Master of Science in Omics Data Analysis

Master Thesis

Identification of chromosomal rearrangements in colorectal cancer

by

Ferran Moratalla Navarro

Supervisor: Dr. Victor Moreno, Unit of Biomarkers and Susceptibility. Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO)

Co-supervisor: Dr. Mireia Olivella, Systems Biology, University of Vic – Central University of Catalonia

Department of Systems Biology

University of Vic – Central University of Catalonia

2015, September the 15th

Acknowledgements

I would like to thank the staff at Unit of Biomarkers and Susceptibility, especially to Rebeca Sanz-Pamplona, Daniel Aguilar, Adriana López-Doriga, Susanna Aussó, David Cordero, Francisco D. Morón-Duran, Henar Alonso, Xavier Junyent and Anna Díez.

I would also like to express my sincere gratitude to my master thesis supervisor, Dr. Victor Moreno, who has given me the chance to perform this project.

And finally thank my partner Marta, for her constant encouragement throughout my study period.

Abstract

Cancer research is continuously shedding light into these worldwide leading diseases. It is mandatory to have higher knowledge in cancer biology to consequently find out new candidate biomarkers and therapeutics. Among all of them, Colorectal cancer is the most commonly seen of human malignant cancers and has the third highest mortality rate[1]. Since the release of the first human genome sequence in 2004, new techniques have revolutionised the study of genetics and its possible applications. A broad type of studies has been carried out; being Single Nucleotide Polymorphisms and Copy Number Variants the most intensively studied analysis. However, other kinds of mutations involving larger parts of the genome, the so-called structural variants, have been substantially less analyzed due to technical limitations. High-throughput sequencing methods seem to have lowered these restrictions.

In this study, gene fusions have been searched in whole exome sequencing samples taking 42 paired normal and cancer tissues. Beginning with short-read files obtained with the mentioned method, they have been aligned against a reference genome to later be analyzed with Breakdancer, a structural variant calling algorithm. After some filtering criteria performed in order to remove a high proportion of false positives, a highly probable list of 22 balanced structural variants (translocations and/or inversions) has been manually studied to get a final result of 20 chromosomal rearrangements, 8 of which are considered gene fusions. In addition, it has been found that one recurrent translocation seen in recent studies is indeed a false positive. Further studies taking into account these results may contribute to the findings of new biomarkers for certain subtypes of colorectal cancer.

Table of Contents

List of Abbreviations

SNP – Single Nucleotide Polymorphism

SV – Structural Variant

BCR-ABL – Breakpoint Cluster Region Abelson Murine Leukaemia Virus Oncogene homolog1

WES – Whole Exome Sequencing

DNA – Deoxyribonucleic acid

CLX – Colonomics project

NCBI – National Cancer for Biotechnology Institution

CRC – Colorectal Cancer

hg19 – Human Genome version 19

hg19/GRCh37 – Human Genome version 19 build 37

GUSTAF - Generic mUlti-SpliT Aligner Finder

FACTERA – Fusion And Chromosomal Translocation Enumeration and Recovery Algorithm

CNV – Copy Number Variants

bp – Base Pairs

NSCLC – Non-Small Cell Lung Cancer

SLC34A2 – Solute Carrier Family 34 member 2

ROS1 – ROS proto-oncogene 1

ALK – Anaplastic Lymphoma receptor tyrosine Kinase

FISH – Fluorescence In Situ Hybridization

EML4-ALK - Echinoderm Microtubule associated protein Like4-Anaplastic Lymphoma receptor tyrosine Kinase

ITX – Intrachromosomal Translocation

CTX – Interchromosomal Translocation

INV – Inversion

UCSC – University of California – Santa Cruz

IGV – Integrative Genomics Viewer

CV – Coefficient of Variation

DFFB – DNA Fragmentation Factor Beta polypeptide

PMF1-BGLAP – Polyamine Modulator Factor 1-Bone gamma carboxyglutamate Gla Protein

OR10K1 – Olfactory Receptor family 10 subfamily K member 1

FARP2 – FERM, RhoGEF And Pleckstrin Domain Protein 2

MYLK – Myosin Like chain Kinase

CEP70 – Centrosomal Protein 70kDa

CUL7 – Cullin 7

PTK7 – Protein Tyrosine Kinase 7

OPN5 – Opsin 5

EIG121L – Estrogen-Induced Gene 121-Like protein

MTSS1 – Metastasis Suppressor 1

CCAT1 – Colon Cancer Associated Transcript 1

PTAR1 – protein Prenyltransferase Alpha subunit Repeat containing 1

FANK1 – Fibronectin type III and Ankyrin repeat domains 1

RNF219-AS1 - RNF219 Antisense RNA 1

NDUFAB1 – NADH Dehydrogenase (Ubiquinone) 1, Alpha/Beta Subcomplex 1

XPO6 – Exportin 6

FBRS – Fibrosin

ZNF423 – Zinc Finger protein 423

DNAH9 – Dynein, Axonemal, Heavy chain 9

DLG2 – Discs, Large Homolog 2

UNC45B – Unc45 Homolog B

AP2B1 – Adaptor-Related Protein complex 2, Beta 1 subunit

HOXB3 – Homeobox B3

CIDEA – Cell Death-Inducing DFFA-Like Effector A

CYB5A – Cytochrome B5 Type A

NKX2-4 – NK2 Homeobox 4

PAX1 – Paired box 1

LOC284801 – Locus 284801

BCL2L1 – B-cell CLL/lymphoma 2 Like 1

SCARNA15 – Small Cajal Body-Specific RNA 15

NELFCD – Negative Elongation Factor Complex Member C/D

LOC284857 – Locus 284857

NR0B1 – Nuclear Receptor subfamily 0, group B, member 1

HUWE1 - HECT, UBA and WWE domain containing 1

rRNA – ribosomal Ribonucleic acid

1. Introduction

Advances in high-throughput sequencing during the last years have enabled the potential to identify almost all kinds of variations in any genomic region[2], from single nucleotide polymorphisms (SNPs) to large structural variants (SVs).

Structural variants can impact on the genome variation and therefore, it is important to consider them in cancer genetics[3]. Structural variants can be classified as germline (being most of them benign) and somatic[4]. In many cancer genomes, recurrent chromosomal rearrangements in specific types of cancer have been detected, such as BCR-ABL fusion gene from a resulting translocation between chromosomes 9 and 22 found in a high proportion of Chronic Myelogenous Leukaemia patients.

Due to the genome changes driven by structural variants (formation of fusion genes, changes in regulatory elements or changes in copy number) significant differences in both overexpression of oncogenes and underexpression of tumor suppressor genes may take place [5]. Therefore, detecting correctly all kinds of variations in a specific cancer genome could help in the understanding of the disease biology. In addition, it could allow designing new personalized therapies as well as identifying oncogenic-driving-mutations that can be therapeutically targetable in the near future.

Colorectal cancer accounts for 9.4% and 10.1% of all types of cancer in men and women, respectively. Is a major case of morbidity and mortality throughout the world although not uniformly spread. Western countries (Europe, United States, Canada, Australia and New Zealand) have higher incidence of colorectal cancer compared with African and Asian countries[6].

Like other types of malign tumours, the overall survival is dramatically dependent on the stage of the disease at diagnosis, ranging from 90% of 5-year survival rates for patients detected at an early stage at diagnosis, to 10% in patients detected at later stages[7]. For this reason it is extremely important to develop not only effective therapies but also high sensitivity biomarkers.

Few articles published recently have found several fusion genes in colorectal cancers, some of which are recurrent events, like NAV2-TCF7L1[8] or TMP3-NTRK1[9]. However, it seems that chromosomal rearrangements are not very common in these diseases[10], in contrast with haematological malignancies, where there have been found thousands of chromosomal rearrangements and hundreds of gene fusions.

This study is an attempt to detect balanced chromosomal rearrangements (translocations and inversions) and especially, those that generate fusion genes in colon cancer DNA samples. Consequently, it can be a method to perform fusion gene detection from whole exome sequencing (WES) data for further studies.

2. Materials and methods

In order to carry out this analysis, an important number of software and applications have been used. The major steps developed to obtain accurate results are represented in a flow diagram showed in Figure 2.1.

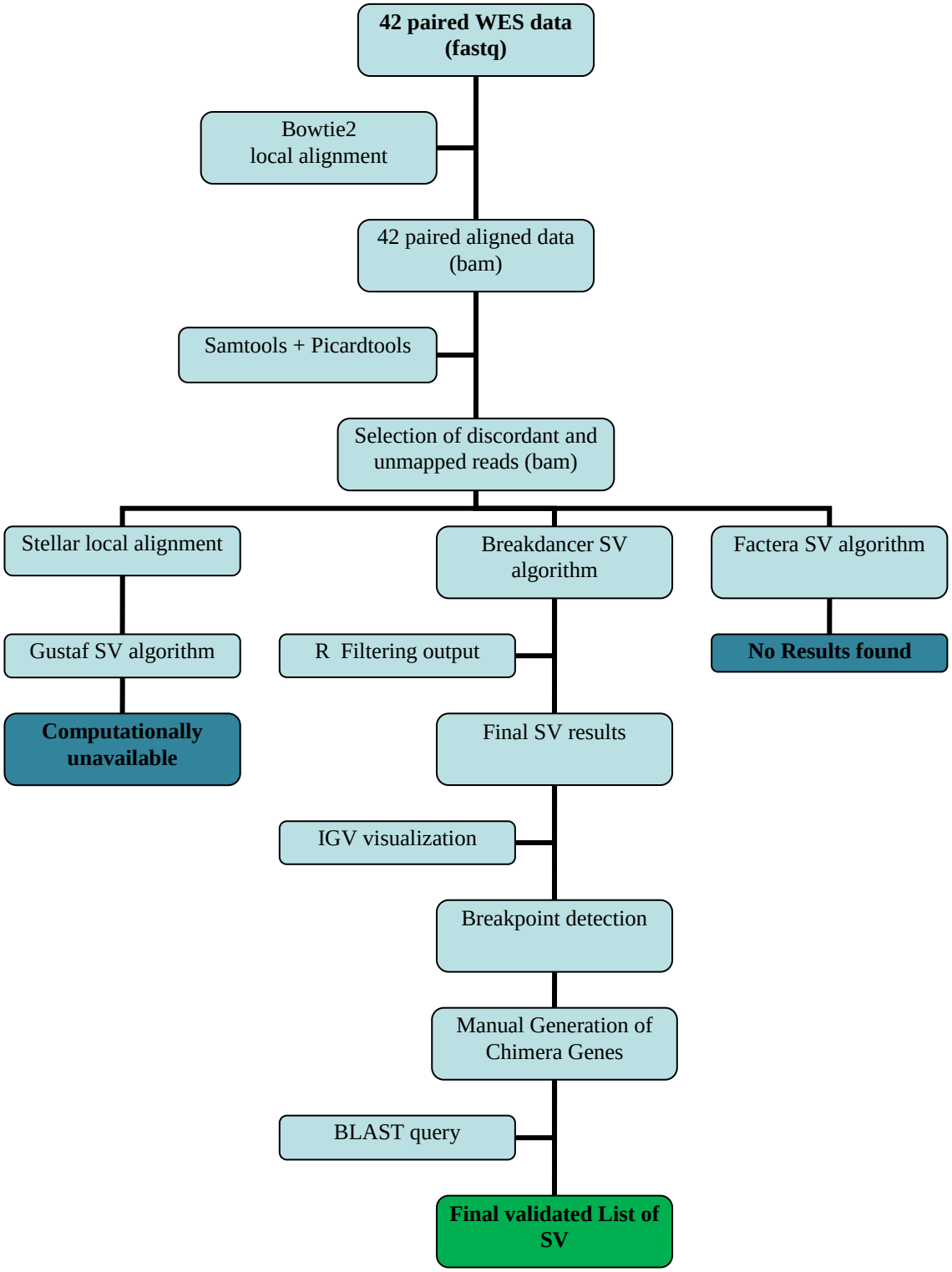


Figure 2.1. Flow diagram of the pipeline followed in this study.

2.1. Patients and samples

The study included a subset of 42 paired adjacent normal and tumor tissues (84 samples) from a previously described set of 100 patients with colon cancer diagnosed at stage II (colonomics project –CLX-: www.colonomics.org; NCBI BioProject PRJNA188510). All patients were recruited at the Bellvitge University Hospital (Barcelona, Spain). Written informed consent was obtained from all patients and the Institution's Ethics Committee approved the protocol. DNA was extracted using a standard phenol-chloroform protocol. To ensure that adjacent and tumor tissues were paired, dynamic arrays were used to genotype 13 SNPs in the 84 samples. All 42 adjacent normal tissues correctly matched with their corresponding tumor. Tumor DNA from an additional series of 227 CRC patients from the same hospital was used for validation purposes. This extended series was not restricted regarding site, stage and microsatellite instability phenotype[11].

2.2. Quality control of samples

CLX samples analyzed here, have been studied before[11]. For this reason, it has not been necessary to perform a quality control of the samples because it has been done previously. Anyway, fastQC[12] software was used to confirm the adequate quality of the samples for the most relevant parameters.

2.3. Alignment of samples

After that, samples were locally aligned against human genome 19 build 37 (hg19/GRCh37) using Bowtie2[13], an ultra-fast read mapping software (able to soft-clipping).

Bowtie2 was set up with a very sensitive alignment because all SVs detection programs use discordant mapped reads if the input samples are paired-end reads. Therefore, the more sensitive the alignment is the more number of discordant mapped reads the software will have to work with. Also, the type of alignment chosen was a local one due to the fact that most of those programs need soft-clipped reads as input (local alignment generates soft-clipped reads).

2.4. Manipulation of alignment files

Using Samtools[14] and Picardtools[15] alignment files have been manipulated to allow SV detection softwares to perform further analysis. Sorted alignments by genomic

coordinates, transformation from human readable format (uncompressed) to binary alignment format (compressed) and indexation have been performed using Samtools. Also, it has been selected only the reads that were unmapped or discordantly mapped in fastq format for a specific study (see Materials and Methods section 2.4.1). Duplicated reads were tagged with Picardtools' Markduplicates, whereas read group names and samples were added with Picardtools' AddOrReplaceReadGroups. Read duplicates have been removed, and those contained in regions of interest have been extracted.

2.5. Structural variant finders

2.5.1. Gustaf analysis

The first SV detection tool used for this analysis is called Gustaf (Generic mUlti-SpliT Aligner Finder)[16]. According to the authors, Gustaf is able to detect and classify all kinds of genomic rearrangements $\geq$ 30 bp with a sensitivity of 0.993 and positive predictive value of 0.946 in 100 bp paired-end read simulated studies. Gustaf takes unmapped and discordant mapped reads (generated with Samtools) as input and then it uses a local aligner called Stellar to detect partial alignments of a read. Finally, Gustaf generates an output table with relevant information (chromosome, position, orientation, type of SV, exact breakpoint and number of supporting reads) using the generated output by Stellar.

2.5.2. FACTERA analysis

FACTERA (Fusion And Chromosomal Translocation Enumeration and Recovery Algorithm) is the second SV detection tool used. It is a brand new software that is able to correctly detect and classify structural variants with base pair resolution in a wide range of lengths[2]. It was used in order to find SVs in paired-end exome-wide sequencing data. FACTERA gives as an output a set of files with all necessary information about the type of the detected SV (translocation, inversion or deletion), chromosome, position, strand, breakpoint, and depth of reads supporting these events.

2.5.3. Breakdancer analysis

Breakdancer[17], a widely used SV finder, was the next software used. Breakdancer has the ability to perform pooled analysis in which all samples are grouped together to reach higher sensitivity. However, due to computational limitations, this option was only

performed to detect SVs within chromosomes. Breakdancer was run for each matched normal and tumor samples to detect SVs between chromosomes (inter-chromosomal translocations) and all results were put together for final filtering. Three quality control steps were done before running the software to account for the coefficient of variation (CV) for each read group, the percentage of inter-chromosomal read pairs, and to assess for the histograms of the insert size distribution of each sample.

Breakdancer analysis is divided in two steps. Firstly, it performs a configuration file with statistics for each read group (those used for the quality control checks). Secondly, the main program uses these statistics to generate a list of putative SVs. Of note, Breakdancer does not find the exact breakpoints but theoretic ones, which can be located within the predicted boundaries with 95% confidence intervals equal to twice standard deviation insert size[18]. For this reason, it was necessary to check manually for the exact breakpoints.

2.6. Filtering process

As a filtering criterion, R[19] has been used to get a high confidence list of SV. The filtering steps performed have been: removal of all SVs found at least once in any of the normal samples, removal of insertions and deletions (CNV), establishment of a cut-off of more than four reads supporting a given SV. Finally; as mentioned in a recent study[18], it has been looked for overlapping regions of low complexity and high repetitive regions for each of the resulting candidates using Simple Repeat and RepeatMasker tracks at UCSC browser[20], respectively.

2.7. Gene annotation

The list of annotated genes for regions overlapping ±10bp the predicted breakpoints was obtained with Bedtools[21].

2.8. SV visualization

Alignment files of the samples containing any of the SVs candidates in the list were visualized with Integrative Genomics Viewer[22] (IGV). The regions of interest have been manually inspected to see the amount of concordant paired reads, the ratio between concordant read pairs and discordant read pairs supporting the given SV, and also, to find out the real breakpoints when possible.

2.9. Defining exact breakpoints

A chimera genome of 4,000bp was generated with 2,000bp before breakpoint 1 and 2,000bp after breakpoint 2 to finally obtain real sequences of the surrounding SVs regions in cases where the real breakpoint was checked. Resulting fasta files had been established as reference genomes. Each sample containing an SV had been aligned with its own set up reference genome. The process to generate the resulting alignments was the same described in Materials and Methods section 2.2, except for a global alignment method.

2.10. Final validation results

As a final step before validating a candidate SV, a blast[23] search against a sequence of about 100bp surrounding each breakpoint was made to see whether these sequences were aligned in any other region of the genome, or in contrast, they were a real fusion gene.

3. Results

3.1. Alignment of the samples

All 84 CLX samples (42 of which are normal tissue samples) have been aligned with Bowtie2 with no warnings or error messages. All samples have showed an overall alignment rate higher than 97%. Somatic structural rearrangements are not expected to be found on normal tissue samples, thus they have been used to filter and remove false positives.

3.2. Gustaf SV analysis

After alignments have been generated and properly reads have been selected (see Materials and Methods section 2.4.1), Stellar have been run in order to find all local alignments between two sequences. After several attempts to successfully generate the desired data, the only results found came from a partial sample of 2,000 reads and it took about 12 hours. Given a pool of 84 samples with most of them of more than 100 million reads, Gustaf has been rejected for such analysis. Stellar is, in this case, a software unable to work efficiently due to time restrictions and Gustaf is designed to work under Stellar output in paired-end read analysis. Gustaf could be an interesting software to look for limited regions of interest, but not to work with WES data.

3.3. Factera SV analysis

Two example sequences have been tested with FACTERA giving the expected results. Both are samples from non-small cell lung cancer (NSCLC) patients, harbouring known rearrangements involving ROS1 or ALK genes and confirmed by FISH[2]. HCC78 carries a chromosomal translocation that creates a gene fusion between SLC34A2 (chromosome 4) and ROS1 (chromosome 6). H3122 carries an inversion in chromosome 2 that produces an EML4-ALK gene fusion.

However, when CLX samples have been run under FACTERA with all the specifications correctly checked, no gene fusion has been found. Several hypotheses could explain this fact: low coverage in samples, too short reads in samples for the algorithm to work properly, the algorithm is not appropriate to analyze these samples. Notably, FACTERA has only been used in one published article by now.

3.4. Breakdancer SV analysis

Breakdancer first step has generated a configuration file from which three quality control checks have been done in order to know whether there are or not some bam formatting errors or flag issues. Firstly, the CV of the insert size for each library is computed and should be about 0.2-0.3. Secondly, it has been computed the percentage of interchromosomal reads, which normally should not be larger than 3%. However, when analyzing tumor samples, the chances to find interchromosomal translocations increase substantially. Therefore, this percentage could be larger than 3% without meaning issues with sequencing or library construction. Finally, histograms of insert size distribution for each library have also been created. Ideally, a normal distribution is expected. In contrast, a bimodal distribution is not expected. Supplementary Table 1 shows the CV for each sample, which are in the expected range. Supplementary Table 2 shows the interchromosomal read flags. Supplementary Figure 1 shows the histograms of insert size distribution for a subset of samples (all showed the same distribution shape).

When Breakdancer has been used to analyze the two mentioned example sequences, the expected gene fusions plus several *de novo* putative SVs are found. After applying filtering criteria (see Materials and Methods Section 2.5), the list of theoretical SVs is reduced substantially (see Table 3.1). However, the expected gene fusions are not filtered.

Table 3.1. Filtered putative list of SVs from example sequences H3122 and HCC78

Chr1	Pos1	Chr2	Pos2	Type	Size	# Reads	# Reads lib
chr2	29,446,006	chr2	29,448,464	ITX	-116	49	H3122|49
chr2	29,448,459	chr2	42,526,790	INV	-215	120	H3122|120
chr2	80,816,295	chr2	80,816,537	ITX	-112	9	H3122|9
chr2	141,665,342	chr2	141,665,677	ITX	-113	8	H3122|8
chr2	212,248,055	chr2	212,248,768	ITX	-112	9	H3122|9
chr2	212,543,272	chr2	212,544,021	ITX	-112	8	H3122|8
chr2	212,566,607	chr2	212,566,865	ITX	-112	8	H3122|8
chr2	212,587,034	chr2	212,587,262	ITX	-112	8	H3122|8
chr4	25,666,650	chr6	117,657,990	CTX	-215	166	HCC78|166
hr4	25,666,856	chr6	117,658,325	CTX	-215	180	HCC78|180

Note. Pos1/Pos2: estimated breakpoint position for breakpoints 1 and 2, respectively. # Reads: number of reads supporting each SV. # Reads lib: number of reads supporting each SV from given samples. ITX: Intrachromosomal translocation. INV: Inversion. CTX: Interchromosomal translocation.

After analyzing the example sequences, all CLX samples have been run under the same procedure. Regarding the output, initially 195,416 intrachromosomal SVs and 1,822 interchromosomal SVs are found by Breakdancer. After filtering SVs found in normal samples, the list of putative SVs dropped to 14,206 intrachromosomal and 547 interchromosomal. When removing insertions and deletions (only found in intrachromosomal output), 1,637 ITX and INV were found. Finally, establishing a cut-off larger than 5 supporting reads for a given SV, the final list of putative SV was dropped to 13 intrachromosomal and 9 interchromosomal SVs (see Table 3.2). None of these variants overlap with both RepeatMasker and Simple Repeat data from the UCSC browser.

Table 3.2 Filtered putative list of SVs from Colonomics Samples.

Chr1	Pos1	Chr2	Pos2	Type	Size	# Reads	# Reads_lib
chr1	158,437,081	chr1	158,462,926	INV	18795	14	M2052_T|14
chr3	138,244,667	chr3	138,311,552	INV	66231	8	H2019_T|8
chr6	43,013,359	chr6	43,084,604	INV	57329	31	P2078_T|31
chr6	47,759,047	chr6	47,773,202	ITX	12984	7	J2037_T|7
chr7	86,570,184	chr7	86,583,225	ITX	12099	8	P2009_T|8
chr8	125,578,901	chr8	128,229,741	ITX	1588774	25	Q2040_T|25
chr9	72,334,210	chr9	72,347,754	INV	13691	21	H2019_T|21
chr16	30,674,685	chr16	49,701,174	ITX	9511246	6	N2036_T|6
chr17	34,049,990	chr17	340,54,582	ITX	356	27	A2027_T|27
chr18	12,253,638	chr18	71,890,011	ITX	59635220	30	R2002_T|30
chr20	21,378,936	chr20	21,737,240	INV	239813	6	Z2084_T|6
chr20	57,565,196	chr20	59,304,571	INV	1694335	39	L2020_T|39
chrX	30,327,808	chrX	53,825,218	INV	6713456	21	A2027_T|1 E2023_T|20
chr1	156,186,627	chr20	26,189,745	CTX	-208	9	S2016_T|9
chr11	85,195,079	chr17	33,478,203	CTX	-213	11	T2093_T|4 D2079_T|5 Q2040_T|2
chr1	3,789,601	chr20	30,274,613	CTX	-211	8	P2009_T|8
chr15	45,814,571	chr16	28,099,399	CTX	-211	23	B2035_T|23
chr16	23,593,973	chr17	11,630,939	CTX	-214	8	D2079_T|8
chr17	46,654,141	chr20	41,955,507	CTX	-223	6	E2023_T|6
chr2	242,357,530	chr3	123,590,919	CTX	-208	8	S2016_T|8
chr3	138,248,358	chr13	78,768,008	CTX	-220	8	H2019_T|8
chr9	72,333,617	chr10	127,593,850	CTX	-220	22	H2019_T|22

Note. Pos1/Pos2: estimated breakpoint position for breakpoints 1 and 2, respectively. # Reads: number of reads supporting each SV. # Reads lib: number of reads supporting each SV from given samples. ITX: Intrachromosomal translocation. INV: Inversion. CTX: Interchromosomal translocation.

3.5. Gene annotation

The annotated genes for the predicted breakpoints have been obtained with Bedtools (see Table 3.3). Of note, nine of these positions are located in intergene regions. Three of these nine breakpoints have supporting reads lying within exon boundaries. This can happen if those reads are located close to the end of an exon. In such case, it could not be possible to construct the whole gene fusion sequence. For the other six predicted breakpoints, two possibilities could explain this fact. On the one hand, a given SV has aroused between a gene exon and any other region of the genome (i.e. intron, non-coding region). On the other hand, if the two breakpoints of a certain SV are located out of exon boundaries, it could be a false positive result due to a mistake in the alignment method.

Table 3.3. Genes overlapping breakpoint positions

Chromosome	Breakpoint	Gene Symbol
chr1	3,789,601	DFFB
chr1	156,186,627	PMF1-BGLAP
chr1	158,437,081	OR10K1
chr1	158,462,926	AK057554
chr2	242,357,530	FARP2
chr3	123,590,919	MYLK
chr3	138,244,667	
chr3	138,248,358	CEP70
chr3	138,311,552	
chr6	43,013,359	CUL7
chr6	43,084,604	PTK7
chr6	47,759,047	OPN5
chr6	47,773,202	
chr7	86,570,184	KIAA1324L
chr7	86,583,225	
chr8	125,578,901	MTSS1
chr8	128,229,741	CCAT1
chr9	72,333,617	
chr9	72,334,210	PTAR1
chr9	72,347,754	
chr10	127,593,850	FANK1
chr11	85,195,079	DLG2
chr13	78,768,008	RNF219-AS1
chr15	45,814,571	SLC30A4
chr16	23,593,973	NDUFAB1
chr16	28,099,399	XPO6*
chr16	30,674,685	FBRS
chr16	49,701,174	ZNF423
chr17	11,630,939	DNAH9
chr17	33,478,203	UNC45B
hr17	34,049,990	AP2B1
chr17	34,054,582	
chr17	46,654,141	HOXB3
chr18	12,253,638	CIDEA
chr18	71,890,011	CYB5A
chr20	21,378,936	NKX2-4
chr20	21,737,240	PAX1*
chr20	26,189,745	LOC284801
chr20	30,274,613	BCL2L1
chr20	41,955,507	SCARNA15*
chr20	57,565,196	NELFCD
chr20	59,304,571	LOC284857*
chrX	30,327,808	NR0B1
chrX	53,825,218	HUWE1*

* intergenic regions. Gene symbols annotated are the nearest ones

3.6. Chimeric gene fusion construction

It has been tried to construct a fusion between DLG2 and UNC45B genes, a predicted gene fusion found in three tumor samples, and another one between MTSS1 and CCAT1, this last one found in one tumor sample (see table 3.2). For this purpose, it has been necessary to correctly check with IGV for the exact breakpoints, which in such cases they have been set up at chr11:85,195,232 and chr17:33,478,275 for DLG2-UNC45B and chr8:125,580,720 and chr8:128,229,159 for MTSS1-CCAT.

A chimera genome of 4,000bp has been generated with 2,000bp from chromosome 11 (chr11:85,193,982-85,195,232) and 2,000bp from chromosome 17 (33,477,025-33,478,275). The resulting fasta file has been established as a reference genome to which the three samples containing the mentioned translocation have been aligned with.

The same procedure has been done to generate another chimera genome for the candidate gene fusion between MTSS1 and CCAT1.

3.7. Visualization of results

Using IGV, it has been possible to visualize the fusion genes with reads aligning in both sides of the breakpoints and reads aligning throughout the breakpoint. DLG2-UNC45B gene fusion has some mismatches with the quimeric reference whereas MTSS1-CCAT1 does not, as it is possible to see in Figures 3.1 and 3.2, respectively. Equivalent alignments are represented in Supplementary Figures 2 and 3 in Annex A2 for the other 2 samples with the recurrent putative gene fusion.

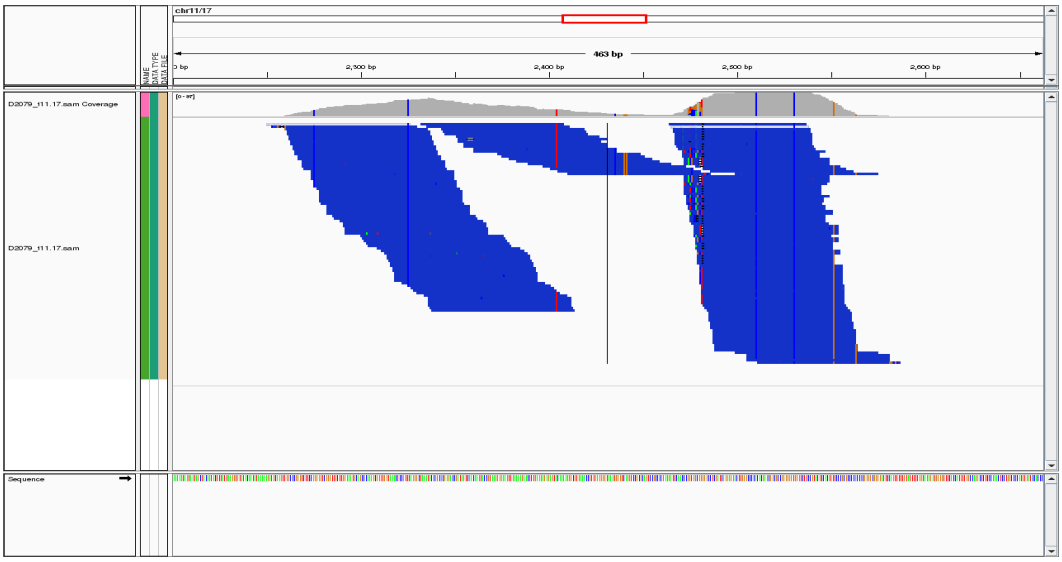


Figure 3.1. Discordant reads surrounding SV detected in sample D2079 aligned against DLG2-UNC45B gene fusion chimera construct. Mismatches are represented as colored vertical lines.

Figure 3.2. Reads surrounding SV detected in sample Q2040 aligned against MTSS1-CCAT1 gene fusion chimera construct. In this case there are no mismatches.

3.8. Final filtering of false positives

A blast search against 100bp surrounding the breakpoint of each gene fusion has been performed. To correctly validate a putative gene fusion, it is expected to find partial alignments with only its two separate sequences. However, in the case of DLG2-UNC45B gene fusion, some other matches have been found (see Appendice A3). In particular this query sequence aligns with a similarity of 100% against rRNA genes (found in chrUn_gl0000220).

3.9. Final validation of results

To validate the hypothesis that DLG2-UNC45B fusion could be a false positive, all reads from the quimeric alignment and all the reads from the two original sites on the alignment against hg19 have been extracted. All these reads (5988) have been now aligned against chrUn_gl0000220 with Bowtie2 (same procedure explained in Materials and Methods section 2.2, except for a global alignment instead of a local one).

When visualized the region that matched with DLG2-UNC45B in the blast query with IGV, a high number of mapping reads with a perfect alignment of the sequences against

chrUn_gl0000220 have been seen, as showed in Figure 3.3. Supplementary Figure 4 and 5 in Annex A2 show equivalent alignments for the other two samples.

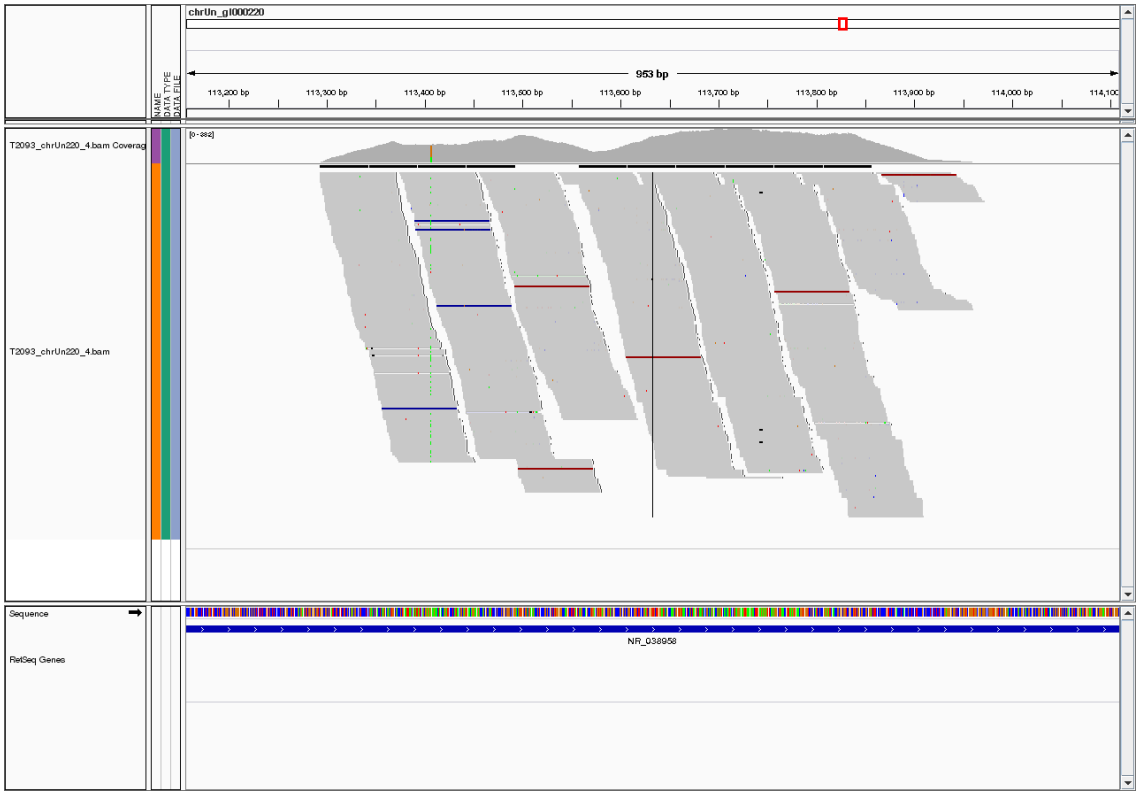


Figure 3.3. Reads surrounding SV detected in sample D2079 aligned against chrUn_gl0000220. In contrast with Figure 3.1, now there are no mismatches and the number of supporting reads is higher.

After manually finding real breakpoints for those samples where exact breakpoints could be determined, it has been found a total of 8 fusion genes, 4 of which have a short sequence surrounding the breakpoint without being sequenced (none of them reported previously), 3 rearrangements affecting the same gene and 7 rearrangements between genes and intergenic regions. In table 3.4 it is summarized this information.

Table 3.4. Summary table of all exact breakpoints found in SV calling, fusion gene formation and samples where they have been identified.

Sample	Breakpoint 1	Breakpoint 2	Fusion Gene
H2019	**Chr3:138,248,076**	**Chr13:78,768,197**	**CEP70 – RNF219-AS1**
H2019	**Chr9:72,333,392**	**Chr10:127,594,138**	**PTAR1 – FANK1**
J2037	**Chr6:47,759,662**	**Chr6:47,772,955***	**OPN5 – OPN5**
P2009	**Chr7:86,570,571**	**Chr7:86,583,017***	**KIAA1324L – KIAA1324L**
P2009	**Chr1:3,789,651**	**Chr20:30,274,613***	**DFFB – BCL2L1**
Q2040	**Chr8:125,580,720**	**Chr8:128,229,159**	**MTSS1 – CCAT1**
N2036	**Chr16:30,678,029***	**Chr16:49,701,137**	**FBRS – ZNF423**
A2027	**Chr17:34,050,541***	**Chr17:34,054,490**	**AP2B1 – AP2B1**
R2002	Chr18:71,889,951	Chr18:12,254,445	CIDEA – n.c.r
B2035	Chr15:45,814,389	Chr16:28,099,588	SCL30A4 – n.c.r.
S2016	Chr2	Chr3	FARP2 – MYLK
M2052	Chr1	Chr1	OR10K1 – n.c.r
D2079	Chr16	Chr17	NDUFAB1 – DNAH9
H2019	Chr3	Chr3	CEP70 – CEP70
H2019	Chr9	Chr9	PTAR1 – PTAR1
P2078	Chr6	Chr6	CUL7 – PTK7
Z2084	Chr20	Chr20	NKX2-4 – n.c.r
L2020	Chr20	Chr20	NELFCD – n.c.r.
E2023	ChrX	ChrX	NROB1 – n.c.r.
E2023	Chr17	Chr20	HOXB3 – n.c.r.
S2016	Chr1	Chr20	PMF1/BGLAP – LOC284801**
D2079 / Q2040 /T2093	Chr11:85,195,232	Chr17:33,478,275	DLG2 – UNC45B**

Note: In bold those fusion genes with exact known breakpoint. Unknown exact breakpoints represented only by chromosome number. * intronic positions. ** False positive fusion genes.

4. Discussion

There are evidences that gene fusion events are not enough to explain malignant transformation in haematological tumors. In contrast, it seems that certain gene fusions associated with certain types of sarcoma are enough to develop malignancy in mice. However, it is not yet known whether carcinomas have the same behavior than haematological disorders or sarcomas[10].

Depending on several aspects of the chimeric gene formation, two options are expected: a loss of functionality for one or both genes in a fusion due to high structural changes or a change in the activity in the second of the genes in a fusion because of the change in epigenetic regulation[10]. Consequently; this somatically-acquired mutations could partially explain the role of certain genes (those presents in fusions) in the onset and progression of colon cancer.

From all SVs found with exact breakpoint positions, three of them have both breakpoints within the same gene (see Table 3.4). Thus, instead of being considered gene fusions, they are considered as genes which products will differ from original ones in a variable degree. Those are OPN5, AP2B1 and KIAA1324L, found in samples J2037, A2027, and P2009, respectively.

OPN5, or neuropsin, is an opsin gene whose product is a member of G-protein-coupled receptor family, mainly expressed in neural tissues, eye and testes[24]. The modification seen in one tumoral sample from colonomics affects exon 3. Nevertheless, due to its lack of expression in colon, it seems that OPN5 variant found is not a good candidate to explain onset or progression of this particular cancer.

In contrast, AP2B1 is moderately expressed in both colon and rectum. It is involved in protein transport and among its related pathways are signaling by fibroblast growth factor receptor (FGFR). The variant found in colonomics sample, however, is only affected in the last exon, so little changes in gene product should be expected. Interestingly, this variant approaches the end of the mentioned gene to the beginning of RASL10B, a RAS-like protein with GTPase activity that could change their expression pattern.

KIAA1324L, also known as EIG121L, is involved in epithelial differentiation in embryonic development[25]. In addition, it is expressed in colon and rectal tissues[26] but it is not clear its function, so it is difficult to know whether this variant could make some effect in the disease. The variant found here, is composed by the first four exons, just as spliced variant found in testes.

Regarding the other variants (proper gene fusions) there are several different scenarios.

MTSS1-CCAT1 fusion is an interesting rearrangement due to some aspects. MTSS1 (metastasis suppressor 1) has been found overexpressed in colorectal cancer and it is positively correlated with low 5-year survival rates[27]. CCAT1 (colon cancer associated transcript 1) is a non-protein coding gene implicated in the transcriptional regulation of MYC[28] and its expression is upregulated in colorectal cancer tissues[29]. The fusion found here is under regulation of CCAT1, so it should be expected that MTSS1 would be then overexpressed, and hence, this fusion could become a biomarker for predicting bad prognosis in CRC patients.

In the case of PTAR1-FANK1, the fusion occurs between the last exon of PTAR1 and the first intron of FANK1, but given the fact that they are head to head genes, this fusion is going to generate two different products. The first one will be PTAR1 joined to an unpredictable sequence composed by part of the first intron and the first exon of FANK1. PTAR1 has prenyltransferase activity, which is known to be needed for oncogenic proteins such as Ras[30], so it is expected that this fusion will have this activity. Furthermore, it will be under regulation of PTAR1. The product of this fusion gene will be regulated by FANK1, which is a gene involved in apoptosis[31], and will be composed by the first exon of FANK1 and part of the last exon of PTAR1 joined to FANK1 intron 1. Taken together, it seems that cells producing these variants will partially loss apoptotic properties.

DFFB-BCL2L1 fusion involves two genes with apoptotic properties which are moderately expressed in colon and rectum tissues. BCL2L1 gene have both actions pro-apoptotic and anti-apoptotic depending on the splice variant expressed[32]. Again, as in the case of PTAR1-FANK1, these are head to head genes, so the gene fusion will

generate two gene products, each one with the first gene properly coded and the second one coded in the wrong direction, giving an unexpected protein product.

CEP70-RNF219AS1 is a fusion which joints a centrosomal protein coding gene (CEP70), which has mitotic functions, with an antisense RNA gene. The fusion will modify CEP70 protein but it is difficult to predict its consequences.

With regards to FBRS-ZNF423 fusion, again are involved head to head genes. FBRS codes for fibrosin, a limphokine that induces fibroblast proliferation[33]. On the other hand, ZNF423 is a zinc finger protein that works as a transcription factor[34]. The products of this fusion will be partially translated genes linked with an unknown sequence.

Regarding the other fusion genes, those where getting the exact breakpoint was not possible, 3 different fusion genes have been found, all of them between head to head genes. Therefore, the gene products in all cases will be part of one gene properly coded and part of another one coded in the wrong direction, with unknown result.

FARP2-MYLK is composed of a gene related with Ras signaling pathway and involved in cytoskeleton modelling (FARP2)[35] and MYLK, which codes for the light chain kinase of myosin[36].

CUL7-PTK7 is a fusion gene which will code for both a component of an ubiquitin-protein ligase complex[37] and an inactive tyrosine kinase[38].

In contrast to the last fusions, NDUFAB1-DNAH9 will have only one gene product because DNAH9 is not expressed in colon tissue. Thus, the gene product will joint an unknown sequence from DNAH9 with the first 3 out of 4 exons of NDUFAB1, which product is a subcomplex of NADH dehydrogenase ubiquinone 1[39].

5. Limitations

This study has few limitations. Firstly, due to the amount of DNA needed to perform this sequencing technologies and the fact that cancer tissues are a combination of cells with different number of mutations, the sequencing of the whole exome is in fact the sequence of a mixture of DNA from different cells. This could decrease the power of detection of somatic structural rearrangements.

Secondly, Breakdancer output has a tendency to generate a large list of false positives SVs due to the nature of the genome, where there are sites of low complexity or high repetitivity. For this reason, it is necessary to perform several non-standardized filtering steps where it is possible to remove some true positive values. In addition, breakpoints generated by Breakdancer are not real but approximated, which makes not viable any automatic analysis from this point on. With regards of the amount of supporting reads for a particular SV, it tends to give a lower value than what it is expected because all reads surrounding a real breakpoint plus their paired reads are not considered SVs supporting reads by this algorithm. In Figure 5.1 is shown an example of this.

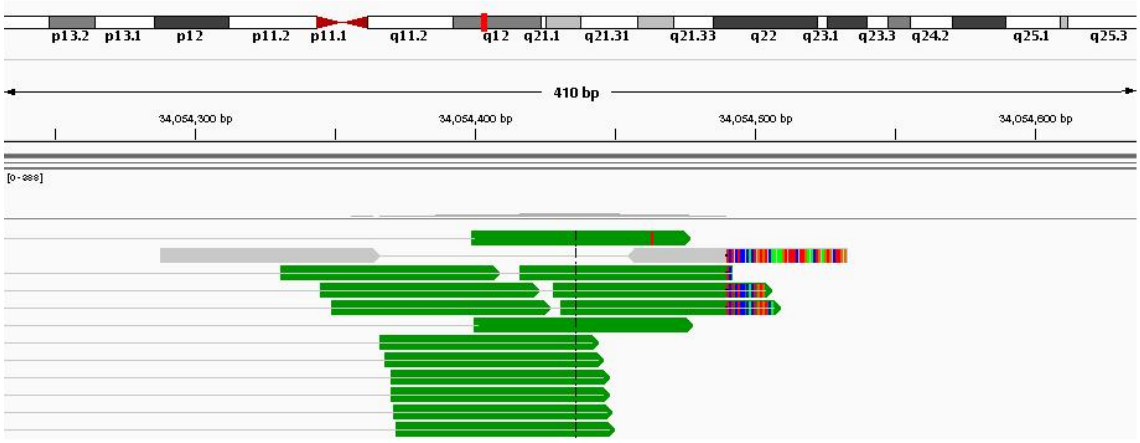


Figure 5.1: Alignment of reads from sample A2027T in 3'UTR of AP2B1 gene in chr17.q12. Green reads are considered discordant whereas grey reads are not. Rainbow sites are mismatches indicating the exact breakpoint of this SV. Breakdancer does not consider paired reads in grey as supporting reads for this SV because they are located in a proper distance between them. However, it has been demonstrated that rainbow-coloured part in rightmost grey read is in fact part of an ITX. Therefore, these pair of reads should be considered as supporting reads for this SV.

Finally, several other programs (Gustaf, FACTERA, Breakmer, SVdetect), some of which are being developed right now, seem to find real breakpoints from soft-clipped read ends. Even though having a large list of SVs, it would be possible to follow further analysis without spending so much time trying to elucidate their breakpoints. However, these softwares are not working properly yet for whole exome sequencing analysis. They are working fine for particular genome data (i.e. type of tumor data) and for longer reads (100bp or more). Also, their algorithms are really good at finding SVs in a very limited size of sequences (GUSTAF).

6. Conclusions

It has been performed a pipeline analysis to get a high probability list of 22 SVs from a set of whole exome sequencing samples. Once set up, this method could be used in the future to perform similar analysis with lots of samples and finally get quick results.

20 out of 22 SVs have been considered true positives, 8 of which are gene fusions. Although none of them are recurrent chromosomal rearrangements, it should be necessary to validate experimentally these results and further analyze whether these gene fusions are able to generate protein products or not.

MTSS1-CCAT1 could be an interesting gene fusion to be studied in the future due to their properties, the novel structure and regulation it has taken after the fusion and their expression in colorectal cancer.

It has also been hypothesized that gene fusion DLG2-UNC45B found in CLX samples, as well as in TCGA Glioblastoma samples[40] and in metastatic cervical carcinoma samples[41], is indeed a false positive generated by Breakdancer due to a high similarity with an rRNA gene containing a highly repetitive sequence. This discovery highlights the necessity of align samples against all the chromosomes sequences (including random and unmapped sequences) every time an Structural Variant finding analysis is going to be performed.

References

1. Jemal A, *et al*. Global cancer statistics. CA Cancer J Clin 2011;61:69-90.

2.- Newman AM, *et a*l. FACTERA: a practical method for the discovery of genomic rearrangements at breakpoint resolution. Bioinformatics. 2014;30(23):3390-3393.

3.- Raphael BJ Chapter 6: Structural Variation and Medical Genomics. PLoS Comput Biol 2013;8(12): e1002821.

4.- Quinlan AR, Hall IM. Characterizing complex structural variation in germline and somatic genomes. Trends Genet. 2012;28:43–53.

5.- Mijuskovic M, *et al.* A streamlined method for detecting structural variants in cancer genomes by short read paired-end sequencing. PLoS One 2012;7(10):e48314.

6. Haggar FA, Boushey RP. Colorectal Cancer Epidemiology: Incidence, Mortality, Survival, and Risk Factors. Clinics in Colon and Rectal Surgery. 2009;22(4):191-197

7. Labianca R. *et al.* Colorectal cancer: screening. Ann Oncol 2005;16(Suppl 2): ii127–ii132

8.- The Cancer Genome Atlas Network. Comprehensivemolecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-337

9. Ardini E. *et al*. The TPM3-NTRK1 rearrangement is a recurring event in colorectal carcinoma and is associated with tumor sensitivity to TRKA kinase inhibition. Mol Oncol 2014;8(8):1495-1507.

10. Mitelman F, Johansson B, and Mertens F. The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer. 2007;7(4):233-245.

11.- Sanz-Pamplona R, *et al.* Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer. Clin Cancer Res 2015, doi: 10.1158/1078-0432.CCR-15-0159

12.- http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

13.- Langmead, B. Salzberg, S. Fast gapped-read alignment with Bowtie2. Nature Methods. 2012, 9:357-359

14.- Li, H. et al. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 2009;25, 2078-2079.

15.- http://picard.sourceforge.net.

16.- Trappe K, *et al*. Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics 2014;30(23):3484-3490.

17.- Chen K, *et al*. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009;6(9):677-681.

18.- Xian F, *et al*. BreakDancer – Identification of Genomic Structural Variation from Paired-End Read Mapping. Curr Protoc Bioinformatics. 2014; doi: 10.1002/0471250953.bi1506s45

19.- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2013, Vienna, Austria. URL http://www.R-project.org/

20.- https://genome.ucsc.edu/

21.- Qinlan, AR. Hall, IM. Bedtools: A flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26(6):841-842.

22.- Robinson, JT. et al. Integrative Genomics Viewer. Nature Biotechnology 2011;29:24-26.

23.- Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21097/

24.- Tarttelin, Emma E, *et al.* Neuropsin (Opn5): a novel opsin identified in mammalian neural tissue. FEBS Letters 2003;554(3):410-416.

25.- Araki T, Kusakabe M, Nishida E. A Transmembrane Protein EIG121L Is Required for Epidermal Differentiation during Early Embryonic Development. The Journal of Biological Chemistry 2011;286(8):6760-6768.

26.- http://www.proteinatlas.org/ENSG00000164659-KIAA1324L/tissue

27.- Wang D. MTSS1 overexpression correlates with poor prognosis in colorectal cancer. J Gastrointest Surg 2011 Jul;15(7):1205-1212.

28.- Xiang J-F. *et al*. Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Research 2014;24:513-531.

29.- Zhenyu Y, *et al.* Expression of lncRNA-CCAT1, E-cadherin and N-cadherin in colorectal cancer and its clinical significance. Int J Clin Exp Med 2015;8(3):3707-3715.

30.- Ochocki JD, Distefano MD. Prenyltransferase Inhibitors: Treating Human Ailments from Cancer to Parasitic Infections. Med Chem Comm. 2013;4(3):476-492.

31.- Wang H. et al. Fank1 interacts with Jab1 and regulates cell apoptosis via the AP-1 pathway. Cell Mol Life Sci 2011;68(12):2129-2139.

32.- Sillars-Hardebol AH, et al. BCL2L1 has a functional role in colorectal cancer and its protein expression is associated with chromosome 20q gain. J Pathol 2012;226(3):442-450.

33.- Prakash S, Robbin PW. Cloning and analysis of the cDNA for human fibrosin, a novel fibrogenic lymphokine. DNA Cell Biol 1998;17(10):879-884.

34.- Turner J, Crossley M. Mammalian Kruppel-like transcription factors: more than just a pretty finger. Trends Biochem Sci 1999;24:236–240.

35.- Kubo T. A novel FERM domain including guanine nucleotide exchange factor is involved in Rac signaling and regulates neurite remodeling. J Neurosci 2002;22(19):8504-8513.

36.- Watterson DM, *et al.* Analysis of the kinase-related protein gene found at human chromosome 3q21 in a multi-gene cluster: organization, expression, alternative splicing, and polymorphic marker. J Cell Biochem 1999;75(3):481-491.

37.- Fu J. *et al.* Ubiquitin ligase cullin 7 induces epithelial-mesenchymal transition in human choriocarcinoma cells. J Biol Chem 2010;285(14):10870-10879.

38.- Shin WS, *et al.* Soluble PTK7 inhibits tube formation, migration, and invasion of endothelial cells and angiogenesis. Biochem Biophys Res Commun 2008;371(4):793-798.

39.- Bartoloni L. *et al.* Axonemal beta heavy chain dynein DNAH9: cDNA sequence, genomic structure, and investigation of its role in primary ciliary dyskinesia. Genomics 2001;72(1):21-33.

40.- Cameron W. Brennan, *et al*. The Somatic Genomic Landscape of Glioblastoma. Cell 2013;155:462-477.

41.- Winnie S. Liang, *et al*. Simultaneous Characterization of Somatic Events and HPV-18 Integration in a Metastatic Cervical Carcinoma Patient Using DNA and RNA Sequencing. Int J Gynecol Cancer 2014; DOI: 10.1097/IGC.0000000000000049

Appendices

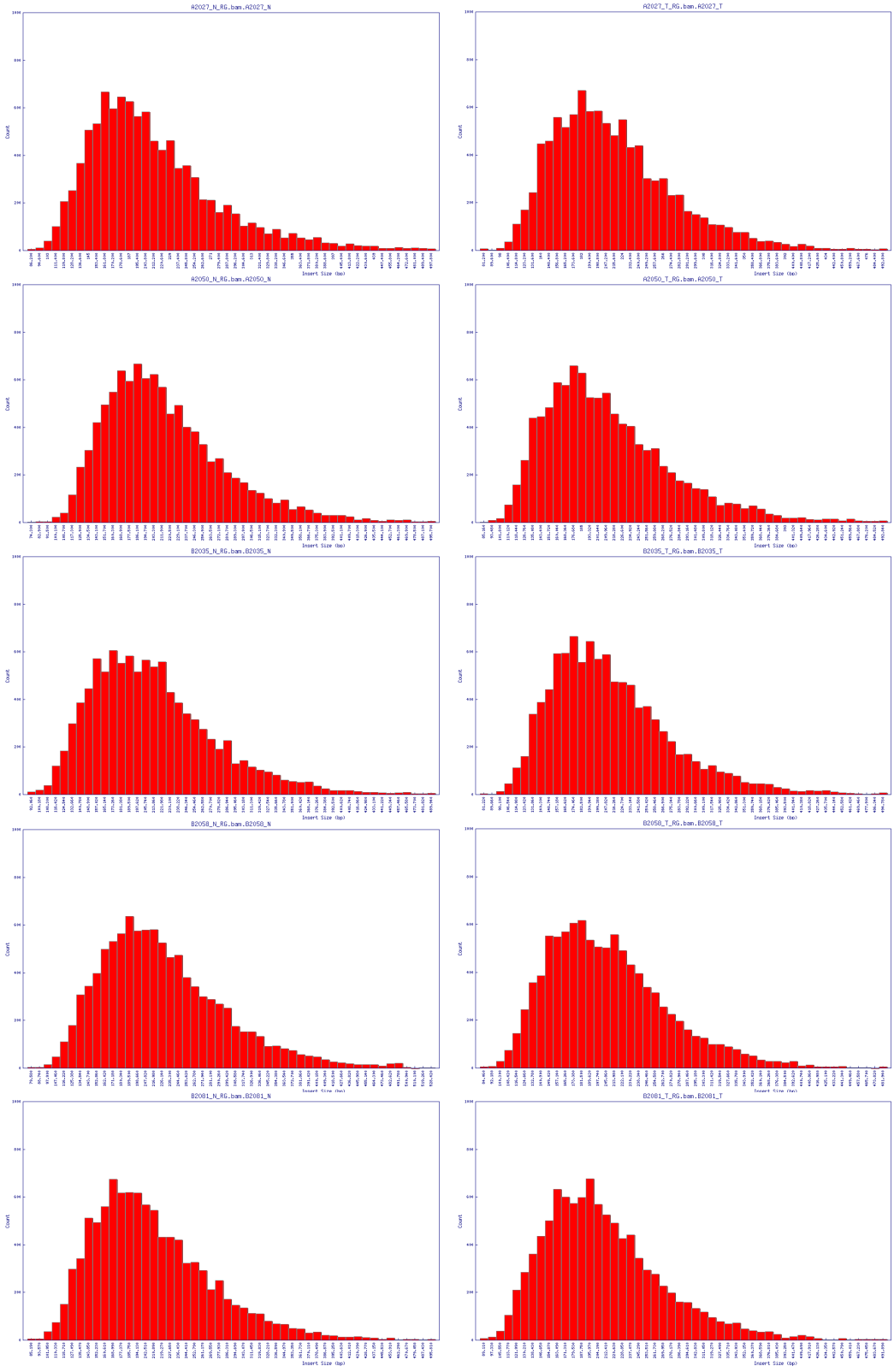
A1. Breakdancer quality controls

Supplementary Table 1.
Coefficient of Variation.

	Normal Sample	Tumoral Sample
A2027	0.316	0.292
A2050	0.295	0.306
B2035	0.289	0.293
B2058	0.306	0.286
B2081	0.292	0.284
C2067	0.309	0.298
D2079	0.283	0.305
E2023	0.285	0.323
E2046	0.283	0.283
E2069	0.293	0.300
F2008	0.298	0.291
F2077	0.300	0.274
H2019	0.325	0.300
H2042	0.284	0.302
H2065	0.292	0.306
J2014	0.291	0.292
J2037	0.286	0.287
K2068	0.317	0.285
L2020	0.281	0.283
L2089	0.293	0.293
M2052	0.292	0.314
N2013	0.294	0.297
N2036	0.283	0.294
P2009	0.284	0.311
P2032	0.285	0.314
P2078	0.319	0.296
Q2040	0.287	0.295
R2002	0.275	0.302
R2025	0.279	0.298
R2048	0.289	0.291
S2016	0.296	0.316
S2062	0.325	0.319
T2047	0.304	0.308
T2093	0.279	0.290
V2041	0.279	0.286
W2026	0.303	0.285
X2034	0.290	0.306
X2057	0.284	0.291
X2080	0.311	0.297
Y2076	0.281	0.286
Z2038	0.321	0.307
Z2084	0.299	0.316

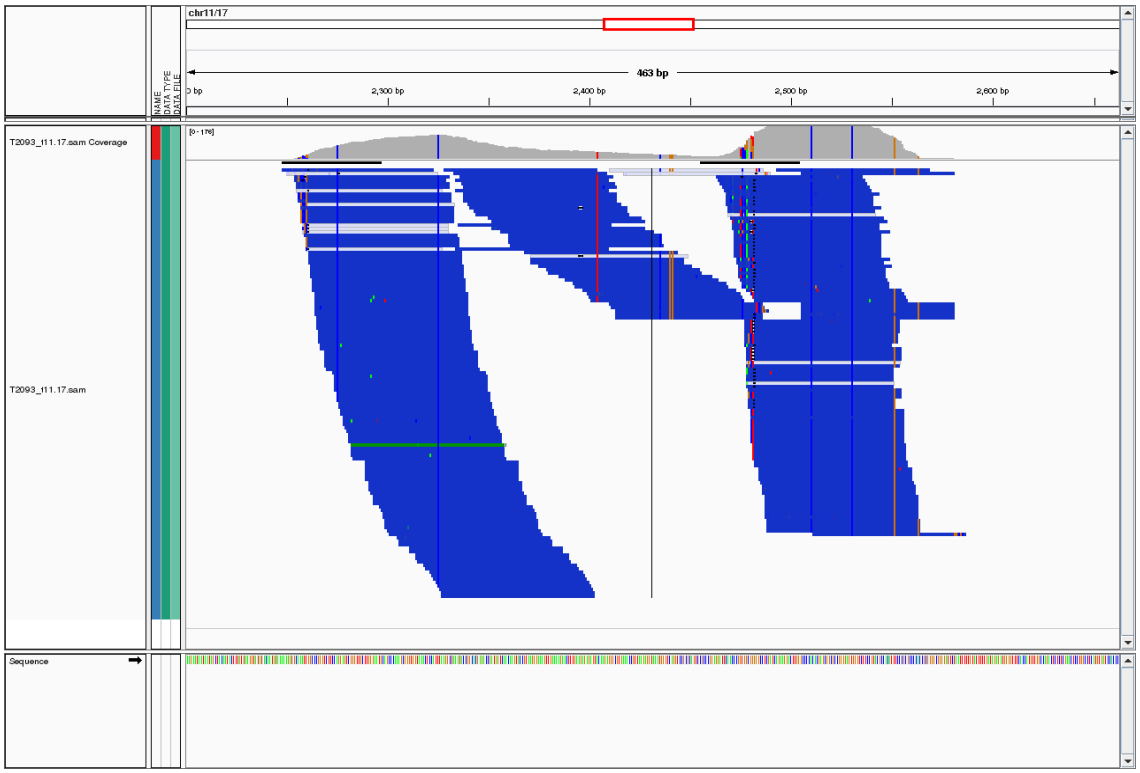
Supplementary Table 2.
Percentage of Intrachromosomal reads pairs.

	Normal Sample	Tumoral Sample
A2027	7.60%	11.52%
A2050	5.69%	1.62%
B2035	7.04%	11.00%
B2058	8.44%	11.44%
B2081	0.95%	2.04%
C2067	1.32%	1.65%
D2079	8.03%	7.11%
E2023	2.16%	3.52%
E2046	7.61%	8.51%
E2069	1.03%	1.24%
F2008	0.98%	9.70%
F2077	1.56%	2.58%
H2019	6.51%	10.40%
H2042	2.16%	16.86%
H2065	0.87%	3.56%
J2014	6.63%	10.68%
J2037	0.77%	1.04%
K2068	11.26%	10.02%
L2020	9.12%	1.73%
L2089	4.91%	4.00%
M2052	3.18%	2.41%
N2013	0.59%	6.78%
N2036	1.61%	8.14%
P2009	7.92%	13.39%
P2032	0.97%	1.37%
P2078	6.55%	7.47%
Q2040	12.85%	8.83%
R2002	6.07%	6.33%
R2025	5.50%	0.97%
R2048	1.31%	1.04%
S2016	7.84%	14.70%
S2062	2.07%	2.94%
T2047	7.37%	1.85%
T2093	6.52%	6.80%
V2041	6.10%	3.09%
W2026	9.28%	7.65%
X2034	1.73%	12.45%
X2057	5.86%	1.09%
X2080	10.04%	14.35%
Y2076	1.23%	2.19%
Z2038	7.81%	7.99%
Z2084	2.96%	2.42%

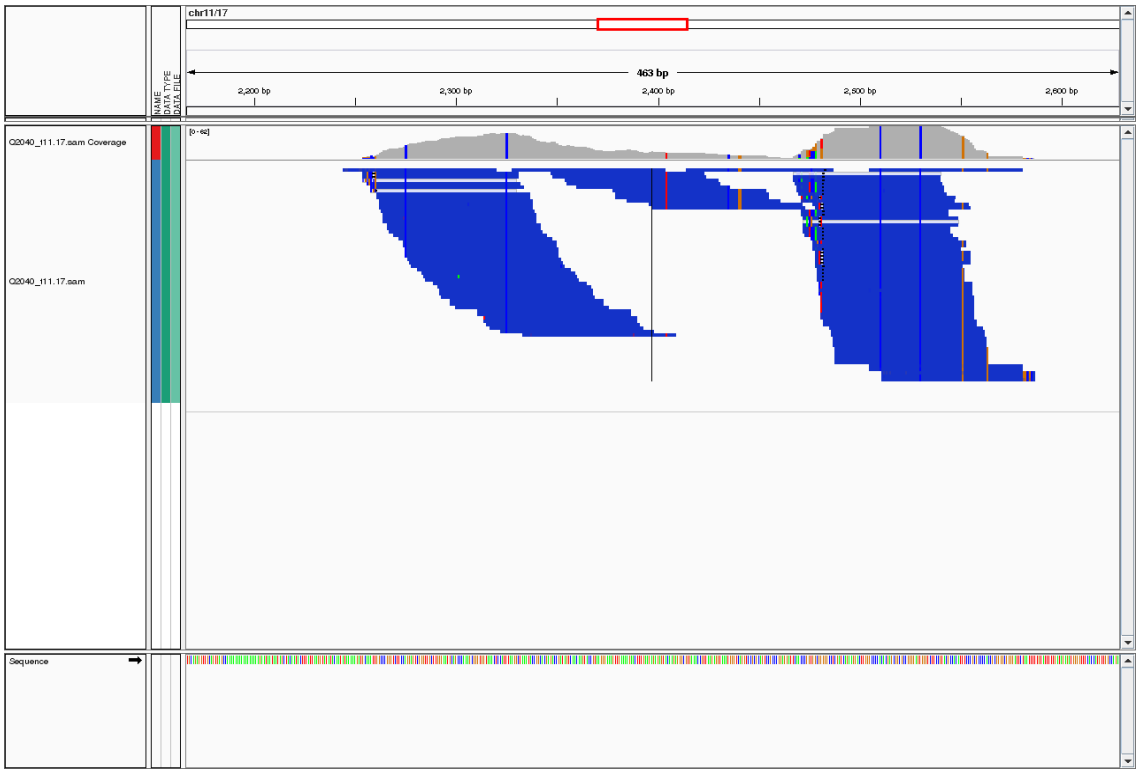


Supplementary Figure 1. Insert size distribution histograms for the five first pair of samples (A2027 to B2081), the others are very similar. Column 1: Normal Samples, column 2: Tumor Samples

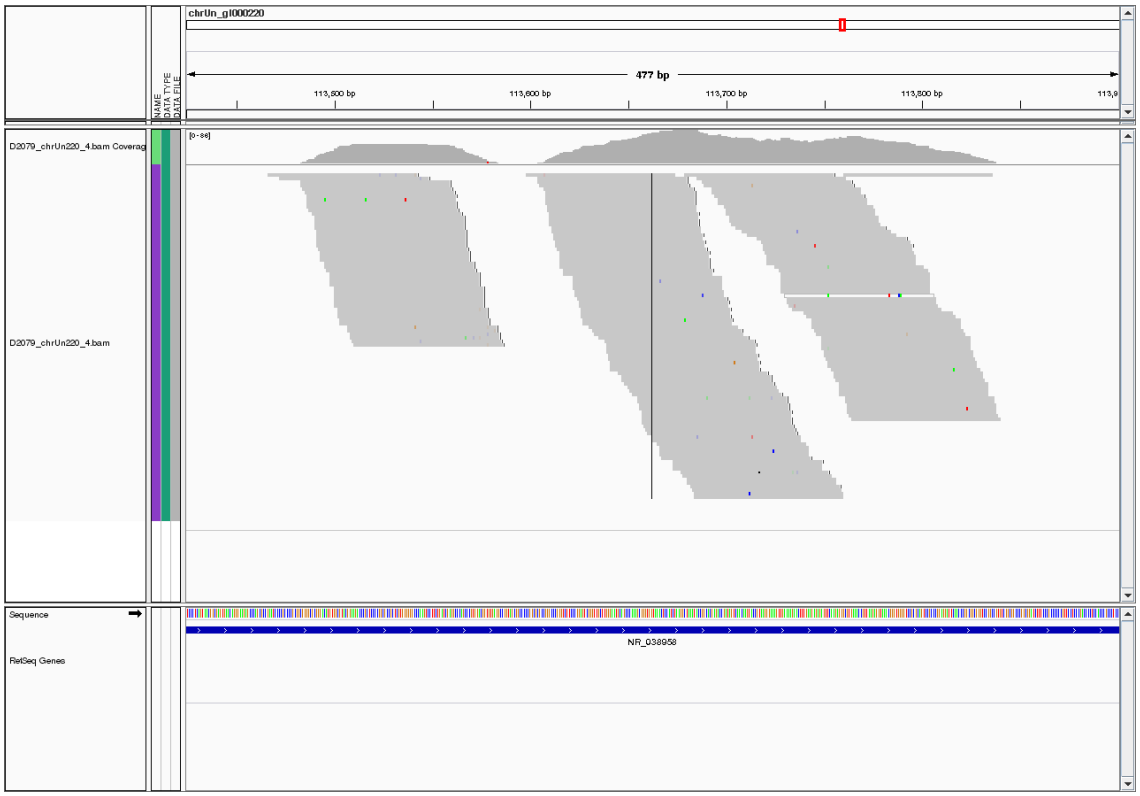
A2. Integrative Genomic Viewer supplementary alignments



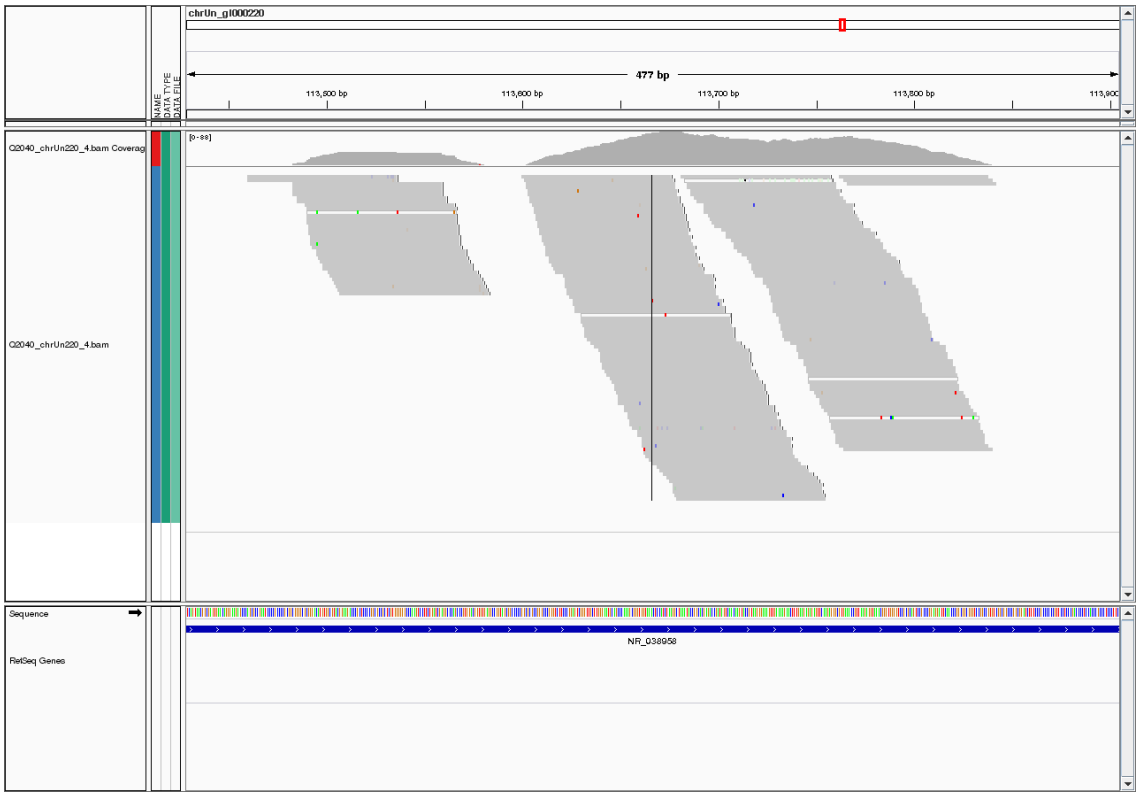
Supplementary Figure 2. Discordant reads surrounding SV detected in sample T2093 aligned against DLG2-UNC45B gene fusion chimera construct. Mismatches are represented as colored vertical lines.



Supplementary Figure 3. Discordant reads surrounding SV detected in sample Q2040 aligned against DLG2-UNC45B gene fusion chimera construct. Mismatches are represented as colored vertical lines.



Supplementary Figure 4. Reads surrounding SV detected in sample T2093 aligned against chrUn_gl0000220.



Supplementary Figure 5. Reads surrounding SV detected in sample Q2040 aligned against chrUn_gl0000220.

A3. BLAST report for short chimeric sequence carrying DLG2-UNC45B gene fusion

BLAST ®

Basic Local Alignment Search Tool

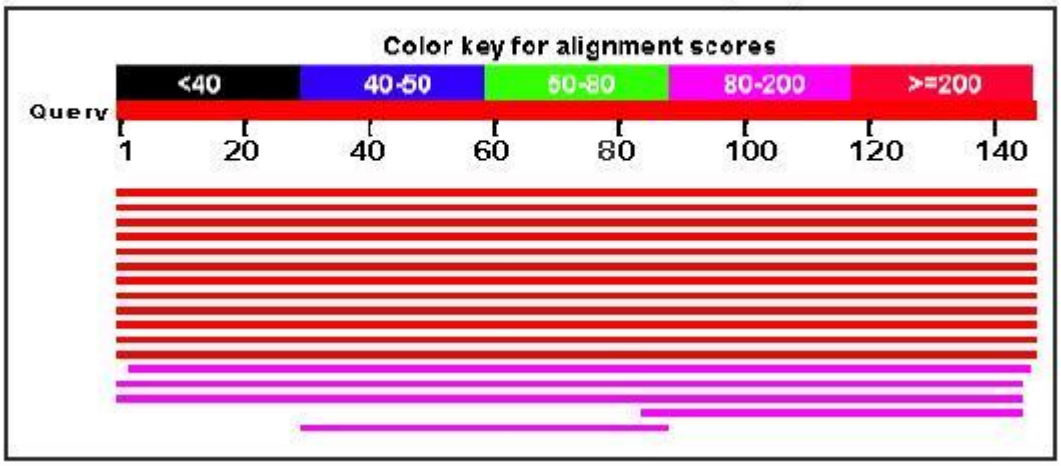
NCBI/ BLAST/ blastn suite/ **Formatting Results - ZD22HFEB015**
Formatting options
Download
Blast report description

chr11_17_DLG2-UNC45B

RID	ZD22HFEB015 (Expires on 09-15 16:41 pm)		
Query ID	lcl|Query_91361	**Database Name**	nr
Description	chr11_17_DLG2-UNC45B	**Description**	Nucleotide collection (nt)
Molecule type	nucleic acid	**Program**	BLASTN 2.2.32+
Query Length	145		

Graphic Summary

Distribution of 20 Blast Hits on the Query Sequence



Descriptions

Sequences producing significant alignments:

Description	Max score	Query cover	E value	Ident	Accession
Chain 5, Structure Of The H. Sapiens 60s Rrna	268	100%	8e-70	100%	3J3F_5
Homo sapiens RNA, 45S pre-ribosomal 5 (RNA45S5), ribosomal RNA	268	100%	8e-70	100%	NR_046235.1
Homo sapiens FOSMID clone ABC12-46987300E12 from chromosome unknown, complete sequence	268	100%	8e-70	100%	AC231275.2
Human DNA sequence from clone CH507-528H12 on chromosome 21, complete sequence	268	100%	8e-70	100%	FP236383.15
Homo sapiens RNA, 28S ribosomal 5 (RNA28S5), ribosomal RNA	268	100%	8e-70	100%	NR_003287.2
Human DNA sequence from clone CH507-146P16 on chromosome 21, complete sequence	268	100%	8e-70	100%	CT476837.18
Human DNA sequence from clone RP11-164K15 on chromosome 22, complete sequence	268	100%	8e-70	100%	AL353644.34
Human DNA sequence from clone RP11-337M7, complete sequence	268	100%	8e-70	100%	AL592188.60
Human ribosomal DNA complete repeating unit	268	100%	8e-70	100%	U13369.1
Human 28S ribosomal RNA gene	268	100%	8e-70	100%	M11167.1
Homo sapiens discs, large homolog 2 (Drosophila) (DLG2), RefSeqGene on chromosome 11	246	100%	4e-63	97%	NG_021375.1
Homo sapiens genomic DNA, chromosome 11q, clone:RP11-482L11, complete sequence	246	100%	4e-63	97%	AP003035.2
Homo sapiens BAC clone RP11-288M20 from 2, complete sequence	148	97%	1e-33	86%	AC068542.5
Homo sapiens chromosome 19 clone CTD-2017D11, complete sequence	137	98%	2e-30	84%	AC092279.2
Homo sapiens chromosome 19 clone LLNLR-242A12, complete sequence	137	98%	2e-30	84%	AC011534.3
{cellular sequence adjacent to integrated hepatitis B virus DNA, clone 13T-198} [human, hepatocellular carcinomas, Genomic, 198 nt]	111	41%	1e-22	100%	S76050.1
Human carcinoma cell-derived Alu RNA transcript, clone ALU496	100	40%	3e-19	98%	M87943.1

A4. Scripts used for the analysis

1. Breakdancer filtering R.

```
setwd("/home/46962313Q/extra/align/breakdancer/prova")

# loading interchromosomal table
sv<- read.table("colonomics_CTX2", sep="\t")

# loading intrachromosomal table
sv2<- read.table("colonomics_full_bd", sep="\t")
col<- c("Chr1", "Pos1", "Orientation1", "Chr2", "Pos2", "Orientation2", "Type", "Size",
"Score", "num_Reads", "num_Reads_lib")
colnames(sv)<- col
colnames(sv2)<- col

sv$Chr1<- as.character(sv$Chr1)
sv$Chr2<- as.character(sv$Chr2)
sv$num_Reads_lib<- as.character(sv$num_Reads_lib)
sv$num_Reads_lib<- gsub("/home/46962313Q/extra/align//", " ", sv$num_Reads_lib)

sv2$Chr1<- as.character(sv2$Chr1)
sv2$Chr2<- as.character(sv2$Chr2)
sv2$num_Reads_lib<- as.character(sv2$num_Reads_lib)
sv2$num_Reads_lib<- gsub("/home/46962313Q/extra/align//", " ",
sv2$num_Reads_lib)

sv_sorted<- sv[order(sv$Chr1, sv$Pos1, sv$Chr2, sv$Pos2),]
sv_filt<- data.frame(sv_sorted[1,])

# detecting common CTX in different samples
i=1
while (i <(nrow(sv_sorted))) {
  for (j in i+1:nrow(sv_sorted)) {
    sv_filt<- rbind(sv_filt, sv_sorted[j,])
    if (sv_sorted$Chr1[i] == sv_sorted$Chr1[j] & sv_sorted$Chr2[i] ==
sv_sorted$Chr2[j] & sv_sorted$Pos1[i]-100 < sv_sorted$Pos1[j] &
sv_sorted$Pos1[i]+100 > sv_sorted$Pos1[j] & sv_sorted$Pos2[i]-100 <
sv_sorted$Pos2[j] & sv_sorted$Pos2[i]+100 > sv_sorted$Pos2[j]) {
      sv_filt[i,10]<- sv_filt[i,10] + sv_sorted[j,10]
      sv_filt[i,11]<- paste(sv_filt[i,11], sv_sorted[j,11], sep=" ")
      sv_filt[j,11]<- paste(sv_filt[j,11], "duplicated", sep=" ")
      } else {
        i<- j
        break
    }
  }
}

# subset for removing duplicated CTX
```

```
sv_filt<- subset(sv_filt, !grepl("*duplicated", sv_filt$num_Reads_lib))

# merging both Breakdancer outputs
sv_full<- rbind(sv2, sv_filt)

# applying filtering criteria
select<- c("INV", "CTX", "ITX")
sv_balanced<- sv_full[sv_full$Type %in% select,]
# dim(sv_balanced)
sv_highsup<- subset(sv_balanced, num_Reads>4)
# dim(sv_highsup)
sv_somatic<- subset(sv_highsup, !grepl("*_N", sv_highsup$num_Reads_lib))
# dim(sv_somatic)
sv_final<- sv_somatic[c(-1, -3, -4, -5, -9, -10, -12, -13, -15, -16, -18, -19, -20, -21, -22, -
23, -24, -25, -26, -27, -29, -32, -35),]
# dim(sv_somatic)

# final filtered output
write.table(sv_final, "sv_candidates.txt", sep="\t", quote=FALSE, col.names=TRUE)

annot1<- data.frame(sv_final$Chr1, sv_final$Pos1, sv_final$Pos1+1)
colnames(annot1)<- c("Chr", "Pos_s", "Pos_e")
annot2<- data.frame(sv_final$Chr2, sv_final$Pos2, sv_final$Pos2+1)
colnames(annot2)<- c("Chr", "Pos_s", "Pos_e")
annot<- as.data.frame(rbind(annot1, annot2))
```

2. Alignment pipeline for Illumina paired-end reads.

```
#!/bin/bash
#$ -cwd
#$ -o ./
#$ -j y
#$ -S /bin/bash

EXOMES="/home/46962313Q/exomes/fastq_clx/"
ALIGN="/home/46962313Q/chromosomes/"
PICARD="/share/apps/java/jre1.6.0_21/bin/java -jar /share/apps/picard/picard.jar"
PERL="/share/apps/Perl/bin/perl"
HG19="/home/46962313Q/reference/bowtie/UCSC_hg19b/hg19"
BOWTIE2="/share/apps/bowtie2/bowtie2"
SAMTOOLS="/share/apps/samtools/samtools"
BREAK="/home/46962313Q/breakdancer/perl/bam2cfg.pl"
DANCER="/home/46962313Q/breakdancer/build/bin/breakdancer-max"
PATH=$PATH:/home/46962313Q/bin:/share/apps/Perl/bin

# Tumoral samples only
cd $EXOMES

samples=`ls *T_1.fastq*`
samples=${samples//_1.fastq/}

CORES=$((`grep -c ^processor /proc/cpuinfo` -1))

# test
#samples=(${samples})
#samples=${samples[0]}
echo $samples

cd $ALIGN

for i in $samples
do

# Bowtie2 local alignment / very sensitive
$BOWTIE2 --very-sensitive-local -p $CORES -x $HG19 -1 $EXOMES/${i}_1.fastq -2
$EXOMES/${i}_2.fastq | $SAMTOOLS view -uS - | $SAMTOOLS sort -m 1000000000 -
$ALIGN/${i}

# Picardtools MarkDuplicates for taggin duplicated reads
$PICARD MarkDuplicates INPUT=$ALIGN/${i}.bam OUTPUT=$ALIGN/${i}_dup.bam
METRICS_FILE=$ALIGN/${i}.txt

# Removal of duplicated and or unmapped reads
$SAMTOOLS view -u -h -F1036 $ALIGN/${i}_dup.bam > $ALIGN/${i}_fil.bam

# Addition of Read Groups
$PICARD AddOrReplaceReadGroups I=$ALIGN/${i}_fil.bam O=$ALIGN/${i}_RG.bam
ID=$ALIGN/${i} SM=1 LB=$ALIGN/${i} PL=illumina PU=1

# Index Bam file
$SAMTOOLS index $ALIGN/${i}_RG.bam
```

```
done

# Normal samples this time
cd $EXOMES

samples=`ls *N_1.fastq*`
samples=${samples//_1.fastq/}

CORES=$((`grep -c ^processor /proc/cpuinfo` -1))

echo $samples

cd $ALIGN

for i in $samples
do

# Bowtie2 local alignment / very sensitive
$BOWTIE2 --very-sensitive-local -p $CORES -x $HG19 -1 $EXOMES/${i}_1.fastq -2
$EXOMES/${i}_2.fastq | $SAMTOOLS view -uS - | $SAMTOOLS sort -m 1000000000 -
$ALIGN/${i}

# Picardtools MarkDuplicates for taggin duplicated reads
$PICARD MarkDuplicates INPUT=$ALIGN/${i}.bam OUTPUT=$ALIGN/${i}_dup.bam
METRICS_FILE=$ALIGN/${i}.txt

# Removal of duplicated and or unmapped reads
$SAMTOOLS view -u -h -F1036 $ALIGN/${i}_dup.bam > $ALIGN/${i}_fil.bam

# Addition of Read Groups
$PICARD AddOrReplaceReadGroups I=$ALIGN/${i}_fil.bam O=$ALIGN/${i}_RG.bam
ID=$ALIGN/${i} SM=1 LB=$ALIGN/${i} PL=illumina PU=1

# Index Bam file
$SAMTOOLS index $ALIGN/${i}_RG.bam

done
```

3. Breakdancer script

```
#!/bin/bash
#$ -cwd
#$ -o ./
#$ -j y
#$ -S /bin/bash

ALIGN="/home/46962313Q/extra/align/"
BREAK="/home/46962313Q/breakdancer/perl/bam2cfg.pl"
DANCER="/home/46962313Q/breakdancer/build/bin/breakdancer-max"
PERL="/share/apps/Perl/bin/perl"
PATH=$PATH:/home/46962313Q/bin:/share/apps/Perl/bin

cd $ALIGN

CORES=$((`grep -c ^processor /proc/cpuinfo` -1))

# SV calling for interchromosomal events
samples=`ls *T_RG.bam*`
samples=${samples//T_RG.bam/}

for i in $samples
do

$BREAK -g -h ${i}T_RG.bam ${i}N_RG.bam > config_${i}
$DANCER -t -a config_${i} > col_CTX_${i}

done


# SV calling for intrachromosomal events
$BREAK -g -h `ls *RG.bam` > colonomics_config

$DANCER -o chrX -a colonomics_config > col_chrX_bd
$DANCER -o chrY -a colonomics_config > col_chrY_bd

for i in {1..22}
do

$DANCER -o chr${i} -a colonomics_config > col_${i}_bd

done
```