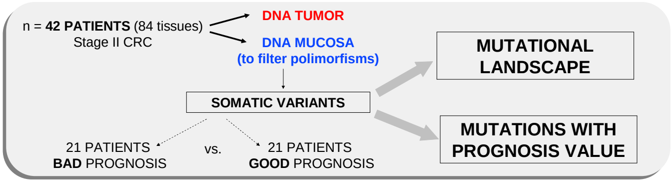
Exome sequencing reveals highly heterogeneous mutational patterns in stage II colorectal tumors

Rebeca Sanz-Pamplona[1], Adriana López-Dóriga[1], Laia Paré-Brunet[1], Sergi Beltran[2], Francesc Castro[2], Marta Gut[2], Lídia Agueda[2], Francisco D. Morón-Duran[1], Kira Lázaro[1], Marco Chierici[3], Cesare Furlanello[3] , Víctor Moreno[1,4]

[1]Unit of Biomarkers and Susceptibility (IDIBELL)-Catalan Institute of Oncology (ICO) and CIBERESP. [2]Centre Nacional d'Anàlisi Genòmica (CNAG). Barcelona, Spain. [3]Predictive Models for Biomedicine & Environment (PMBE), Foundazione Bruno Kessler (FBK). Trento. Italy. [4]University of Barcelona. Barcelona, Spain.

EXPERIMENTAL DESIGN & AIM

EXPERIMENTAL DESIGN



n = **42 PATIENTS** (84 tissues)
Stage II CRC

DNA TUMOR

DNA MUCOSA
(to filter polimorfisms)

SOMATIC VARIANTS

21 PATIENTS
BAD PROGNOSIS

vs.

21 PATIENTS
GOOD PROGNOSIS

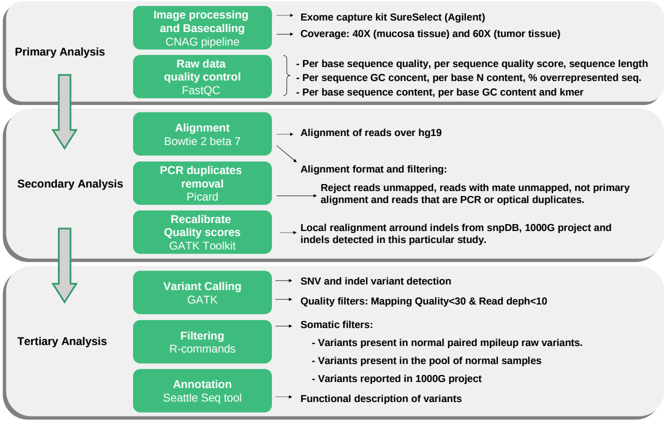
MUTATIONAL LANDSCAPE

MUTATIONS WITH PROGNOSIS VALUE

To characterize the mutational landscape of pre-metastasic tumors, the exonic DNA of a set of 42 stage II, MSS colon tumors and their paired mucosa were sequenced. Exome sequencing is a useful technique to discover mutations in CRC still unknown. We hypothesized that genetic alterations selected because they confer malignant advantage to cells are already present in the primary stages of colon carcinoma.

OBJECTIVES: The main objective is to search for novel recurrent mutations in stage II CRC that may drive tumor progression, and to check its putative association with prognosis. A derived objective is to functionally characterize genes harboring mutations in an attempt to better understand the CRC pathobiology.

SEQUENCING AND ANALYSIS PIPELINE



Primary Analysis

Image processing and Basecalling
CNAG pipeline
- **Exome capture kit SureSelect (Agilent)**
- **Coverage: 40X (mucosa tissue) and 60X (tumor tissue)**

Raw data quality control
FastQC
- **Per base sequence quality, per sequence quality score, sequence length**
- **Per sequence GC concent, per base N content, % overrepresented seq.**
- **Per base sequence content, per base GC content and kmer**

Secondary Analysis

Alignment
Bowtie 2 beta 7
- **Alignment of reads over hg19**
- **Alignment format and filtering:**

PCR duplicates removal
Picard
- **Reject reads unmapped, reads with mate unmapped, not primary alignment and reads that are PCR or optical duplicates.**

Recalibrate Quality scores
GATK Toolkit
- **Local realignment arround indels from snpDB, 1000G project and indels detected in this particular study.**

Tertiary Analysis

Variant Calling
GATK
- **SNV and indel variant detection**
- **Quality filters: Mapping Quality<30 & Read deph<10**

Filtering
R-commands
- **Somatic filters:**
  - **Variants present in normal paired mpileup raw variants.**
  - **Variants present in the pool of normal samples**
  - **Variants reported in 1000G project**

Annotation
Seattle Seq tool
- **Functional description of variants**

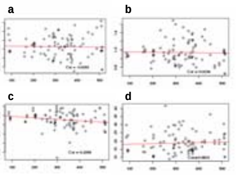
STAGE II COLORECTAL CANCER SOMATIC LANDSCAPE

SOMATIC SINGLE NUCLEOTIDE VARIANTS (SNV)

A. SNV EXAMPLE

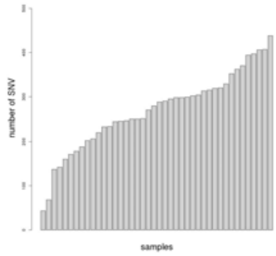
Example showing a KRAS mutation (C>T) in the tumor tissue not present in their paired mucosa (IGV software).

B. QUALITY CONTROL



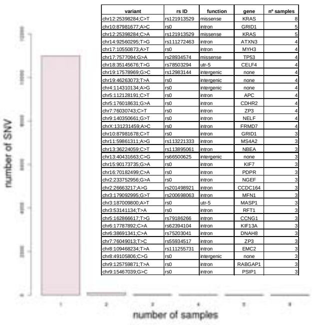
No correlation existed between the number of mutations found per sample and the quality parameters "number of reads" (a), "number of no matched reads" (b), "percentage of unique aligned reads" (c) and "number of duplicate reads" (d), in both mucosa and tumoral tissues.

C. SNV MUTATIONS / SAMPLE



A total of 11126 somatic single nucleotide variants (SNVs) had been found within the 42 analyzed tumors. Per sample, the number of SNVs ranged from 43 (the tumor with less mutations) to 438 (the tumor with more mutations), with a mean of 270 mutations per sample.

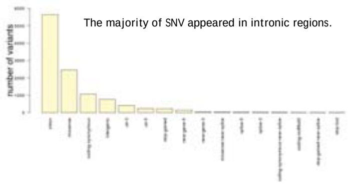
D. MUTATIONS DISTRIBUTION



variant	rs ID	function	gene	nº samples
chr12:25398284;C>T	rs121913529	missense	KRAS	8
chr10:87981677;A>C	rs0	intron	GRID1	5
chr12:25398284;C>A	rs121913529	missense	KRAS	5
chr14:92560295;T>G	rs111272463	intron	ATXN3	4
chr17:10550873;A>T	rs0	intron	MYH3	4
chr17:7577094;G>A	rs28934574	missense	TP53	4
chr18:35145676;T>G	rs78503294	utr-5	CELF4	4
chr19:17578969;G>C	rs12983144	intergenic	none	4
chr19:46263073;T>A	rs0	intergenic	none	4
chr4:114310134;A>G	rs0	intergenic	none	4
chr5:112128191;C>T	rs0	intron	APC	4
chr5:176018631;G>A	rs0	intron	CDHR2	4
chr7:76030743;C>T	rs0	intron	ZP3	4
chr9:140350661;G>T	rs0	intron	NELF	4
chrX:131231459;A>C	rs0	intron	FRMD7	4
chr10:87981678;C>T	rs0	intron	GRID1	3
chr11:59861311;A>G	rs113221333	intron	MS4A2	3
chr13:36224059;C>T	rs113895061	intron	NBEA	3
chr13:40431663;C>G	rs66500625	intergenic	none	3
chr15:90173735;G>A	rs0	intron	KIF7	3
chr16:70182499;C>A	rs0	intron	PDPR	3
chr2:233752956;G>A	rs0	intron	NGEF	3
chr2:26663217;A>G	rs201498921	intron	CCDC164	3
chr3:179092995;G>T	rs200698063	intron	MFN1	3
chr3:187009800;A>T	rs0	utr-5	MASP1	3
chr3:53141134;T>A	rs0	intron	RFT1	3
chr5:162866617;T>G	rs79186266	intron	CCNG1	3
chr6:17787892;C>A	rs62394104	intron	KIF13A	3
chr6:38691341;C>A	rs75203041	intron	DNAH8	3
chr7:76049013;T>C	rs55934517	intron	ZP3	3
chr8:109468234;T>A	rs111255731	intron	EMC2	3
chr8:49105806;C>G	rs0	intergenic	none	3
chr9:125759871;T>A	rs0	intron	RABGAP1	3
chr9:15467039;G>C	rs0	intron	PSIP1	3

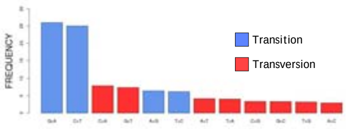
The vast majority of mutations were patient-exclusive so only appeared in one tumor. Only 138 SNV were shared by 2 or more tumors. The most recurrent mutation was the well-known KRAS G12D (8 out of 42 tumors).

E. FUNCTIONAL EFFECT



The majority of SNV appeared in intronic regions.

F. BASE CHANGE

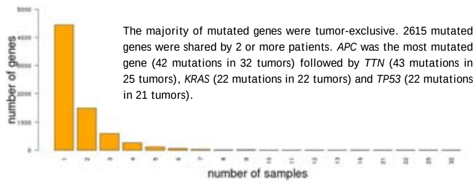


Transitions were more frequent than transversions.

MUTATED GENES

gene	nº mut	nº samples
APC	42	32
TTN	43	25
KRAS	22	22
TP53	22	21
LRP1B	22	16
SYNE1	20	13
RYR2	15	13
CDH4	14	13
SDK1	16	12
FAT4	13	12
MUC16	13	11
USH2A	13	11
COL5A1	12	10
MUC19	12	10
HMCN1	11	10
CCDC129	10	10
ERBB4	10	10
ZFHX4	10	10
CSMD1	14	9
COL22A1	12	9
GRID1	12	9
ATM	11	9
CSMD3	11	9
DNAH8	11	9
FRAS1	11	9
LRP2	11	9
HYDIN	10	9
PCDHA9	10	9
PKHD1L1	10	9
PXDNL	10	9
ABCA13	9	9
C5orf42	9	9
FBXW7	11	8
NBEA	11	8
RYR3	11	8
CACNA1E	9	8
PRPF6	9	8
CACNA1G	8	8
LPHN3	8	8
NALCN	8	8
OBSCN	8	8
RIMS2	8	8
ROBO2	8	8
WDFY4	8	8

A. SNV DISTRIBUTION ACROSS GENES



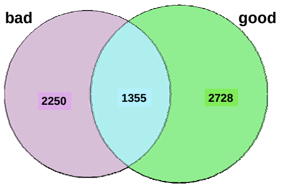
The majority of mutated genes were tumor-exclusive. 2615 mutated genes were shared by 2 or more patients. *APC* was the most mutated gene (42 mutations in 32 tumors) followed by *TTN* (43 mutations in 25 tumors), *KRAS* (22 mutations in 22 tumors) and *TP53* (22 mutations in 21 tumors).

B. VALIDATION

Gene	% mutated samples our data (n = 42)	% mutated samples Kann et al (n = 100)
ADORA1	1 / 42 = 2.38 %	4 / 100 = 4 %
APC	32 / 42 = 76.19 %	76 / 100 = 76 %
ARHGAP36	2 / 42 = 4.76 %	8 / 100 = 8 %
FBXW7	8 / 42 = 19.04 %	21 / 100 = 21 %
FZD3	1 / 42 = 2.38 %	10 / 100 = 10 %
KRAS	22 / 42 = 52 %	30 / 100 = 30 %
MESDC2	1 / 42 = 2.38 %	3 / 100 = 3 %
NRG1	6 / 42 = 14.29 %	10 / 100 = 10 %
PI15	2 / 42 = 4.76 %	4 / 100 = 4 %
SALL4	3 / 42 = 7.14 %	13 / 100 = 13 %
SMAD4	3 / 42 = 7.14 %	10 / 100 = 10 %
ST8SIA2	3 / 42 = 7.14 %	5 / 100 = 5 %
TP53	21 / 42 = 50 %	31 / 100 = 31%

Our results were in concordance with those described by Kann et al. (*Domain landscapes of somatic mutations in cancer, BMC Genomics, 2012*) Moreover, well-known recurrent mutations in *KRAS* and *TP53* genes were validated using the Biomark platform achieving good results.

C. MUTATED GENES DISTRIBUTION BETWEEN BAD AND GOOD GROUPS



Venn diagram showing the intersection between mutated genes in good and bad prognosis groups. 2250 genes were mutated only in bad tumors and 2728 only in good tumors. 1355 genes were mutated in both groups. Intronic and intergenic mutations were taken into account.

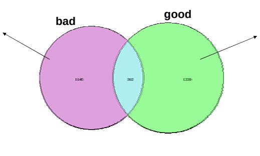
FUNCTIONAL ANALYSIS

A. CRC MUTATED GENES

Term	adj p-value	Biological process term description
GO:0007399	3.42E-88	nervous system development
GO:0009653	3.42E-88	anatomical structure morphogenesis
GO:0006464	1.91E-86	protein modification process
GO:0043687	1.10E-82	post-translational protein modification
GO:0048468	5.12E-74	cell development
GO:0006796	5.25E-73	phosphate-containing compound metabolic process
GO:0007155	6.00E-70	cell adhesion
GO:0006468	1.81E-69	protein phosphorylation
GO:0016310	1.86E-65	phosphorylation
GO:0009887	2.24E-62	organ morphogenesis
GO:0010646	3.02E-59	regulation of cell communication
GO:0022008	1.64E-53	neurogenesis
GO:0006811	4.83E-53	ion transport
GO:0048523	1.46E-52	negative regulation of cellular process
GO:0048699	1.91E-50	generation of neurons
GO:0009966	1.81E-49	regulation of signal transduction
GO:0030030	1.05E-46	cell projection organization
GO:0030182	1.32E-45	neuron differentiation
GO:0009893	6.82E-45	positive regulation of metabolic process
GO:0006812	1.03E-44	cation transport
GO:0009790	1.70E-44	embryo development
GO:0030001	4.48E-44	metal ion transport
GO:0016192	5.14E-43	vesicle-mediated transport
GO:0051649	2.84E-42	establishment of localization in cell
GO:0007417	5.12E-41	central nervous system development
GO:0042221	5.12E-41	response to chemical stimulus
GO:0042592	5.12E-41	homeostatic process
GO:0009888	6.25E-40	tissue development
GO:0000904	1.07E-38	cell morphogenesis involved in differentiation
GO:0006928	3.78E-38	cellular component movement
GO:0000902	1.41E-37	cell morphogenesis
GO:0006366	3.69E-37	transcription from RNA polymerase II promoter
GO:0007267	8.40E-37	cell-cell signaling
GO:0010628	9.44E-37	positive regulation of gene expression
GO:0048666	1.78E-36	neuron development
GO:0048878	4.54E-36	chemical homeostasis
GO:0019226	4.74E-36	transmission of nerve impulse
GO:0050793	8.03E-36	regulation of developmental process
GO:0048646	9.92E-36	anatomical structure formation involved in morphogenesis
GO:0048870	4.18E-35	cell motility
GO:0007167	6.04E-35	enzyme linked receptor protein signaling pathway
GO:0031328	1.26E-34	positive regulation of cellular biosynthetic process
GO:0008104	5.03E-34	protein localization
GO:0016477	2.21E-33	cell migration
GO:0008283	3.90E-33	cell proliferation
GO:0008219	5.34E-33	cell death

B. DIFFERENTIAL FUNCTIONS IN BAD AND GOOD PROGNOSIS GROUPS

Term	adj p-value	Biological process term description
GO:0030154	2.09E-25	cell differentiation
GO:0048513	1.57E-22	organ development
GO:0006464	1.57E-22	protein modification process
GO:0009653	1.32E-21	anatomical structure morphogenesis
GO:0043687	1.93E-20	post-translational protein modification
GO:0006796	1.48E-18	phosphate-containing compound metabolic process
GO:0016310	5.48E-17	phosphorylation
GO:0042221	2.67E-08	response to chemical stimulus
GO:0048523	4.31E-08	negative regulation of cellular process
GO:0007169	7.53E-08	transmembrane receptor protein tyrosine kinase signaling pathway
GO:0050848	2.07E-07	regulation of calcium-mediated signaling
GO:0051649	2.44E-07	establishment of localization in cell
GO:0048839	8.08E-07	inner ear development
GO:0009952	9.31E-07	anterior/posterior pattern specification
GO:0010212	1.59E-06	response to ionizing radiation
GO:0006996	3.03E-06	organelle organization
GO:0048469	3.20E-06	cell maturation
GO:0050679	3.25E-06	positive regulation of epithelial cell proliferation
GO:0043583	3.37E-06	ear development
GO:0001894	4.60E-06	tissue homeostasis
GO:0007015	7.58E-06	actin filament organization
GO:0035282	8.07E-06	segmentation
GO:0006983	8.64E-06	ER overload response
GO:0008016	8.97E-06	regulation of heart contraction
GO:0009605	9.25E-06	response to external stimulus
GO:0006942	1.08E-05	regulation of striated muscle contraction
GO:0008219	1.21E-05	cell death
GO:0051606	1.31E-05	detection of stimulus
GO:0009101	1.38E-05	glycoprotein biosynthetic process
GO:0008154	2.16E-05	actin polymerization or depolymerization
GO:0032956	2.34E-05	regulation of actin cytoskeleton organization
GO:0050678	2.67E-05	regulation of epithelial cell proliferation
GO:0032259	4.61E-05	methylation
GO:0048489	5.18E-05	synaptic vesicle transport
GO:0046907	5.30E-05	intracellular transport
GO:0060047	5.49E-05	heart contraction
GO:0051493	5.73E-05	regulation of cytoskeleton organization
GO:0042063	7.48E-05	gliogenesis
GO:0055117	7.80E-05	regulation of cardiac muscle contraction
GO:0045744	8.75E-05	negative regulation of G-protein coupled receptor protein signaling pathway
GO:0016458	9.18E-05	gene silencing
GO:0060048	9.19E-05	cardiac muscle contraction

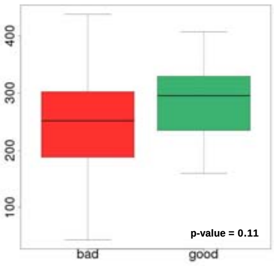


Term	adj p-value	Biological process term description
GO:0043588	4.91E-07	skin development
GO:0043279	6.02E-07	response to alkaloid
GO:0007416	6.11E-07	synapse assembly
GO:0009892	3.29E-06	negative regulation of metabolic process
GO:0009308	3.65E-06	amine metabolic process
GO:0048663	4.88E-06	neuron fate commitment
GO:0060433	5.09E-06	bronchus development
GO:0017156	5.89E-06	calcium ion-dependent exocytosis
GO:0035264	8.24E-06	multicellular organism growth
GO:0060078	1.12E-05	regulation of postsynaptic membrane potential
GO:0021953	1.22E-05	central nervous system neuron differentiation
GO:0045124	1.28E-05	regulation of bone resorption
GO:0003007	1.41E-05	heart morphogenesis
GO:0042246	1.80E-05	tissue regeneration
GO:0048863	1.83E-05	stem cell differentiation
GO:0014031	1.93E-05	mesenchymal cell development
GO:0048762	1.93E-05	mesenchymal cell differentiation
GO:0046850	2.25E-05	regulation of bone remodeling
GO:0035176	2.25E-05	social behavior
GO:0045780	2.30E-05	positive regulation of bone resorption
GO:0060438	2.30E-05	trachea development
GO:0002076	2.59E-05	osteoblast development
GO:0021904	3.75E-05	dorsal/ventral neural tube patterning
GO:0042220	3.75E-05	response to cocaine
GO:0001946	4.05E-05	lymphangiogenesis
GO:0048730	4.74E-05	epidermis morphogenesis
GO:0031645	4.94E-05	negative regulation of neurological system process
GO:0035094	5.88E-05	response to nicotine
GO:0001570	6.26E-05	vasculogenesis
GO:0001945	6.53E-05	lymph vessel development
GO:0009744	6.53E-05	response to sucrose stimulus

FatiGO software (www.babelomics.org) were used to make a functional analysis enrichment. Only somatic mutations with a functional impact were taken into account.

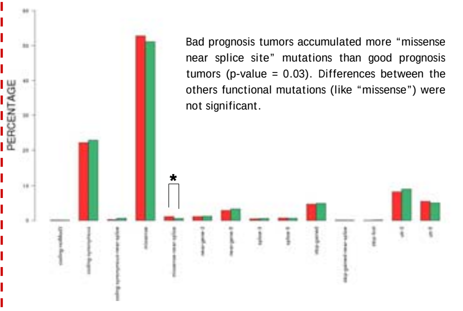
PROGNOSIS ASSOCIATION

A. NUMBER OF MUTATIONS PER GROUP



No significant differences in the total number of mutations between the two good and bad prognosis groups existed. Similar result was achieved if intronic and intergenic mutations were excluded from the analysis. Due to their low recurrence (the vast majority of SNVs were patient exclusive) and subsequently lack of statistical power, none of these mutations were associated with prognosis.

B. PERCENTAGE OF MUTATIONS PER GROUP



Bad prognosis tumors accumulated more "missense near splice site" mutations than good prognosis tumors (p-value = 0.03). Differences between the others functional mutations (like "missense") were not significant.

CONCLUSIONS

1- Our analysis reveals a high mutational heterogeneity across patients. The vast majority of SNVs were patient-exclusive so only present in a unique sample.

2- The well-known KRAS G12D mutation is the most recurrent one appearing in 8 out of 42 samples (19%). Collapsing by genes, *APC* is the most mutated one followed by *TTN*, *KRAS* and *TP53*. *KRAS* and *TP53* mutations has been validated in the same samples using an alternative technique.

3- No single mutation was associated with prognosis. Bad prognosis tumors do not accumulate more somatic mutations than good prognosis ones but indeed accumulate more "missense near splice site mutations".

4- A functional analysis showed differences in mutated functions and pathways between good and bad prognosis groups of tumors.

ONGOING WORK

1- Apart from well-known CRC mutations such as those in *KRAS*, *TP53* or *APC*, novel somatic and recurrent mutations found in our data need further validation:

- An independent validation using exome sequencing data from TCGA is now in progress, though the information about prognosis is limited.

-To assess the reliability of the technique, a independent set of tumors (n=200) has been recruited to validate recurrent somatic mutations using Sanger sequencing.

2- We plan to better characterize exclusive pathways found in good and bad prognosis groups integrating exome sequencing data with expression data from the same set of tumors (see our web page www.colonomics.org).