SUPPLEMENTARY DATA

Supplementary Methods

Identification of somatic copy number alterations by SNP-array data

Matched germline and tumor DNA were profiled using SNP6.0 arrays (Affymetrix). Initial QC of the arrays was performed using the Birdsuite software.[1] The QC acceptance criteria were contrast QC >40, SNP call % >97, and MAPD <0.4. Tumor specific copy number alterations (CNAs) were derived from each tumor/normal pair by running the "CRMA (v2): Paired total copy number analysis" from the aroma.cn package in R.[2] The package is available at aroma-project.org.

Illumina mate pair libraries and next generation sequencing

Initially, we used the Illumina Mate Pair Library Preparation Kit v2 together with the TruSeq ™ DNA Sample Preparation Kit to allow indexing of mate pair libraries. Briefly, 2.5 ug of high molecular weight genomic DNA (gDNA) was fragmented by Covaris Adaptive Focused Acoustics™ (AFA) sonication device (S2, Covaris, Inc.) to a fragment size of 2000-5000 bp. (Duty cycle 20%, intensity 0.1, 1000 cycles burst, 5 minutes). Sonication was performed in AFA miniTUBE (Part # 520066). Samples were speedvaced and the fragment size analyzed on an Agilent DNA chip (Bioanalyzer). Following Biotinylation, according to the Illumina protocol, the samples were indexed by performing end repair, A-Tailing, and adapter ligation as described in the TruSeq™ protocol. Finally the libraries were enriched by 18 cycles of PCR. The DNA concentration of the libraries was evaluated by Q-PCR (KAPA Library Amplification Kit, KK2611; KAPABIOSYSTEMS) and the insert size distribution was measured by Agilent DNA 1000 Analyzer Chip. Lately, we used the Nextera Mate Pair Sample Preparation Kit (FC-132-1001, Illumina) that utilizes a gel free protocol and reduces the input to 1 ug of genomic DNA. We adjusted the concentration of each library to 2 nM and prepared clusters on Illumina paired end flow cells using the manufacturer’s instructions and the appropriate Illumina cluster and sequencing kits. Illumina sequencing was performed on the Illumina HiSeq2000 platform. Here, we employed the sequencing kits TruSeq SBS Kit v3 - HS (200-cycles) (FC-401-3001, Illumina) and TruSeq PE Cluster Kit v3 - cBot - HS (PE-401-3001, Illumina) to generate 2x101 bp paired end sequencing. Average sequence read depth/physical read depth is listed in Supplementary Table 2.

Data Analysis

Fastq files were prepared with CASAVA (v1.8.2) and quality checked using FastQC and FastqScreen (http://www.bioinformatics.babraham.ac.uk/projects/). For data generated by the Illumina Mate Pair Library Preparation Kit v2 the read length was cut down to 50 base pairs to minimize the number of chimeric reads as a result of reading through the circulation ligation point. For the Nextera DNA Sample Preparation Kit the internal transposon adapters were removed using Cutadapt (https://code.google.com/p/cutadapt/) and Illumina adapters were removed using AdapterRemoval (v1.5).[3] Next the reads were reverse complemented using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and mapped to hg19 using BWA (v0.6.1-r104).[4] Insert size and PCR and optical duplicates were marked in each library using the Picard package v1.88 (http://picard.sourceforge.net). To clean the data, unmapped reads, PCR duplicates, and read pairs with insert sizes below 1000 bp were removed. BreakDancer[5] was applied to the final alignments with a mapping quality cut-off at 10 for the germline samples and 35 for tumor and metastasis samples to identify discordant mapping reads and to annotate them according to the underlying type of structural variant e.g. deletion (DEL), intra chromosomal translocation (ITX). Inter chromosomal translocation (CTX), inversion (INV), and insertion (INS). Bed-files were produced to visually inspect and compare the BreakDancer output with the SNP array copy number data in the IGV browser.[6] To check for sample mixup, SNPs were called using samtools mpileup and hierarchical clustering of all libraries was performed using the mitochondrial SNPs.

Identification of somatic structural variants

We applied two different pipelines to identify SSVs (Figure 1). Pipeline 1 was based on SNP array data to identify somatic copy number alterations (CNA) by comparing tumor and blood DNA. CNAs were taken further if they were supported by 15 or more probes, had a genomic size above 20,000 bp, and had a mean log2 copy number ratio amplitude >|0.2|. The somatic CNAs were compared to the tumor mate pair analysis to identify read pairs supporting the SNP findings. The number of read pairs supporting each CNA was logged. Four to eight SSVs from each patient were selected for validation based on the amplitude of the SNP analysis, the number of supporting reads in the mate pair analysis, and the possible biological function of the involved genes. SSVs with a high absolute log2 copy number ratio amplitude and many supporting reads in the mate pair analysis was preferred, as this indicate that many tumor cells contain this allele, or the region is amplified in all or some of the tumor cells. SSVs affecting known drivers of CRC tumorigenesis were also preferred as these are likely to have occurred early in tumor development and therefore to be present in the majority of tumor cells and subsequent metastasis. In pipeline 2, the SSVs were identified by analysis of matched germline DNA, but otherwise the selection criteria were the same as in pipeline 1. Potential SSVs where evaluated by PCR using tumor and germline DNA. PCR products were visualized on an agarose gel and Sanger sequenced across the SSV junction to annotate the SSV breakpoint at base-pair resolution.

Isolation and quantification of DNA

DNA was extracted from fresh frozen tissue using the Puregene DNA purification kit (Gentra Systems), from FFPE tissue using the QiaAMP DNA FFPE Tissue kit (56404) with O/N incubation with proteinase K at 56˚C and 550 rpm, from 2-4 mL plasma by QIAamp DNA Blood Midi Kit (51185), modified to be able to allow two mL of plasma on one column, or QIAamp Circulating Nucleic Acid Kit (55114) according to the manufactures instruction.

Amplification of SSVs by multiplex Nested PCR

For the analysis of patients 10 and 16, twelve cycles of Nested PCR were carried out with seven sets of primers (Supplementary Table 3) using 90% the DNA purified from the plasma samples (the remaining 10% were used to estimate CPP1, cfDNA quantity, and leucocyte DNA contamination) as template in a final volume of 50 µl and the following final concentrations: primer mix 0.5 µM, dNTP 0.2 mM, PCR buffer 1x, $MgCL_2$ 1.5 mM, Tempase enzyme 0,1U/µl. The Nested product was diluted between four and two hundred times before being used as template in ddPCR.

SUPPLEMENTARY RESULTS

Detection limit

For all samples we stablished a procedure for assessing the minimal ctDNA/cfDNA ratio detectable in a given sample. For a sample with no leucocyte DNA contamination and for which cfDNA purification was 100% efficient the minimal detectable fraction (the detection limit) is 1/(estimated cfDNA quantity). Applying this detection limit evaluation procedure to the marker negative samples revealed that our ability to detect ctDNA in them was similar to that of the marker positive samples (Supplementary Figure 1-2 and 3A-C).

Tumor and metastasis heterogeneity

We analyzed the initial fresh frozen sample, utilized for mate-pair sequencing, and seven additional tumor FFPE biopsies reflecting different topological locations in the primary tumor by ddPCR (Figure 4). Three SSVs C16D2, C20D1, and C20D2 were found in all biopsies indicating that these two SSVs occurred early in tumor development. Consistent with these observations C16D2, C20D1, and C20D2 also had the most supporting read pairs in the mate-pair analysis and could be detected in the pre-op plasma samples. Analysis of FFPE punch biopsies from the early and late liver lesions revealed that C20D1 and C20D2 were present in both and C16D2 only in the early metastasis (Figure 4D). The sequencing analysis of the late metastasis confirmed that it did not carry the C16D2 deletion at the RBFOX1 locus seen as the most frequent SSV in both the primary tumor and the early metastasis. Nor did it carry any of the other RBFOX1 deletions seen in the primary tumor (Figure 4A, Supplementary Table 4). A plausible explanation for the observed findings is that early on in tumor development tumorigenic cells carrying only the C20D1 and C20D2 SSVs metastasized to the liver, but lay dormant. Meanwhile, the C16D2 deletion occurred in the primary tumor, in cells harboring the C20D1 and C20D2 SSVs. One of these cells metastasized to the liver, where it expanded to a clinical manifest lesion, which subsequently became completely eliminated by the partial hepatectomy at month 3. Eventually, the dormant C20D1 and C20D2 positive cells started proliferating and formed the late-occurring metastasis.

REFERENCES

1 Korn JM, Kuruvilla FG, McCarroll SA*, et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature genetics 2008;**40**:1253-60.

2 Bengtsson H, Wirapati P, Speed TP. A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. Bioinformatics 2009;**25**:2149-56.

3 Lindgreen S. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC research notes 2012;**5**:337.

4 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;**25**:1754-60.

5 Chen K, Wallis JW, McLellan MD*, et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature methods 2009;**6**:677-81.

6 Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics 2013;**14**:178-92.