Ultra-Sensitive Mutation Detection and Genome-Wide DNA Copy Number Reconstruction by Error-Corrected Circulating Tumor DNA Sequencing

Sonia Mansukhani,¹ Louis J. Barber,¹ Dimitrios Klefogiannis,¹ Sing Yu Moorcraft,² Michael Davidson,² Andrew Woolston,¹ Paula Zuzanna Proszek,³ Beatrice Griffiths,¹ Kerry Fenwick,² Bram Herman,⁵ Nik Matthews,³ Ben O’Leary,⁴ Sanna Hulkk,² David Gonzalez De Castro,⁷ Anisha Patel,⁷ Andrew Wotherspoon,⁵ Aleruchi Okachi,² Isma Rana,² Ruwaida Begum,² Matthew N. Davies,¹,⁴ Thomas Powles,¹ Katharina von Loga,¹ Michael Hubank,³ Nick Turner,⁶,¹¹ David Watkins,⁷ Ian Chau,² David Cunningham,² Stefano Lise,¹ Naureen Starling,² and Marco Gerlinger¹,²

BACKGROUND: Circulating free DNA sequencing (cfDNA-Seq) can portray cancer genome landscapes, but highly sensitive and specific technologies are necessary to accurately detect mutations with often low variant frequencies.

METHODS: We developed a customizable hybrid-capture cfDNA-Seq technology using off-the-shelf molecular barcodes and a novel duplex DNA molecule identification tool for enhanced error correction.

RESULTS: Modeling based on cfDNA yields from 58 patients showed that this technology, requiring 25 ng of cfDNA, could be applied to >95% of patients with metastatic colorectal cancer (mCRC). cfDNA-Seq of a 32-gene, 163.3-kbp target region detected 100% of single-nucleotide variants, with 0.15% variant frequency in spike-in experiments. Molecular barcode error correction reduced false-positive mutation calls by 97.5%. In 28 consecutively analyzed patients with mCRC, 80 out of 91 mutations previously detected by tumor tissue sequencing were called in the cfDNA. Call rates were similar for point mutations and indels. cfDNA-Seq identified typical mCRC driver mutations in patients in whom biopsy sequencing had failed or did not include key mCRC driver genes. Mutations only called in cfDNA but undetectable in matched biopsies included a subclonal resistance driver mutation to anti-EGFR antibodies in KRAS, parallel evolution of multiple PIK3CA mutations in 2 cases, and TP53 mutations originating from clonal hematopoiesis. Furthermore, cfDNA-Seq off-target read analysis allowed simultaneous genome-wide copy number profile reconstruction in 20 of 28 cases. Copy number profiles were validated by low-coverage whole-genome sequencing.

CONCLUSIONS: This error-corrected, ultradeep cfDNA-Seq technology with a customizable target region and publicly available bioinformatics tools enables broad insights into cancer genomes and evolution.

CLINICALTRIALS.GOV IDENTIFIER: NCT02112357

© 2018 American Association for Clinical Chemistry

Many tumors release cell-free DNA (cfDNA)¹² into the circulation, allowing the analysis of cancer genetic aberrations from blood samples (7–6). Such “liquid biopsies” can inform tailored therapies (7) or predict recurrences after surgery (8, 9). cfDNA analysis also permits subclonal mutation detection that is often missed by biopsies because of spatial intratumor heterogeneity (10, 11). Genetic techniques with high analytical sensitivity and low false-positive error rates are crucial for accurate circulating free DNA sequencing (cfDNA-Seq) because of low

¹ Center for Evolution and Cancer, Division of Molecular Pathology, The Institute of Cancer Research, London, UK; ² Gastrointestinal Cancer Unit, The Royal Marsden NHS Foundation Trust, London and Sutton, UK; ³ Centre for Molecular Pathology, The Royal Marsden NHS Foundation Trust, Sutton, UK; ⁴ Tumour Profiling Unit, The Institute of Cancer Research, London, UK; ⁵ Diagnostics and Genomics Group, Agilent Technologies Inc., Santa Clara, CA; ⁶ Breast Cancer Now Research Centre, The Institute of Cancer Research, London, UK; ⁷ Centre for Cancer Research and Cell Biology, Bellatine University, UK; ⁸ Department for Radiology, The Royal Marsden NHS Foundation Trust, London and Sutton, UK; ⁹ Department of Histopathology, The Royal Marsden NHS Foundation Trust, London and Sutton, UK; ¹⁰ Barts Cancer Institute, Queen Mary University of London, London, UK; ¹¹ Breast Cancer Unit, The Royal Marsden NHS Foundation Trust, London and Sutton, UK.

¹ Current address: Achilles Therapeutics; Francis Crick Institute, London, UK.

© 2018 American Association for Clinical Chemistry

1 S. Mansukhani and L. J. Barber contributed equally to this work.
2 Current address: Achilles Therapeutics; Francis Crick Institute, London, UK. Received March 21, 2018; accepted July 17, 2018. Previously published online at DOI: 10.1373/clinchem.2018.289629

© 2018 American Association for Clinical Chemistry

1 Nonstandard abbreviations: cfDNA-Seq, circulating free DNA sequencing; cfDNA, cell-free DNA; ddPCR, digital droplet PCR; NGS, next-generation sequencing; VAF, variant allele frequency; MBC, molecular barcode; mCRC, metastatic colorectal cancer; FORMAF, Feasibility Of Molecular characterization Approach to Treatment; HD, healthy donor; SNP, single-nucleotide polymorphism; IGV, Integrative Genomics Viewer; VClOT, variant call quality threshold; CNA, copy number aberration.
Methods

PATIENTS AND SAMPLES
Plasma samples and clinical data were available from the Feasibility of Molecular Characterization Approach to Treatment (FOrMAT) trial ((16), Chief Investigator: N Starling ClinicalTrials.gov NCT02112357). Healthy donor (HD) cfDNA was obtained through the Tissue Collection Framework to Improve Outcomes in Solid Tumors (Chief Investigator: T Powles). Both trials were approved by UK ethics committees, and all patients provided written informed consent. Details of clinical trials, patients, samples, sample processing, and experimental techniques are provided in the Methods in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol64/issue11.

cfDNA SEQUENCING
SureSelectXT-HS (Agilent) was used to prepare sequencing libraries with our optimized protocol (see Methods file in the online Data Supplement) and a custom-designed SureSelect bait library (see Table 1 in the online Data Supplement). Sequencing libraries were clustered with the cBot and sequenced with paired-end 75 reads on an Illumina HiSeq2500 in rapid mode. Sequencing fastq files have been deposited into the NCBI Sequence Read Archive (SRA submission code SUB3510375).

SureCall software (version 4.0.1.45, Agilent) was used to trim and align fastq reads to the hg19 reference genome with default parameters and for MBC deduplication, permitting one base mismatch within each MBC. Consensus families comprising single reads were removed; on-target depths were assessed and variants were called with SureCall SNPPEL.

To identify variants supported by duplexes, we developed the freely available DuplexCaller bioinformatics tool (17).

All variant positions identified in patient cfDNA were assessed in 6 HD samples with bam-readcount (18). Most called variants were absent in HD samples (see Table 2 in the online Data Supplement) but mutations with VAF less than double that of an identical variant in HD were removed as false positives.

BAM files resulting from MBC deduplication before removal of single-read consensus families were used to generate genome-wide DNA copy number profiles with CNVkit (19), with antitarget average size set to 30 kb. HD samples were used as the normal reference pooled dataset.

LOW-COVERAGE WHOLE-GENOME SEQUENCING
Genomic libraries were constructed from 10 ng of cfDNA with the NEBNext Ultra II kit and sequenced with 100 bp single-end reads on HiSeq2500 in rapid mode (0.42× median coverage). Data were aligned (hg19 reference genome) with Bowtie (v. 0.12.9) and processed as described (20). Logratios were normalized against a sex-matched pooled dataset from HD cfDNA (9 male, 8 female) before segmentation and median centering.

ddPCR
ddPCR was performed in case 8 (BRAFV600E) and for validation of discordant variants between cfDNA and tumor tissue. Four of 11 such cases had sufficient remaining cfDNA to validate subclonal variants (see Methods file in the online Data Supplement).

cfDNA SEQUENCING WITH A COMMERCIAL KIT
25 ng, 17 ng, and 25 ng of cfDNA (cases 3, 15, and 23, respectively) were processed with the Roche Avenio Ex-
mented kit as per the manufacturer’s protocol. Libraries were sequenced with 151-bp paired-end reads on Illumina NextSeq500 to 2689–6420× depth after deduplication. Data were analyzed with the Roche Avenio circulating tumor DNA Analysis Software v1.0.0 with default parameters.

Results

cfDNA SEQUENCING OPTIMIZATION

Modeling based on cfDNA yields from 58 patients with mCRC showed that 25 ng of cfDNA could be extracted from 20–30 mL blood from >95% of cases (see Fig. 2C in the online Data Supplement). Twenty-five nanogram was therefore chosen as our standard cfDNA input quantity. We designed a solution hybrid-capture panel targeting 32 genes, including all major CRC driver genes (163.3 kb; see Table 1 in the online Data Supplement), and used Agilent SureSelectXT-HS kit, which tags each DNA strand with a random 10-base MBC, for sequencing library preparation. The SureSelectXT-HS protocol was optimized to perform reliably with 25 ng of cfDNA input (see Methods file in the online Data Supplement). The fraction of on-target reads is usually low when using small, targeted sequencing panels and low input DNA, so we first assessed how the on-target fraction could be optimized by varying the stringency of the postcapture wash. Two library preparations were started in parallel from each of 4 cfDNA samples, according to the 1.5 h fast-hybridization protocol. Then, postcapture washes were performed at 65 °C in one library and at 70 °C in the other. Sequencing generated similar read numbers (65 °C, 92820887; 70 °C, 102582694 median reads/sample) and the on-target fraction significantly (P = 0.001) increased from 30%–35% to 71%–74% with the 70 °C protocol (Fig. 1A). Hence, the more stringent conditions were chosen for our standard protocol. Target exon coverage was even with this solution hybrid-capture technique and was not subject to the gaps commonly seen with commercial amplicon sequencing designs (Fig. 3 in the online Data Supplement). This would be particularly advantageous for the analysis of tumor-suppressor genes for which driver mutations often spread across large parts of the gene.

We next used MBCs to deduplicate sequencing data and perform error correction. SureCall creates families of reads with matched MBCs that also align to the same genomic position and then identifies the most likely consensus sequence for each family (Fig. 1B). This reduces random errors arising during PCR and sequencing, as these are not common to all reads of a family. Consensus families contained a median of 8–15 supporting reads in samples sequenced with the optimized protocol (see Fig. 4 in the online Data Supplement), which was within the optimal range for barcode error correction (21). After MBC deduplication, the median on-target depth with the 70 °C protocol was 1782×. This was theoretically sufficient to achieve a detection limit as low as 1 mutated DNA fragment in 1782 molecules (0.056%). However, the analytical sensitivity for de novo mutation detection is lower in practice because more than one read is required to support robust bioinformatics calling. Thus, we designed a mixing experiment to test the ability to detect and bioinformatically call mutations with low VAFs.

ASSAY SENSITIVITY AND SPECIFICITY

cfDNA from 2 donors that differed in 16 homozygous single-nucleotide polymorphisms (SNPs) within the targeted region were used to prepare a dilution series with 0.15%, 0.075%, and 0.0375% cfDNA from donor A added into cfDNA from donor B. Sequencing a median of 74030118 reads/sample generated a median on-target depth of 21651× before deduplication. Data from each sample were then processed in 2 ways: first, we used MBCs for deduplication and calling of consensus sequences; second, we performed standard deduplication using only the genomic position of each read pair. The median on-target depth was higher after MBC deduplication (MBCs 2420× vs 1587× with standard deduplication; Fig. 1C). This was anticipated because different MBCs tag distinct DNA fragments that would otherwise be counted as duplicates. For example, the forward and reverse strands of each original “duplex” dsDNA molecule were separately tagged by MBC and so were retained as independent consensus families (Fig. 1D). Standard deduplication cannot distinguish these reads from PCR duplicates.

We first investigated whether the added SNPs could be reidentified in the MBC-deduplicated BAM files with the Integrative Genomics Viewer (IGV) (22) and tried to understand patterns associated with true-positive variants. All 16 SNPs were detected in the 0.15% mix, 14 of 16 at 0.075% and 11 of 16 at 0.0375% mixing ratios (Fig. 1E). Thus, our ultradeep cfDNA-Seq assay allowed robust detection of variants at 0.15% and retained a high detection capability at 0.075%. We then assessed if MBC error correction improved the bioinformatics calling accuracy of ultralow frequency variants, which is more challenging than reidentification of known variants. While interrogating sequencing data manually in IGV, we had observed that all true variants were at least supported by 2 consensus families mapping to the same genomic position but differing in whether the variant was seen in read 1 or read 2 in paired-end sequencing (Fig. 1D). These reads were highly likely to represent the forward and reverse strand of the double-stranded input cfDNA molecule, as observed previously (15). Based on this observation, we developed the DuplexCaller bioinformatics tool that identified variants supported by du-
Fig. 1. Optimization of cfDNA-Seq with incorporation of molecular barcode technology for enhanced error correction.

Percentage of reads on-target before deduplication in samples prepared with 65 °C vs 70 °C postcapture washes (A). Graphic depicting the principles of MBC error correction (B). Reads with the same MBC that map to the identical genomic location are grouped into a consensus family. If a variant (pink) occurs in all reads, then the consensus read sequence will be variant for that base (top). However, if a variant (green) is only detected in a small fraction of the reads in the family, it will be disregarded, and the consensus read sequence will be wild type (bottom). cfDNA mixing experiment: 25 ng of mixtures of donor A added into donor B at 0.15%, 0.075%, and 0.0375% (C). Illustration of duplex read pair detection (D). A double-stranded (ds) cfDNA fragment (black) containing a variant (green) is depicted, ligated to Y-shaped, MBC-tagged adapters (gray). Expected and observed VAFs and genomic positions for the 16 SNPs in the cfDNA mixing experiment (E). Effect of MBC error correction on true-positive and false-positive calls (F). The top panels show the number of true-positive variants (expected SNPs) that were bioinformatically called in the mixing experiment with standard deduplication (left) and MBC deduplication (right) with different VCQTs. The lower panel shows the number of likely false-positive variant calls (not observed in the deep sequencing of either cfDNA sample used in the mixture) for standard deduplication (left) and MBC deduplication (right).
plex reads (see Methods file in the online Data Supplement) and added the requirement for such a “duplex-configuration” to be present to accept a mutation as genuine. The presence of a variant in at least one additional family with a different alignment position was also added to the postcall filters to ensure high specificity. Thus, a variant had to be present in ≥3 consensus DNA families to be accepted as a mutation call in the MBC-deduplicated data. For a meaningful comparison, mutations in the standard deduplicated data were also required to be present in ≥3 reads.

We then compared SureCall calls for the mixing experiment on standard- vs MBC-deduplicated data and quantified how many of the homozygous SNPs from sample A that were present at 0.15% in the cfDNA mixture were called. Although samples A and B differed at 16 homozygous SNP positions, only 9 variant SNPs in added sample A could be assessed for capability to call at low frequency against the reference genome. The other 7 SNPs were reference wild type in added sample A and so could not be called. Mutation calling after standard deduplication with low-stringency caller settings [variant call quality threshold (VCQT) = 40] detected 5 of 9 homozygous SNPs (Fig. 1F) but also generated 156 additional calls. These additional variants were likely false positives, since they had not been identified by deep sequencing of the individual cfDNA samples used in the mixing experiment. Stepwise increase of the VCQT reduced false positives, but this reduction was accompanied by a loss of analytical sensitivity. When the same data were called with MBCs and a low-stringency VCQT = 40 (Fig. 1F), 4 of the added SNPs were called with only 2 likely false-positive variants. We assessed why calling with MBC error correction failed to identify the 5 other SNPs. Each of these had VAFs <0.1% when visualized in IGV (22), which was below the minimum VAF of 0.1% that can be called by SureCall. We also assessed the number of false-positive calls in standard deduplicated data at the maximum VCQT that identified the 4 true-positive variants detected with MBCs: 81 likely false positives were called, compared to just 2 with MBCs. Hence, at the same analytical sensitivity, deduplication using the MBCs dramatically decreased false positives by 97.5%. Mutation calling in 6 HD samples subjected to cfDNA-MBCs dramatically decreased false positives, but this reduction was accompanied by a loss of analytical specificity of this MBC technology.

CONCORDANCE OF cfDNA AND TUMOR SEQUENCING IN mCRC PATIENTS

cfDNA from 28 patients with mCRC were consecutively analyzed. Seven were sequenced with the 65 °C protocol and 21 with the 70 °C protocol. The median sequencing depth was higher with 70 °C (2087×) than with 65 °C (1205×) (Fig. 2A). We then analyzed the concordance and discordance of mutation calls within the target regions common to the tumor biopsy sequencing assay and our cfDNA-Seq panel. Biopsies of 23 cases had been sequenced with the ForMAT NGS panel (see Table 4 in the online Data Supplement), and 4 biopsies had been subjected to routine clinical amplicon sequencing of 5 genes (BRAF, KRAS, NRAS, PIK3CA, and TP53). One case had failed tissue sequencing.

Eighty-eight percent (80 of 91) of all mutations that had been found by tumor sequencing were called in the cfDNA (Fig. 2A). All 11 mutations not called in cfDNA were from only 3 cases. Inspection of the sequencing data on IGV revealed that 5 out of 11 mutations were present in cfDNA at VAFs below the SureCall detection limit (Fig. 2B). Sufficient cfDNA remained from case 8 for orthogonal analysis by ddPCR. Using manufacturer-validated ddPCR probes for the BRAFV600E mutation, we identified 2830 wild-type DNA fragments but no mutated fragments (data not shown). This confirmed that the absence of sufficiently abundant tumor-derived cfDNA molecules, rather than technical failure, explained the inability to detect mutations.

We next assessed mutations called by cfDNA-Seq in genes that had not been sequenced in corresponding tumor tissue. APC mutations were detected in each of 4 cases whose tumors had only been analyzed with the 5-gene amplicon panel (Fig. 2A). Furthermore, one mutation was found in each of FBXW7, CTNNB1, TCF7L2, ATM, and SMAD4. We also detected mutations in APC, TP53, and KRAS in case 28 that had failed prior tumor tissue sequencing attempts. In total, 11 of these 13 mutations (85%) encoded protein changes previously reported in the COSMIC cancer mutation database (23), and all variants in the tumor-suppressor genes APC and FBXW7 were truncating and hence likely driver mutations. This demonstrated that our assay could detect biologically and clinically important cancer mutations directly from cfDNA.

We then investigated mutations that had been called in cfDNA but were absent when the same gene had been analyzed in tumor tissue: 7 in TP53, 7 in ATM, 3 in PIK3CA, 2 in SMAD4, and one each in KRAS, FBXW7, and TCF7L2. All 4 mutations called in the oncogenes KRAS and PIK3CA were canonical cancer driver mutations. Eight of 18 mutations (44%) located in tumor-suppressor genes were nonsense mutations or encoded for amino acid changes found recurrently in cancer (24), suggesting that these were also driver mutations. Together, 54.5% (12 of 22) of variants detected only in cfDNA were likely cancer driver mutations. The VAFs of mutations that were only detected in cfDNA but not in tumor tissue were a mean 105-fold lower than the VAF of the most abundant mutation detected in the same cfDNA sample (see Fig. 1 in the online Data Supple-
Fig. 2. Analysis of mutation calls in cfDNA samples from 28 patients with metastatic colorectal cancer.
Concordance of mutations identified by cfDNA-Seq and by sequencing of tumor material (A). Mutations identified in both cfDNA-Seq and tumor sequencing are green. Novel variants called by cfDNA-Seq and not by tumor sequencing are blue. Variants not detected by cfDNA-Seq that were detected in tumor sequencing are orange. Pink indicates clonal hematopoiesis. Red outlines indicate mutations reported as tumorigenic in COSMIC. Variants in gray have been identified in the cfDNA of patients that either had been sequenced with the limited 5-gene amplicon panel or failed FORMAT sequencing. Percentages indicate VAF in cfDNA. Read depth and number of consensus family reads supporting each of the 11 variants in cases 7, 8, and 21 that had not been called in cfDNA but had previously been detected in tumor tissue. Median VAF, 0.066% (B). ddPCR validation of the KRAS c.183A>C mutation that results in the amino acid change Q61H in case 10 (C). Green dots, droplets with wild-type DNA; blue dots (outlined by the red quadrant), droplets with mutant DNA; black dots, droplets that have no incorporated DNA. ddPCR validation of 6 subclonal mutations called in cfDNA but not in tumor tissue (D).
ment); these variants likely originated from small cancer subclones. However, 2 TP53 mutations present in cfDNA but not in matched tumor tissue (cases 9 and 13) were also detected with similar VAF in DNA from blood cells (see Table 5 in the online Data Supplement). These TP53 mutations hence originated from a clonal expansion of blood cells (9), termed clonal hematopoiesis (24, 25).

An activating mutation in KRAS (Q61H) was detected with a VAF of 0.37% in cfDNA but not in the matched tumor (case 10). This was the only patient who had received treatment with the anti-EGFR antibody cetuximab before blood collection, and the KRAS mutation was likely a driver of acquired resistance that evolved during therapy (26). ddPCR testing of cfDNA provided orthogonal validation (Fig. 2C), showing that our technology is suitable for the detection of subclonal resistance driver mutations. Suspected driver mutations in PIK3CA were frequently discordant with 3 out of 7 mutations only detectable in cfDNA (E545K, H1046R, R1023*). Two cases (17, 26) harbored parallel evolution events, as further activating PIK3CA mutations were present in the tumors and the cfDNA. These results are consistent with studies showing that intratumor heterogeneity of PIK3CA mutations is common in mCRCs, whereas heterogeneity is rare for mutations in APC and, in tumors not previously treated with anti-EGFR antibodies, for KRAS, NRAS, and BRAF mutations (27).

Mutations in ATM tumor-suppressor gene were called in 8 of 28 cfDNA samples. Sequencing of matched tumor showed wild-type sequence in 7 of these, and one tumor had only been sequenced with the 5-gene panel. All ATM mutations had low VAFs (median, 0.17%) and only 2 of 8 encode protein changes previously cataloged in cancer (24), making it difficult to interpret their functional relevance. No ATM mutations were called in 6 HDs, indicating that the mutation calls in cfDNA from mCRC patients are unlikely to be the result of a high false-positive call rate in this gene.

Next, we used ddPCR to validate further subclonal mutations called in cfDNA but not in tumor tissue. All subclonal variants with VAF <2% from samples in which sufficient cfDNA material was available and in which a custom ddPCR assay could be designed were assessed (see Methods file in the online Data Supplement). ddPCR validated all 6 tested mutations, and VAFs were similar to those found by our error-corrected cfDNA technology (Fig. 2D, see Table 6 in the online Data Supplement).

Additionally, we resequenced 3 cfDNA samples containing low VAF (<2%) mutations (cases 3, 15, 23) with the commercially available Avenio circulating tumor DNA kit. Nine of 10 point mutations in genes targeted by both panels were concordant (see Table 7 in the online Data Supplement). The low-frequency TP53 R175H variant in case 3 was not called by Avenio software but was seen to be present on manual review of the BAM file. Three indels in APC (cases 3 and 23) were not called by Avenio analysis. This comparison further confirmed the reliable performance of our customizable cfDNA assay.

**GENOME-WIDE DNA COPY NUMBER ABERRATION ANALYSIS**

We finally assessed if we could maximize the information gain from a targeted cfDNA assay through simultaneous reconstruction of genome-wide copy number aberration (CNA) profiles. Applying the CNVkit package (19) that uses off-target reads to infer copy number changes, we generated genome-wide CNA profiles for 20 of 28 cases (71%) (Fig. 3, A and B). Chromosome arm losses (Chr17p and 18q) and gains (Chr1q, 7, 8q, 13, and 20), which are typical for mCRC, were observed (28). All 8 samples with a flat CNA profile had very low maximum VAFs \( \leq 5.6\% \). A high-level targetable amplification involving the ERBB2 oncogene was detected despite a low tumor-derived cfDNA fraction (8.6% VAF) in case 11 (Fig. 3C). This amplification had also been detected in the matched tumor; this detection validated the ability of our cfDNA-Seq technology to profile CNAs. No other amplifications had been detected in tumor biopsies with the FOrMAT NGS panel. Low-coverage whole-genome sequencing is an established approach for genome-wide copy number profiling, and we applied this to 18 samples with sufficient cfDNA. This independent validation showed a median-weighted Spearman correlation of 0.886 with the profiles generated from cfDNA-Seq using CNVkit (see Fig. 5 in the online Data Supplement).

**Discussion**

Our ultradepend and error-corrected cfDNA-Seq protocol, which uses off-the-shelf MBCs in combination with a custom-designed solution hybrid-capture panel, detected 100% of the known variants with VAFs of 0.15% in a mixing experiment. The use of MBC error correction and the requirement for variants to be supported by a duplex pair of consensus families reduced false-positive mutation calls by 97.5% while maintaining true positives. We developed the DuplexCaller bioinformatics tool, which can be run directly after MBC deduplication to facilitate mutation calling; all bioinformatics tools for the analysis of data generated with this technology are hence freely available. Our approach did not rely on background error correction models that are constructed from large numbers of HD samples and are therefore impractical for applications requiring frequently changing custom gene panels, including clinical assay development.

Importantly, the 1.5-h fast-hybridization step (standard protocol, 16h) used in our assay dramatically re-
Fig. 3. Reconstruction of genome-wide copy number aberration profiles from cfDNA-Seq assay.

Genome-wide CNAs can be detected from targeted cfDNA-Seq, even where tumor content is low (A). Representative log copy ratio plots for 5 cases (green number) in our cohort with tumor content ranging from 53.5% to 8.6% (red number indicates max VAF) are shown. Genome-wide heat map of segmented copy number raw log ratio data after amplitude normalization (B). Gains are red, and losses are blue. Profiles are ordered (left to right) from highest to lowest tumor content (based on maximum VAF) for all 20 cases that had a visible CNA profile. Focused log copy ratio plot of chromosome 17 for case 11, which had a high-level amplification of ERBB2 (C).
duces library preparation time, which is advantageous when fast turnaround is critical. Increasing the wash temperature after capture dramatically reduced off-target reads. The higher temperature likely relaxes the target-bait bond in hybridized molecules with a higher number of mismatches, reducing the nonspecific carryover of DNA fragments into the library. cfDNA-Seq of 28 mCRC patients demonstrated that 88% of mutations detected by clinical-grade tumor tissue sequencing were also called in cfDNA. This detection capability is similar to that reported for MBC error-corrected cfDNA-Seq with a 5-gene assay using amplicons (87.2%) (1) and a 54-gene assay using target-capture (85%) (14, 29). Furthermore, indels are more difficult to call than point mutations. Yet, our cfDNA assay called 23 of 26 indels (88.5%) that were known based on tumor sequencing, showing a similar performance to point mutation detection (87.7% called).

cfDNA-Seq detected several additional driver mutations not reported by tumor sequencing. Seven were in *TP53*. Two were also observed in the matched blood cells, indicating that they originated from clonal hematopoiesis. The discovery of clonal hematopoiesis in 7% of our cohort demonstrates the importance of sequencing DNA extracted from blood cells to avoid misinterpreting such variants as cancer-associated mutations. In one patient who received cetuximab therapy, we detected a *KRAS* Q61H variant that was absent from the matched tumor and likely represents the evolution of a drug-resistant subclone. Multiple *PIK3CA*-activating mutations detected in 2 anti-EGFR therapy-naive patients represent parallel evolution events. These examples show that our cfDNA assay can provide insights into cancer evolution. Because the minimally invasive nature of cfDNA-Seq allows application at multiple time points, this could be used to monitor the evolution of subclonal drug resistance driver mutations without prior knowledge of specific loci where resistance mutations will occur. We finally demonstrate that cfDNA-Seq allows genome-wide CNA reconstruction and validate this against low-coverage genome sequencing. As the number of targeted therapies increases, custom target-enrichment panels that can be readily adapted and scaled for the tumor type and therapeutic agent in question could be used to investigate the full tumor genomic landscape of point mutations, indels, and CNAs. This would facilitate the identification of novel resistance mechanisms. Importantly, this ultra-sensitive cfDNA-Seq technology can also address the subset of 20% of patients with mCRC who cannot be molecularly profiled because of unobtainable or inadequate biopsy tissues (16, 30).

In conclusion, this cfDNA-Seq approach with customizable and off-the-shelf reagents showed a similar performance to published techniques that use bespoke reagents and more complex analyses.

**Author Declaration:** A version of this paper was previously posted as a preprint on bioRxiv as [https://www.biorxiv.org/content/early/2017/11/02/213306](https://www.biorxiv.org/content/early/2017/11/02/213306). The authors had premarketing access to Agilent SureSelectXT-HS reagents.

**Author Contributions:** All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.

S. Mansukhani, L.J. Barber, and M. Gerlinger conceived the study and wrote the manuscript; S. Mansukhani, L.J. Barber, and B. Griffiths processed samples; S. Mansukhani, L.J. Barber, and B. Herman developed the cfDNA-Seq assay; D. Kleiofigalanis and S. Lise developed the DuplexCaller tool; S. Mansukhani, L.J. Barber, D. Kleiofigalanis, A. Woolston, M.N. Davies, and M. Gerlinger analyzed the data; S.Y. Moorcraft, M. Davidson, A. Patel, A. Okachi, I. Rana, R. Begum, D. Watkins, A. Wotherspoon, K. van Loga, I. Chau, D. Cunningham, N. Starling, and T. Powles provided clinical data and samples. K. Fenwick and N. Matthews sequenced the cfDNA libraries; P.Z. Proszek, D. Gonzalez de Castro, S. Hulkkki, and M. Hubank provided tumor biopsy sequencing data from the FOrMAT panel and ran the Avenio analysis; N. Turner and B. O’Leary provided support for ddPCR.

**Authors’ Disclosures or Potential Conflicts of Interests:** Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

**Employment or Leadership:** B. Griffiths, Institute of Cancer Research; B. Herman, Agilent Technologies UK Ltd.

**Consultant or Advisory Role:** I. Chau, Eli-Lilly, Bristol Meyer Squibb, MSD, Merck-Serono, Roche, Bayer, Five Prime Therapeutics.

**Stock Ownership:** B. Herman, Agilent Technologies.

**Honoraria:** I. Chau, Eli-Lilly.

**Research Funding:** The Moodie Report Foundation, National Institute of Health Research Biomedical Research Centre for Cancer funding (NIHR BRC grant number A3) to the ICR and Royal Marsden Hospital, Wellcome Trust Strategic Grant (105104/Z/14/Z) to the ICR Centre for Evolution and Cancer, charitable donations from Tim Morgan to the Institute of Cancer Research and from Philip Moodie to The Royal Marsden Cancer Charity. S. Mansukhani, Cancer Research UK Clinical PhD Studentship; S.Y. Moorcraft, Clive and Ann Smith Fellowship; M. Gerlinger, Cancer Research UK.

**Expert Testimony:** None declared.

**Patents:** None declared.

**Role of Sponsor:** The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation of manuscript, or final approval of manuscript.

**Acknowledgment:** The authors would like to thank all patients participating in the FOrMAT clinical trial and the clinical research team members at the Royal Marsden Hospital who supported the sample collection.
Error-Corrected Circulating Tumor DNA Sequencing

References

1. Bettgeowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014;6:224ra224.

2. Savaregna G, Marson S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. Nat Rev Clin Oncol 2017;14:531–48.

3. Heitzer E, Ull P, Belic J, Gutschli S, Quehenberger F, Fischereder K, et al. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. Genome Med 2013;5:30.

4. Haber DA, Velculescu VE. Blood-based analyses of cancer: circulating tumor cells and circulating tumor DNA. Cancer Discov 2014;4:650–61.

5. Leary RJ, Sausen M, Kinde I, Papadopoulos N, Carpten JD, Cragg D, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med 2012;4:162ra154.

6. Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. N Engl J Med 2013;368:1199–209.

7. Mok T, Wu NL, Lee JS, Yu CJ, Siu R, Sandoval-Tam J, et al. Detection and dynamic changes of EGFR mutations from circulating tumor DNA as a predictor of survival outcomes in NSCLC patients treated with first-line irinotecan and chemotherapy. Clin Cancer Res 2015;21:3196–3203.

8. Garcia-Murillas I, Schivo G, Weigelt B, Ng C, Hrebien S, Cutts RJ, et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci Transl Med 2015;7:302ra133.

9. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med 2017;9 eaan2415.

10. Abbo A, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantini S, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature 2017;545:446–51.

11. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat Genet 2014;46:225–33.

12. Diehl F, Li M, He Y, Kinzler KW, Vogelstein B, Dressman D. BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. Nat Methods 2006;3:551–9.

13. Perakis S, Speicher MR. Emerging concepts in liquid biopsies. BMC Med 2017;15:75.

14. Lamann RB, Mortimer SA, Zill OA, Sebisanovic D, Lopez R, Blau S, et al. Analytical and clinical validation of a digital sequencing panel for quantitative, highly accurate evaluation of cell-free circulating tumor DNA. PLoS One 2015;10:e0140712.

15. Newman AM, Lovejoy AF, Klass DM, Kurtz KT, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol 2016;34:547–55.

16. Moorcroft SY, Gonzalez de Castro D, Cunningham D, Jones T, Walker BA, Peckitt C, et al. Investigating the feasibility of tumour molecular profiling in gastrointestinal malignancies in routine clinical practice. Ann Oncol 2018;29:230–6.

17. Github. duplexCaller. https://github.com/dklefoogi/duplexFiltering/blob/master/duplexCaller.py (Accessed January 2018).

18. Github. bam-readcount. https://github.com/genome/bam-readcount (Accessed October 2017).

19. Talevich E, Shin H, Botton B, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol 2016;12:e1004873.

20. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, et al. Genome-wide copy number analysis of single cells. Nat Protoc 2012;7:1024–41.

21. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by duplex sequencing. Nat Protoc 2014;9:2586–2606.

22. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol 2011;29:24–6.

23. Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, et al. COSMIC: high-resolution cancer genetics using the catalogue of somatic mutations in cancer. Curr Protoc Hum Genet 2016;91:10.11.11–10.11.37.

24. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat Med 2014;20:1472–8.

25. Goombs CC, Zehr A, Devlin SM, Khitsktaga A, Syd A, Jonsson P, et al. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. Cell Stem Cell 2017;21:374–82.

26. Misiale S, Yaeger R, Huber S, Scala E, Janakiraman M, Loika D, et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. Nature 2012;486:532–6.

27. Biannon AR, Vakiani E, Sylvester BE, Scott SN, McDermott G, Shah RH, et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. Genome Biol 2014;15:454.

28. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330–7.

29. Kim SJ, Lee WS, Lamann RB, Mortimer S, Zill OA, Kim KM, et al. Prospective blinded study of somatic mutation detection in cell-free DNA utilizing a targeted 54-gene next generation sequencing panel in metastatic solid tumor patients. Oncotarget 2015;6:40360–9.

30. Khakoo S, Georgiou A, Gerlinger M, Cunningham D, Starling N. Circulating tumour DNA, a promising biomarker for the management of colorectal cancer. Crit Rev Oncol Hematol 2018;122:72–82.