after obtaining informed consent and following protocols approved by institutional review boards and in accordance with the Declaration of Helsinki protocol. Samples were obtained from either NewYork-Presbyterian/Weill Cornell Medical Center (Institutional Review Board (IRB) numbers 0201005295 (Tumor Biobanking), 1008011210 (GU Tumor Biobanking), 1011011386 (Urothelial Cancer Sequencing), 100701157 (Genomic and Transcriptomic Profiling), 1305013903 (Precision Medicine), 1708018519 (Cardiac Surgery Biobank), 2014-0024 (approved by the Institutional Animal Care and Use Committee at Weill Cornell Medicine), 1610017682 (Circulating tumor DNA for early detection and management of Non Small Cell Lung Cancer), Memorial Sloan Kettering Cancer Center (IRB number 12-245 (Genomic Profiling in Cancer Patients), Massachusetts General Hospital (IRB number 11-181 (Collection of Tissue and Blood Specimens and Clinical Data from Patients with Melanoma and Other Cutaneous Malignancies)), or the Royal Marsden NHS Foundation Trust in the United Kingdom (Supplementary Table 1). Tumor, normal and plasma samples from the Royal Marsden NHS Foundation Trust were obtained under an ethically approved protocol (Melanoma TRACERx, Research Ethics Committee Reference 11/LO/0003). Cancer diagnosis was established according to World Health Organization criteria and confirmed in all cases by an independent pathology review. Patients did not receive any compensation.

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

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

Life sciences study design

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

Sample size

Statistical methods were not used to determine sample size. For error-rate analysis, we obtained billions of sequenced bases per sample, which allowed us to robustly quantify error rates between correction methods (UMI-agnostic, single-stranded UMI correction, duplex correction). For cancer detection, the number of samples was chosen based on the availability of datasets.

Data exclusions

Variants were filtered based on quality control metrics according to sample type (UMI-agnostic denoising, single-stranded UMI-based denoising and duplex denoising). These quality control metrics are described in the Methods section.

Replication

In silico admixtures with 10-50 replicates were analyzed to estimate the error rates and the reproducibility of whole genome sequencing's accuracy of tumor fraction estimate in a tumor-informed model, and whole genome duplex sequencing's accuracy of tumor fraction detection

in a tumor-agnostic model. For whole genome duplex sequencing, tumor-specific signatures present in circulating DNA were assessed against 10 control samples. Attempts at replicating whole genome duplex sequencing experiments were successful.

Randomization Randomization was not included in this retrospective, non-intervention study

Blinding No blinding was performed in this non-intervention study

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
<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 type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<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

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals Patient-derived xenografts (PDX) were established using fresh pathological tissue fragments from patients with lung cancer of diffuse B cell lymphoma, implanted subcutaneously into six- to eight-week-old anesthetized NGS female mice.

Wild animals No wild animals were used in the study.

Reporting on sex N/A

Field-collected samples No field collected samples were used in the study.

Ethics oversight Mouse PDX studies were reviewed and approved by Institutional Animal Care and Use Committee (IACUC, institutional review board number 2014-0024) at Weill Cornell Medicine.

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