Mansukhani *et al.* Supplemental Methods

Patients and samples

Samples from 6 healthy donors (HD) were collected after obtaining written informed consent through the Improving Outcomes in Cancer biobanking protocol at the Barts Cancer Centre (PI: Powles), which was approved by the UK national ethics committee (approval number: 13/EM/0327). 27 ml of blood were collected into EDTA tubes from each donor during a single blood draw.

Blood samples from 58 patients with metastatic colorectal cancer were acquired after obtaining written informed consent through the FOrMAT clinical trial (Feasibility Of Molecular characterization Approach to Treatment, PI: Starling, ClinicalTrials.gov NCT02112357) at the Royal Marsden Hospital, which was approved by the UK national ethics committee (approval number: 13/LO/1274RM). FOrMAT is a single centre translational study assessing the feasibility of next generation sequencing to guide treatment personalisation in patients with advanced gastrointestinal tumors. Blood samples had either been collected from treatment naïve patients (n=19) or at the time tumors started to progress after prior palliative systemic treatment (n=39).

Archival or fresh tumor specimens from patients enrolled in the FOrMAT trial had been sequenced with solution hybrid capture enrichment of 46 cancer driver genes relevant for gastrointestinal cancers as described (Moorcraft SY *et al.* 2018 Annals of Oncology 29:230). Mutation calls, BAM files of tumor and blood sequencing data as well as clinical data were available for our analyses.

Blood sample processing

The EDTA anticoagulated blood was centrifuged within 2 h (10 min, 1600 g). Plasma was aliquotted and stored in sterile 2 ml Nalgene cryogenic vials at -80°C. Upon thawing, samples were centrifuged (10 min, 16000 g, 4°C) and cfDNA was extracted from 4 ml plasma per patient with mCRC and from 2x4 ml from healthy donors using the Qiagen QIAamp Circulating Nucleic Acid Kit. cfDNA was eluted in 30 μl 10 mM Tris 0.1 mM EDTA and stored at -20^{o}C. cfDNA fragments in the range 100-700 bp were quantified on a Bioanalyzer High Sensitivity DNA chip (online Supplemental Figure 2A), and if this showed a high concentration, on a 7500 DNA chip (Agilent).

Determining the optimal cfDNA quantity for mCRC analysis

cfDNA was extracted from the plasma of 58 patients with mCRC enrolled in the FOrMAT trial (online Supplemental Figure 2B). Resulting yields were used to define the cfDNA quantity that could be extracted from >95% of patients with mCRC if 10 ml plasma were collected through a standard blood draw of 20-30 ml (online Supplemental Figure 2C). 25 ng cfDNA were obtained reliably from >95% of mCRC cases. With a haploid human genome mass of ~3.3 pg, this should have contained >7500 genome equivalents which was sufficient for the detection of mutations with frequencies ~0.1% VAF, even if only 20% of the cfDNA fragments were incorporated into sequencing libraries (Cai X *et al.* 2015 Trends in Genetics 31:564). Therefore 25 ng was used as the standard input quantity for cfDNA sequencing.

Library Preparation

We modified the Agilent SureSelect^{XT-HS} protocol in order to assure a reliable performance with 25 ng cfDNA input. All PCR steps were performed on an Eppendorf Mastercycler nexus GSX1/SX1e. 8 cycles of pre-hybridization PCR were optimal for 25 ng of input cfDNA to generate the amount of product required (500–1000 ng) for in-solution capture. The entire product was used as input for hybridization to our custom-designed Agilent SureSelect capture bait library, targeting 163.3kb comprised of 32 genes and 40 SNP positions on chromosome 18q (online Supplemental Table 1). All other reagents were added according to the manufacturer's protocol and 60 cycles Fast Hybridization were performed, taking ~1.5h. Capture was started immediately after the final hybridization cycle and proceeded for 30 minutes at room temperature.

Post-capture washes were performed with two different conditions. The manufacturer's initial protocol recommended 3 incubations of 10 min each at 65^{o}C, followed by 9 post-capture PCR cycles. We subsequently used more stringent post-capture washes with 6 incubations of 5 min each at 70^{o}C, followed by 12 post-capture PCR cycles. All other conditions remained the same while post-capture wash conditions were varied. Our final optimized protocol incorporated higher stringency post-capture washes with 10 post-capture PCR cycles to compensate for the lower amount of non-specific DNA carryover. Seven patients with mCRC were sequenced with the original protocol, and the optimized cfDNA library preparation protocol was used for the remaining 21 patients with mCRC and 6 healthy donors.

10 PCR cycles were used for post-capture amplification and this was followed by two rounds of 1x Ampure XP bead cleanup to remove un-incorporated primers. The final prepared sequencing libraries were profiled on a Bioanalyzer High Sensitivity DNA chip and quantified by qPCR using the Kapa Library Quantification kit before pooling. Pooled libraries were clustered using the Illumina cBot and sequenced with paired-end 75 reads on an Illumina HiSeq2500 in rapid mode.

Variant Calling

Variant calling (single nucleotide variants and indels) using the SureCall SNPPET Caller was performed with the following parameters: Variant score threshold =0.01; Minimum quality for base =30; Variant call quality threshold =40 (manufacturer default setting: 100); Minimum Allele Frequency =0.001; Minimum number of reads per barcode =2; no region padding; and masked overlap between reads. Further filtering of the primary calls was performed after visualizing sequencing data on IGV.

Bioinformatic Identification of duplexes

Each of the two strands of a double-stranded (“duplex”) cfDNA molecule is labeled with a different MBC (blue or red boxes). The resulting paired-end sequencing data will be orientated differently for the two original strands. The original forward strand of the cfDNA duplex will result in read pairs where the forward read (relative to the reference genome) will be read 1 and the reverse read will be read 2 (Figure 1D). The reverse strand of the original cfDNA duplex will result in read pairs with the opposite orientation: i.e. the reverse read is read 1 and the forward read is read 2. Hence the order of the read pair sequences with respect to the reference genome alignment can be used to detect likely duplex DNA strands. To identify variants supported by duplexes we developed the duplexCaller bioinformatics tool. This tool took as input the SureCall variant calls (in a VCF-like format) and the corresponding MBC de-duplicated sample BAM file. SAMtools (Li H *et al.* 2009 Bioinformatics 25:2078) was used to identify only the consensus reads that spanned the genomic positions of interest. The list of consensus reads was parsed in the SAM format, selecting and saving the start position of each consensus read (i.e., field POS with the leftmost mapping position), the start position of the next read in the template (i.e., field PNEXT with the leftmost mapping position of mate read) as well as the length of the template (field TLEN), and the bitwise FLAG field. By applying memory-efficient indexing of the relevant information, a data structure was generated for all consensus reads supporting variants that could potentially form duplexes (i.e. a candidate set of duplexes). In addition, this technique allowed discarding of the non-relevant information for every input variant, and thus it improved the execution time. Note that during the population of the data structure, read pairs were not considered if mates were mapped to different chromosomes (i.e., RNEXT flags not equal to ‘=’).

Next, the deployed data structure was used to process all fragments that start and stop in the same genomic positions. By parsing the list of fragments with the same start and stop positions, the consensus reads being first in the template and second in the template for each fragment could be identified using bitwise operation to the FLAG field of the SAM format. In this way, given a set of fragments with the same start and stop positions, the number of duplexes was the minimum number of pairs of fragments having their leftmost consensus read as first in the template, and second in the template respectively. As an output, the duplexCaller reported the total number of duplexes found for every variant call that passed the initial filtering.

Further Variant Filtering

Variants predominantly located in reads with an alignment score of zero or in reads with multiple non-contiguous non-reference bases were removed because these were usually indicative of misalignment. Where the insert size was < 74 and all 74 bases of the read were aligned, any variants occurring at the end of the read were excluded as these represent erroneous mapping of the adapter sequence. All variant positions identified in patient cfDNA were checked in the dataset of six HD samples and each HD was checked against the rest of five HD samples using bam-readcount (https://github.com/genome/bam-readcount). The majority of called variants were completely absent in the six HD samples (online Supplemental Table 2), however, mutations whose VAF was not at least double that of an identical variant in a healthy donor sample were discounted in order to filter out recurrent false positives.

Standard de-duplication

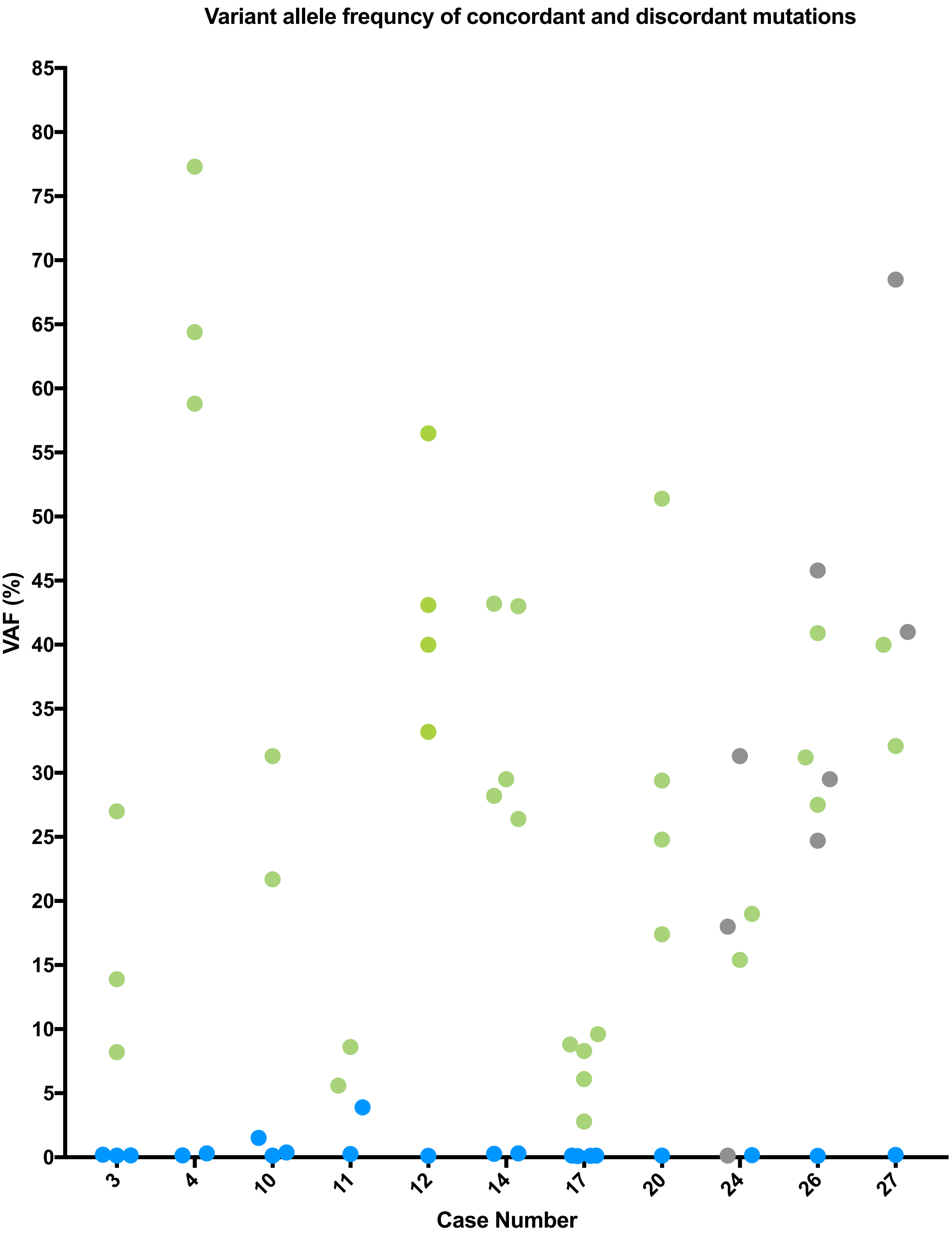
To compare our MBC analysis with current methods, standard de-duplication and subsequent calling was also performed with SureCall using the same parameters but without inclusion of the MBC, such that de-duplication was solely based on genome alignment and insert size. As the post-call filtering of MBC de-duplicated data essentially required a variant to be present in a minimum of three reads (two that form a duplex and one additional read with a different alignment position), we also mandated a minimum of three reads with a specific variant to support a mutation call in the standard-duplicated data to enable a fair comparison.

ddPCR

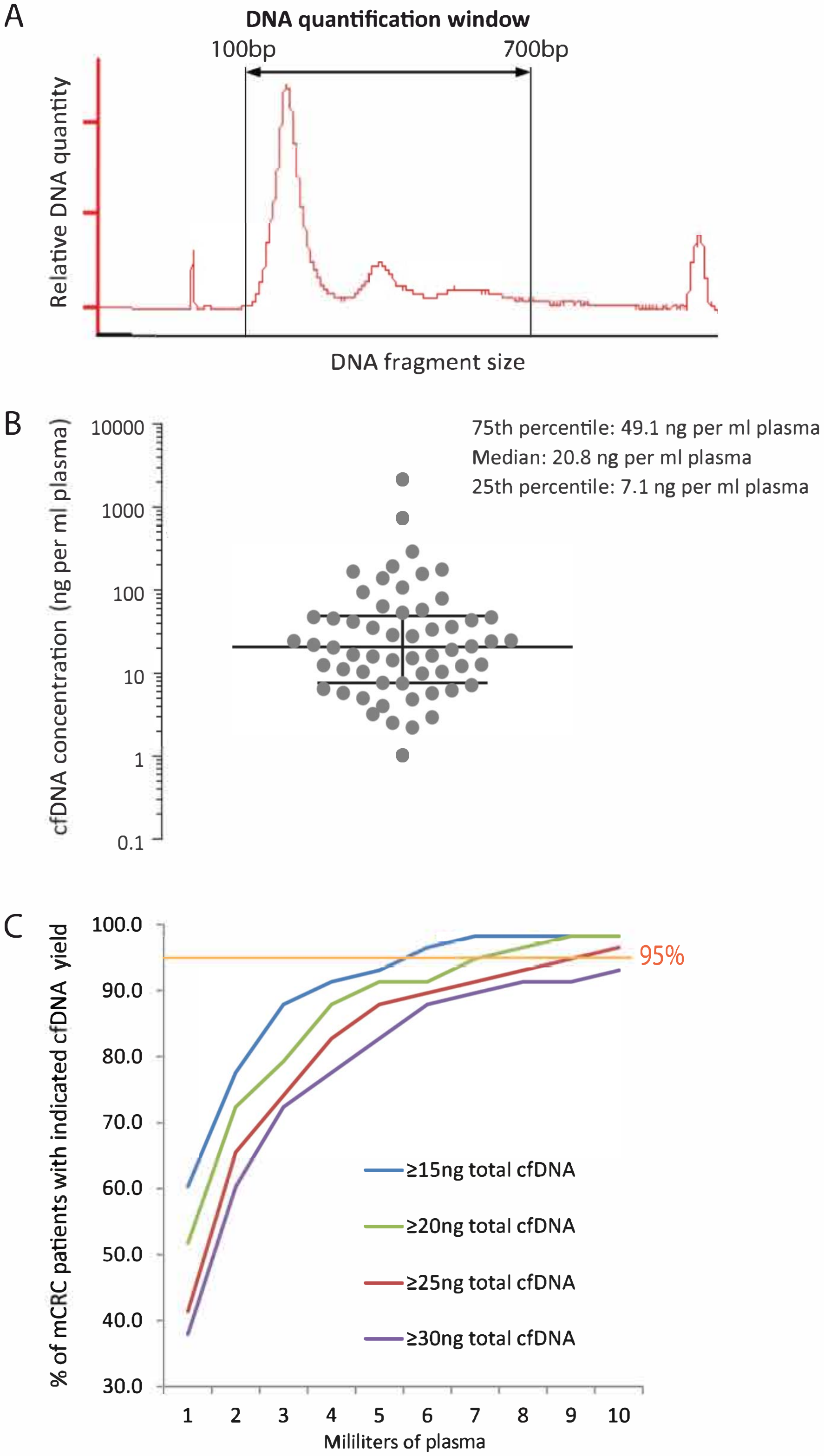
Digital droplet PCR was performed for independent validation of mutations called in cfDNA. The number of variants selected for orthogonal validation was limited by remaining cfDNA material and success of custom ddPCR assay design. Of the 11 cases where low-level subclonal mutations were identified in cfDNA, sufficient DNA for ddPCR was available from 4 cases (4, 10, 14, and 24), for which 9 discordant cfDNA+/tumor- variants were selected for independent validation (case 4: *ATM* L1405S, *TP53* H178P, case 10: *KRAS* Q61H, *ATM* G449A, *TP53* A355S, case 14: *ATM* R2993*, *PIK3CA* E545K, and case 24: *CTNNB1* V676I and *TP53* I254V). Pre-validated commercially available ddPCR SNP Genotyping Assays were used for *BRAF, KRAS* and *PIK3CA* (Life Technologies; Assay ID A44177 *BRAF*_476, A44177 *KRAS*_555, A44177 *PIK3CA*_763) and the remaining probes were custom designed by Thermo Fisher. Custom probes could not be designed for *ATM* L1405S and the *TP53* A355S assay could not be established due to a failure to detect mutant droplets in positive control experiments. 7/9 ddPCR assays were successfully validated with positive and negative controls.

Input cfDNA for each ddPCR experiment varied, depending on amount of residual material available (17 ng for case 10, 20 ng for cases 8 and 14 and >25 ng for the other cases). cfDNA was added to a ddPCR reaction containing 11 μl mastermix (10 μl 2x ddPCR Supermix for Probes and 1 μl 20x target primer/probe mix for both mutant and wild type alleles) and made up to a total volume of 21 μl with nuclease-free water.

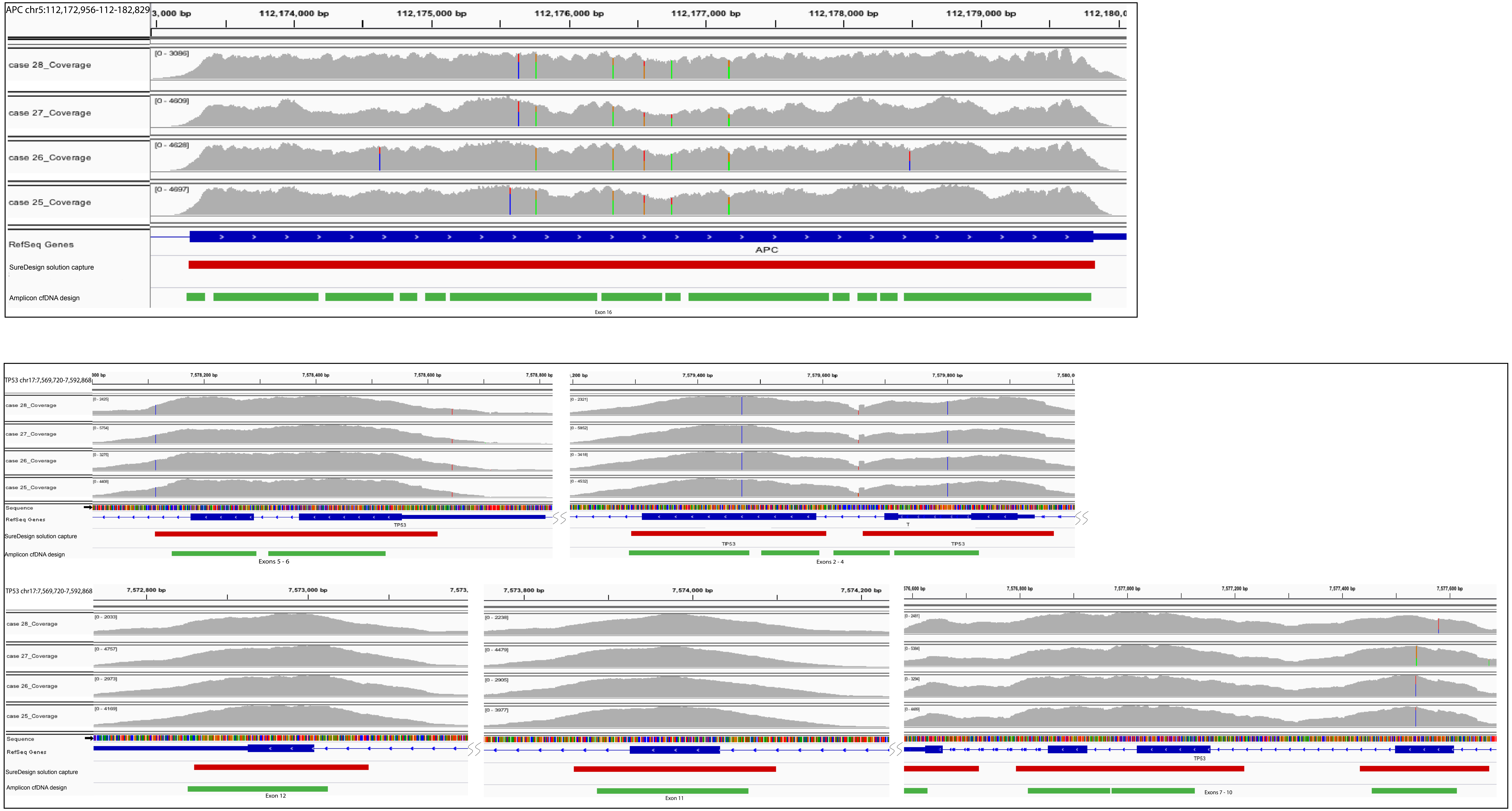
The reaction was partitioned into a median of 17,676 droplets per sample in a Bio-Rad QX-200 droplet generator according to the manufacturer's protocol. Emulsified PCR reactions were run on a 96 well plate on a G-Storm GS4 thermal cycler incubating the plates at 95^{o}C for 10 min followed by 40 cycles of 95^{o}C for 15 s and 60^{o}C for 1 min. Plates were read on the QX200 droplet reader using QuantaSoft analysis software (Bio-Rad) to acquire and analyze data. At least four positive control (patient cfDNA for *KRAS* Q61H and *BRAF* V600E, or gBlock controls) and negative control wells were run to verify assay performance and facilitate thresholding in fluorescence values. For each patient, plasma was analyzed in triplicate and ddPCR results based on the combined data of these wells (online Supplemental Table 6).



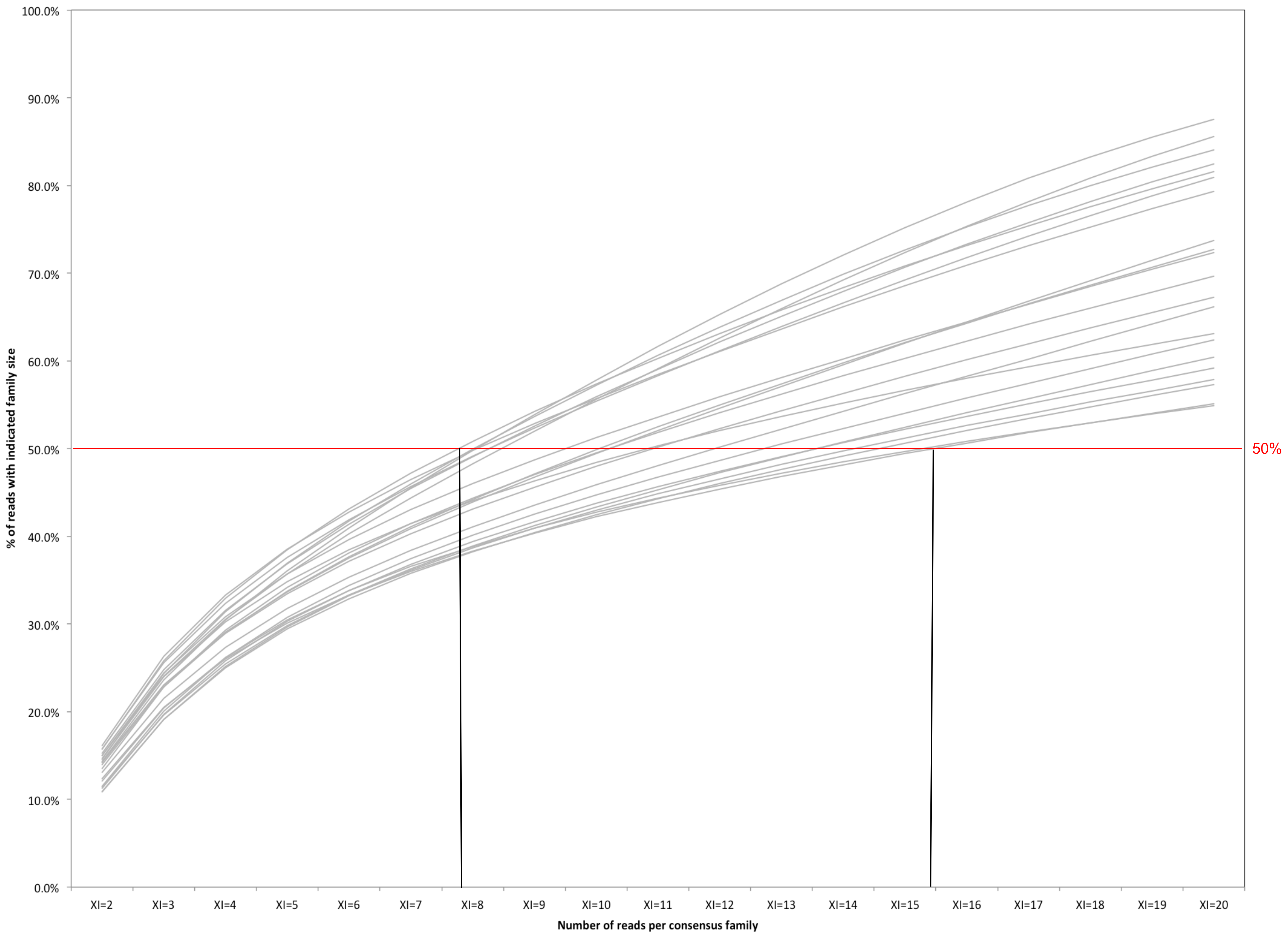
Supplemental Figure 1: Comparison of variant allele frequencies of mutations called in cfDNA and not in matched tumor tissue (blue). Variants identified by tumor sequencing and cfDNA sequencing are colored green. Variants identified by cfDNA-Seq only are coloured grey.



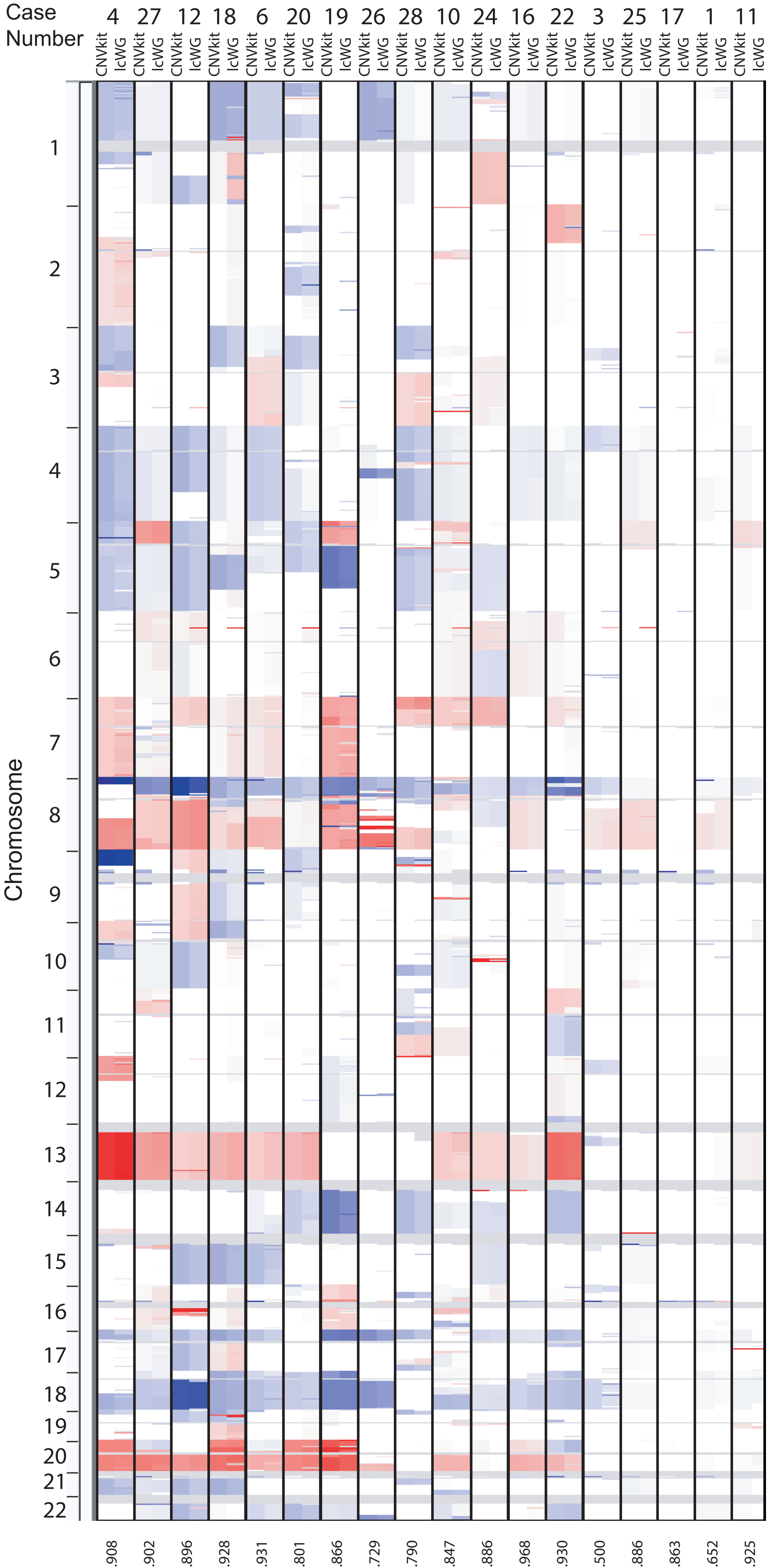
Supplemental Figure 2. (A) Example of a cfDNA sample Bioanalyzer profile. cfDNA was quantified across a window from 100 bp to 700 bp in all samples in this study. (B) cfDNA yields (ng per ml plasma) obtained from 58 patients with metastatic colorectal cancer. A median of 20.8 ng cfDNA per ml of plasma were extracted with a minimum of 1.03 ng/ml. (C) Fraction of patients that achieve the indicated total cfDNA yield (see legend) based on the plasma volume available from each patient.



Supplemental Figure 3: The Amplicon design was generated for the Illumina TruSeq amplicon sequencing platform using ctDNA settings and median stringency for key colorectal cancer driver genes (KRAS, NRAS, BRAF, APC, TP53, CTNNB1, PIK3CA, FBXW7, PTEN, TCF7L2, ATM and SMAD4; all genes are included in our cfDNA-Seq panel). Differences in coverage between solution capture and amplicon design are shown for the large final exon of APC gene, that harbours the majority of mutations in colorectal cancer, and TP53 (exons 2-12).



Supplemental Figure 4: Cumulative consensus family size distribution. The median number of reads (XI) per consensus family varied between 8 and 15 across all 21 samples sequenced with the optimized protocol.



Supplemental Figure 5: Comparison of genome wide heatmap of segmented copy number data using off-target reads from cfDNA-Seq (CNVkit) and low coverage whole genome (lcWG) sequencing. Gains are red and losses are blue. Profiles are ordered (left to right) from highest to lowest tumor content.