Article

Identification of novel mutations in colorectal cancer patients using Ampliseq comprehensive cancer panel

Bader Almuzzaini[1†*], Jahad Alghamdi[2†], Alhanouf Alomani[3], Saleh AlGhamdi[4], Abdullah A. Alsharm[5], Saeed Alshieban[6], Ahood Sayed[2], Abdulmohsen G. Alhejaily[7], Feda S Aljaser[8], Manal Abudawood[8], Faisal Almajed[9], Abdulhadi Samman[10], Mohammed Balwi[1], Mohammad Azhar Aziz[11,*]

1 King Abdullah International Medical Research Center, Medical Genomics Research Department, Ministry of National Guard Health Affairs, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

2 King Abdullah International Medical Research Center, The Saudi Biobank, King Saud Bin Abdulaziz University for Health Sciences, Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia

3 Department of Pathology, College of Medicine, Al-Imam Mohammad Ibn Saud Islamic University(IMSIU), Riyadh, Saudi Arabia

4 Clinical Research Department, Research Center, King Fahad Medical City, Riyadh, Saudi Arabia

5 Comprehensive Cancer Center, King Fahad Medical City, Riyadh, Saudi Arabia

6 King Abdul Aziz Medical city-National Guard Health Affairs (NGHA), King Abdullah International Medical Research Center, King Saud Bin Abdul Aziz University for Health Sciences (KSAU-HS), Saudi Arabia

7 Faculty of Medicine, King Fahad Medical City, Riyadh, Saudi Arabia

8 Department of Clinical Laboratory Sciences, Chair of Medical and Molecular Genetics Research, College of Applied Medical Sciences, King Saud University Riyadh, Riyadh, Saudi Arabia

9 Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, King Saud Bin Abdulaziz University for Health Sciences, Ministry of National Guard Health Affairs, Riyadh, Saudi Arabia

10 Department of Pathology, Faculty of Medicine, University of Jeddah, Jeddah, Saudi Arabia

11 King Abdullah International Medical Research Center, Colorectal Cancer Research Program, Department of Cellular therapy and cancer research, Ministry of National Guard Health Affairs, King Saud Bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

† Contributed equally

* Corresponding Authors:

1. Mohammad A. Aziz, PhD

   King Abdullah International Medical Research Center

   P.O. Box 22490, Riyadh 11426 KSA Mail Code 1515

   Tel. No.: +966 11 4294582

   Email: azizmo@ngha.med.sa

2. Bader Almuzzaini, PhD

   King Abdullah International Medical Research Center

   P.O. Box 22490, Riyadh 11426 KSA Mail Code 1515

   Tel. No.: +966 11 4294533

   Email: MuzainiB@NGHA.MED.SA

Abstract: Biomarker discovery would be an important tool in advancing and utilizing the concept of precision and personalized medicine in the clinic. Discovery of novel variants in local population provides confident targets for developing biomarkers for personalized medicine. We identified the need to generate high quality sequencing data from local colorectal cancer patients and understand the pattern of occurrence of variants. In this report, we used archived samples from Saudi Arabia and used Ampliseq Comprehensive Cancer panel to identify novel somatic variants. We report a comprehensive analysis of next generation sequencing results with a coverage of >300X. We identified 466 novel variants which were previously unreported in COSMIC and ICGC databases. We analyzed the genes associated with these variants in terms of their frequency of occurrence, probable pathogenicity and clinicopathological features. Among pathogenic somatic variants, 174 were identified for the first time in large intestine. APC, RET and EGFR genes were most frequently mutated. Higher number of variants were identified in left colon. Occurrence of variants in ERBB2 was significantly correlated with those of EGFR and ATR genes. Network analyses of the identified genes provide functional perspective of the identified genes and suggest affected pathways and probable biomarker candidates. This report lays the ground work for biomarker discovery and identification of driver gene mutations in local population.

Keywords: Colorectal cancer, personalized medicine, biomarker, Ampliseq

1. Introduction

Colorectal cancer (CRC) is a heterogeneous disease. Inter-patient heterogeneity has been one of the major obstacles towards developing therapeutic strategies. Different populations have been found to show varied response towards standard of care regimens [1]. This variation has largely been attributed to the difference in underlying gene mutations and genetic changes which determines the progression of CRC. CRC progresses with continuing accumulation of genomic and epi-genomic alterations, which eventually induce oncogenic transformation of the normal colon cell into tumor cells followed by metastasis. Pathways responsible to initiate CRC are well known based on the evidence of mutations and chromosomal changes observed in patients. The mechanistic role of signaling pathways in causing CRC have constantly been enriched with better understanding of the underlying gene mutations. These gene mutations have been used as biomarkers to predict disease progression and outcome of therapeutic regimens.

KRAS mutation status is routinely used for administering antibodies to inhibit epidermal growth factor receptor (EGFR). Successful use of these antibodies (cetuximab and panitumumab) only in KRAS wild type patients had set the stage of precision and personalized medicine. However, not all patients with wild type KRAS gene, respond to anti-EGFR therapy. Therefore, there is a pertinent need to identify biomarkers that can capture the population heterogeneity and facilitate the practice of precision and personalized medicine. Earlier studies have taken up population based mutational profiling of CRC to develop the concept of precision medicine [2,3]. Population specific mutational analysis of colorectal cancer is scarce in Saudi Arabia but highly pertinent to develop the precision and personalized medicine paradigm [4-6]. With the technological advancement in detecting mutations at an unprecedented scale, the possibility of practicing precision medicine through biomarkers has further increased. There is better precision and accuracy in detecting mutations in patients that can be used as predictive and prognostic biomarkers.

Next Generation Sequencing (NGS) technology can be used with DNA enrichment methods to generate deep sequencing of target genes or genomic regions of interest, such as the exome or identified cancer “hotspots”. For the targeted detection of mutations in known cancer genes, a comprehensive cancer panel (IonAmpliSeq) is available. Gene panels allow simultaneous detection of relevant mutations with unprecedented accuracy and sensitivity This Comprehensive Cancer Panel (CCP) is designed to target coding DNA sequences (CDS) and splice variants from 409 tumor suppressor and onco- genes that are frequently mutated. The requirement of small amount of input DNA (only 40 ng) per reaction enabled challenging analysis of formalin fixed paraffin

embedded (FFPE) tissues. The use of the IonAmpliSeq™ Cancer Panel and NGS using the IonTorrent platform, provides a fast, easy and cost effective sequencing workflow for detecting genomic “hotspot” regions that are frequently mutated in human cancer. A previous study from Jeddah, Saudi Arabia have used Ion AmpliSeq™ Cancer Hotspot Panel v2 which spans only 50 frequently mutated genes [7].

In this study, we used IonAmpliSeq™ CCP to sequence samples from 99 archived patient samples from two hospitals in Riyadh, Saudi Arabia. The confirmation of well-known mutations point towards chromosomal instability pathway as predominant mechanism of development of CRC in this cohort. We provide comprehensive analyses of novel variants that can be useful for biomarker discovery and identification of driver genes. Discovery of biomarkers and identification of driver genes from local population is critical in developing precision and personalized medicine approach towards addressing colorectal cancer.

Materials and Methods

Patient description and sample collection

100 patient tumor samples were retrospectively recruited in this study and after exclusion of 1 sample due to low DNA quality, we sequenced 99 samples and clinicopathological characteristics were available from 95 patients. Sequencing data from 90 of these samples qualified for coverage requirement and was used for further analyses. All samples were collected in the period between 2016 and 2018 at King Abdulaziz Medical City (KAMC) and King Fahad Medical City (KFMC), Riyadh, Saudi Arabia. All samples were diagnosed as primary colorectal adenocarcinoma at histopathology level. Patients were excluded if: (i) they had been treated with chemotherapy or radiotherapy prior to tumor resection, ii) they had familial adenomatous polyposis (FAP), or hereditary non-polyposis colorectal cancer (HNPCC), (iii) the formalin-fixed, paraffin-embedded tissue (FFPE) samples, patients’ clinical and pathological data, or written informed consent form signed by patient to access the archival samples were not available.

The formalin-fixed paraffin-embedded tissues (FFPE) blocks from patients with colorectal adenocarcinomas were retrieved from the archives of the Department of anatomical pathology laboratory in KAMC and KFMC. All slides were revised and marked by a histopatholgist before DNA extraction.We selected only marked tissue with tumor percentage more than 40% and used 1-2 slides for extraction based on tissue size. Chart reviews were done after obtaining the ethical approval to collect the demographic and clinicopathological features from the hospital information system "BESTCare 2.0 A" at KAMC including age at diagnosis, gender, tumor stage, site and metastasis grade.

Ethical approval

Full Institutional Review Board (IRB) approval was given by King Abdullah International Medical Research center (KAIMRC), Ministry of National Guard, Health Affairs (IRB protocol #RC13/249/R). All patients' data were secured and accessed only by research investigators.

DNA extraction

Genomic DNA was extracted from FFPE samples which were assessed by a pathologist to select the appropriate block to assure presence of colorectal cancer cells and excluded the insufficient necrotic tissue for NGS. DNA was extracted either from slide sample using Ion Ampliseq ™ Direct FFPE DNA Kit (Thermo Fisher Scientific Inc, Canada) according to the manufacturer’s instructions. In case of FFPE block samples, DNA was extracted from FFPE blocks using 8µm of tissue ribbon using QIAamp DNA FFPE Tissue Kit (QIAGEN) following manufacturer’s instruction. Measurement of the DNA quality and concentration was done by using Qubit® 3.0 Fluorometer (Life Technologies).

Comprehensive cancer panel (CCP) and data availability

Pre-designed comprehensive cancer panel (CCP) from Ion AmpliSeq™ (Life Technologies) was used. This panel comprises of 16,000 primer pairs in four primer pools for 409 genes which cover approximately 15,749 somatic mutations reported in The Catalogue of Somatic Mutations in Cancer (COSMIC). For the complete list of 409 genes see Supplementary Table S1. All sequencing data generated from 90 patients is deposited in SRA database (reference PRJNA685957, https://www.ncbi.nlm.nih.gov/sra/PRJNA685957)

Library preparation and NGS data analysis

The library was constructed using Ion AmpliSeq™ (CCP) Library Kit 2.0 (Life Technologies), and Ion Xpress™ Barcode Adapter 1–16 Kit (Life Technologies) according to manufacturer's instructions. Library quantification was done using the Ion Library TaqMan Quantitation Kit (Life Technologies) following standard procedure available. The qualified library was sequenced by the use of Ion S5XL Semiconductor Sequencer following the manufacturer's user guide.

Variant calling and annotation

Variants were called by Torrent Suite Variant (TSV) (version 5.8) [8]. Variants with a coverage of more than 300X and read quality more than 50 were included in this study to enhance the quality of identified somatic variants. Variants that passed this quality metrics were annotated by using Ensemble Variant Effect Predictor (VEP) tool (version 102). This tool uses gnomAD, (version r2.1) and the Catalogue Of Somatic Mutations (COSMIC) databases (version 90) [9]. We excluded common variants previously reported in Ensemble (v102), and only included variants classified as confirmed "somatic" or "pathogenic" by COSMIC database. This classification is based on 'functional analysis through hidden markov models (FATHMM). Further, variants were classified into 'colorectal cancer associated' or 'other organ' sites. The potential damaging effect at protein level of the variants were assessed using prediction software using Sorting Intolerant From Tolerant (SIFT; v5.2.2) and Polymorphism Phenotyping v2 (Polyphen2; v2.2.2) scores [10,11]. These scores predict the impact of detected missense variants on the human protein structure. All variants which showed 'deleterious' on SIFT and/or 'damaging' on Polyphen2 were included for downstream analysis.

Molecular profiling and statistical analysis

Descriptive statistics were applied to summarize patient characteristics based on clinicopathological features. Summary statistics of the identified genetic variants were carried out in PLINK [12] to calculate the minor allele frequencies (MAF) and Hardy-Weinberg Equilibrium p-value [3]. Associations between mutations and CRC or histological features were determined using Fisher's Exact Test. Due to the limited sample size, tumor stages were grouped into early (stage I-II) and late (stage III-IV). Site of tumor was classified as left, right and others. The involvement of lymph nodes and secondary metastases were analyzed as dichotomous trait. All analyses were conducted using JMP Prostatistical software (JMP®, Version 13. SAS Institute Inc., Cary, NC, 1989-2019). Sequence Kernel Association - Optimal unified test (SKAT-O) was used to perform gene-based association analysis [13]. The association of rare variants with tumor stage (defined as late versus early), gender (female versus male), age group (young <50 years versus old), and tumor location (left versus right) was analyzed. The variants were weighted based on their allele frequency, where rare variants were assigned higher weight than common variants. To account for multiple testing, an adjusted p-value of 0.0001 was considered as a significant threshold, reflecting the Bonferroni correction of 409 genes.

Ingenuity pathway analysis

The networks for mutated genes were generated through the use of IPA (QIAGEN Inc., https://www.qiagenbio-informatics.com/products/ingenuity-pathway-analysis) [14]. Networks were created using following filter : Species=Human AND Disease=Cancer AND mutation=hemizygous OR in-frame OR gain-of-function OR frameshift OR missense OR homozygous OR null mutation OR silent OR heterozygous OR loss of function OR knockout OR nonsense. Two networks were generated – One with 27 most frequently mutated genes and another with 75 genes harboring pathogenic mutations reported in large intestine. 'Connect' function was used to investigate the known interactions among these genes. 'Overlay' function was used to find the association of these genes with canonical pathways and finding candidate biomarkers.

Results

Cohort characteristics

The baseline characteristics of the analyzed samples are shown in Table 1. The median age of patients was 62 years, with 58 of them being male (61%). According to TNM staging system, 65% of the patients were classified as T3 with 59% showing no spread to regional lymph node (T0), and 96% were without distant metastases. Highest proportion of patients were diagnosed as stage III (39%) and more tumors were located in left colon (52%).

Novel variants identified in colorectal cancer patient cohort

From a panel of 409 genes, we identified 4,256 variants. Among these, 483 variants were classified as novel as they were not found in COSMIC database. However 17 of these variants were reported in international cancer genome consortium (ICGC) database. All novel variants are provided as Supplementary Table S2. We checked for the probability of these variants to be germline by analyzing their variant allele frequency (VAF). 69 variants presented in at least one patient with a VAF between 49-51 or 99-100, indicating that they could be germline mutations, which is also supported by the MAF (>1%) among 45 of these variants (supplementary table S3).

Among 4,256 identified variants, 299 variants were classified as pathogenic. 174 variants from 299 pathogenic variants were found to be identified for the first time in large intestine representing novel variants in colorectal cancer (Figure 1 and Supplementary Table S4). We employed two different methods (SIFT and polyphen) for classifying 561 somatic variants. Both methods suggest the detected variants to be either synonymous (n=240) or missense (n=247) (Figure 2A) According to Polyphen scoring method, 143 mutations were predicted to be benign and rest could be pathogenic (Figure 2 B). SIFT prediction method also provided similar categorization with 111 variants listed as tolerated and 130 variants were classified as deleterious (Figure 2C).

Novel variants identified in most commonly mutated APC, RET and EGFR genes

Highest mutated genes (n=20) among the patients were identified based on the presence of at least one confirmed pathogenic variant and arranged according to decreasing trend of frequency in the patient cohort (Table 2). 96% of the patient samples had at least one confirmed pathogenic variant within APC gene. We identified 5 novel (defined as previously unreported in COSMIC database) variants out of total 38 variants detected in APC gene. These novel variants include c.1696G>A (p.V566I) missense mutation at exon 14, c.1697delT (p.V566X) frame shift mutation at exon 14, c.2680_2681delGTinsTA (p.Val894Ter) stop gain mutation at exon 16, c.3917delA (p.E1306X), frame shift mutation at exon 16, c.4320-4341del ACCACCTCCTCAAACAGCTCAA (p. PPPPQTAQ1440-1447X). 23 of 38 variants were confirmed as somatic variants in COSMIC. Of the 23 variants, 15 were confirmed as pathogenic, and 12 were confirmed as tissue-specific pathogenic variants for

large Intestine. RET gene mutations were found in 53% of the patient samples. Out of 13 detected variants in RET gene, 3 were somatic and one of them is pathogenic (p.L769=). 53% patient samples harbored EGFR gene mutation. Total 17 variants were detected, of which 6 were somatic and 2 variants were specific for large intestine. One of these variants was a high impact non sense mutation (p.R1068*). This comprehensive analysis and finding of novel variants within known genes would open up avenues to develop biomarkers that will be relevant for local population.

Colorectal cancer specific variants mapped to twenty seven genes

The distribution of pathogenic mutations found in large intestine across gender, age, tumor stage, site, lymph node and metastasis is described in Figure 3. 73 variants specific for 'colon and rectum' were identified within 27 genes. Tissue-specific pathogenic variants in the studied population show that APC gene was the highest mutated with variants detected in 66% of the samples, followed by ERBB2 (51%), ATR (45%), EGFR (40%), and FGFR3 (30%) genes. It is known that APC gene mutation is the initial event in CRC progression and is well depicted in our results. We observed variants in APC, ATR, KRAS,ATM and KIT genes in left colon of young female patients (<50years age) in stage 1. However no mutation was observed in young male patients in left colon in early stage (1&2) but mutations were observed in these patients in right colon and rectum. This detailed catalogue of variants analyzed according to clinicopathological features could be further used for molecular classification of patients.

Left colon exhibits higher mutation load

We identified 27 genes with at least one confirmed pathogenic variant and presented in at least three patients. We found that patients with left side of the colon had higher prevalence of mutated genes, with the exception of ATR, MLH1, ATM, MTOR, PDGFRA, EP300, COL1A1, PTEN and TSHR genes (Figure 4A). Significantly higher number of mutations were observed in FGFR3 gene in left side and EP300, TSHR genes on the right side of the colon. While comparing the early and late stage tumors, the prevalence of mutated genes were almost similar except for significantly higher COL1A1 gene mutations among patients in early stage when compared to late stage (Figure 4B).

Pathogenic variants in ERBB2 were significantly correlated with mutations in EGFR and ATR

Gene correlation analysis showed that occurrence of pathogenic gene mutations was correlated (Figure 5). Presence of pathogenic variants at ERBB2 was significantly correlated with mutations in EGFR and ATR (r^2 = 0.39 and 0.26; p-values = 0.0001 and 0.01, respectively). High correlation was found between KDM6A and UBR5 gene mutations (r^2= 0.47, p-value = 2.3E-6). FGFR3 gene was the most correlated. It was found positively correlated with HNF1A and TP53 whereas EGFR and ATR were negatively correlated.

By testing the association of mutated genes with clinicopathological variables, we found a significant association between ERBB2 mutation and tumor late stage (Fisher's exact t-test; p value = 0.04). Significant association between EP300 and TSHR mutations with right colon tumor (Chi-square; p = 0.02 and 0.01; respectively), and FGFR3 being more mutated in left colon (Chi-square; p = 0.01) were also observed.

For the gene-based rare variants analysis (SKAT-O), no gene was associated with clinicopathological variable at the significant threshold. However, suggestive significance was found between PIK3CB and colorectal cancer on left side (P = 0.0007), androgen receptor (AR) and female gender (P = 0.0002), TGM7 and young patient (p= 0.002), EXT1 and late stage (Table 3 A-D).

Network analysis of mutated genes

Using Ingenuity pathway analysis, we created an information based network of 27 highly mutated genes and found TP53 was the most connected node (Figure 6A). This network identified 16 druggable target genes. Network of 75 genes with pathogenic mutations in large intestine also exhibited TP53 as a highly connected node. 33 of these network genes were identified as target molecules (Figure 6B). Both networks identified TSHR gene as a potential druggable target (Supplementary Table S5).

Genes with pathogenic mutations in large intestine were found to be associated with hepatic fibrosis signaling, CRC metastasis, senescence, NF-κB and regulation of epithelial to mesenchymal transition pathways. Genes associated with these pathways are shown in Figure 6C. Biomarker analysis of these 75 genes revealed 16 candidate molecules, some of which are already in clinical use (Figure 6D and Table 4). These biomarkers have potential use in determining diagnosis, prognosis, efficacy and response to drugs.

Table 1. Clinicopathological features of CRC patients

Age, Years (SD)	62 (14)
Male, n (%)	58 (61%)
Stage	
I, n (%)	17 (18%)
II, n (%)	32 (34%)
III, n (%)	37 (39%)
IV, n (%)	8 (9%)
Primary Tumor	
T1, n (%)	2 (2%)
T2, n (%)	18 (19%)
T3, n (%)	61 (65%)
T4, n (%)	13 (14%)
Lymph Node	
N0, n (%)	55 (59%)
N1, n (%)	33 (35%)
N2, n (%)	6 (6%)
Distant Metastasis	
M0, n (%)	90 (96%)
M1, n (%)	4 (4%)
Site	
Left colon, n (%)	47 (52%)
Right colon, n (%)	30 (33%)
Rectum, n (%)	13 (14%)

T= Tumor, N= Node (0 = no nodes, 1= *1* node, 2= 2 nodes), M=Metastasis (0= no metastasis, 1 = metastasis); SD=Standard deviation

Table 2. List of twenty genes with variants in order of frequency in the sample cohort

Variants and individuals for the top 20 genes										
Gene	Variants								Individual[a]	
	Total	Novel	Pathogenic		Somatic	PolyPhen Damaging	SIFT Deleterious	Non-Synonymous	Pathogenic %	Somatic %
			All tissue	Tissue-Specific						
APC	38	5	15	12	23	1	2	30	0.96	0.99
RET	13	0	2	0	3	3	3	6	0.53	0.83
EGFR	17	0	4	2	6	0	0	7	0.53	0.69
LRP1B	70	12	10	1	22	3	1	48	0.52	0.86
ERBB2	14	3	3	3	4	1	2	10	0.51	0.52
ATR	31	6	4	2	11	1	0	19	0.46	0.68
CSMD3	42	5	4	0	8	5	3	34	0.44	0.52
RALGDS	13	1	2	1	3	1	2	6	0.36	0.42
HIF1A	8	0	3	0	3	0	0	6	0.36	0.36
FGFR3	18	2	2	2	3	1	0	11	0.33	0.34
KRAS	7	1	1	1	2	0	1	5	0.28	0.28
PIK3CG	14	1	4	1	5	1	4	7	0.22	0.23
TP53	24	1	16	16	18	1	9	23	0.21	0.31
HNF1A	13	0	5	3	4	0	0	9	0.20	0.27
PIK3R1	10	2	2	0	3	0	0	6	0.19	0.30
KDM6A	11	1	3	2	3	1	0	8	0.19	0.19
ATM	31	3	6	3	3	0	4	26	0.17	0.16
MLH1	9	1	2	1	3	0	1	8	0.16	0.17
PRDM1	5	0	2	0	2	0	1	3	0.16	0.16
JAK1	11	0	3	0	7	0	0	3	0.14	0.44

[a] Percentage of samples with at least one pathogenic or somatic variant within the gene

Table 3A: Top genes associated with Female versus male (232)

SetID	P.value	N.Marker.All	N.Marker.Test	MAC	m	Method.bin	MAP
AR	**0.0002**	**5**	**5**	**37**	**19**	**QA**	**-1**
BTK	0.003	1	1	5	5	ER	0.003523
SAMD9	0.003	14	14	66	30	QA	-1
PAX7	0.006605	6	6	41	26	QA	-1
KDM6A	0.010983	11	11	58	27	QA	-1

Table 3B: Top Genes associated with young group (<50 year old), versus old (285)

SetID	P.value	N.Marker.All	N.Marker.Test	MAC	m	Method.bin	MAP
TGM7	**0.002186**	**4**	**4**	**85**	**43**	**QA**	**-1**
MRE11A	0.006959	8	8	177	62	QA	-1
NBN	0.011436	11	11	171	47	QA	-1
VHL	0.015733	2	2	2	2	ER	0.015733
IDH1	0.019368	7	7	16	12	ER	1.32E-10

Table 3C: Top genes associated with left colorectal cancer, versus right (204)

SetID	P.value	N.Marker.All	N.Marker.Test	MAC	m	Method.bin	MAP
PIK3CB	**0.0007**	**9**	**9**	**9**	**8**	**ER**	**8.76E-05**
PIK3CA	0.001	12	12	86	39	QA	-1
RNF213	0.001	61	56	709	78	MA	-1
AURKB	0.003	5	5	101	39	QA	-1
ERBB4	0.005	11	11	41	27	QA	-1

Table 3D: Top genes associated with late stage, versus early stage (180)

SetID	P.value	N.Marker.All	N.Marker.Test	MAC	m	Method.bin	MAP
EXT1	**0.003027**	**7**	**7**	**73**	**51**	**MA**	**-1**
RNASEL	0.013275	4	4	34	29	ER.A	-1
CDH5	0.018248	8	8	135	62	MA	-1
BUB1B	0.021125	18	18	178	48	MA	-1
MUTYH	0.027108	11	11	108	51	MA	-1

N.Marker.All ; number of all variants within that gene; N.Marker.Test: number of variants entered the analysis (In our case we did not exclude common variants, but we assigned them lower weight. So it will be similar as N.Marker.All); MAC: total minor allele count (MAC); m: the number of individuals with minor alleles; method.bin: a type of method to be used to compute the p-value. MAP: minimum possible p-values. The number in the bracket shows the number of effective tests; (We choose to select a p value that is equal to 0.05/409)

Table 4. Candidate biomarkers from the list of 75 genes with pathogenic mutations in large intestine

Symbol	Entrez Gene Name	Location	Family	Entrez Gene ID for Human
APC	APC regulator of WNT signaling pathway	Nucleus	enzyme	324
BRAF	B-Raf proto-oncogene, serine/threonine kinase	Cytoplasm	kinase	673
CTNNB1	catenin beta 1	Nucleus	transcription regulator	1499
EGFR	epidermal growth factor receptor	Plasma Membrane	kinase	1956
ERBB2	erb-b2 receptor tyrosine kinase 2	Plasma Membrane	kinase	2064
FLT1	fms related receptor tyrosine kinase 1	Plasma Membrane	kinase	2321
IGF1R	insulin like growth factor 1 receptor	Plasma Membrane	transmembrane receptor	3480
KDR	kinase insert domain receptor	Plasma Membrane	kinase	3791
KIT	KIT proto-oncogene, receptor tyrosine kinase	Plasma Membrane	transmembrane receptor	3815
KRAS	KRAS proto-oncogene, GTPase	Cytoplasm	enzyme	3845
MLH1	mutL homolog 1	Nucleus	enzyme	4292
PDGFRA	platelet derived growth factor receptor alpha	Plasma Membrane	kinase	5156
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	Cytoplasm	kinase	5290
PTEN	phosphatase and tensin homolog	Cytoplasm	phosphatase	5728
SMAD4	SMAD family member 4	Nucleus	transcription regulator	4089
TP53	tumor protein p53	Nucleus	transcription regulator	7157

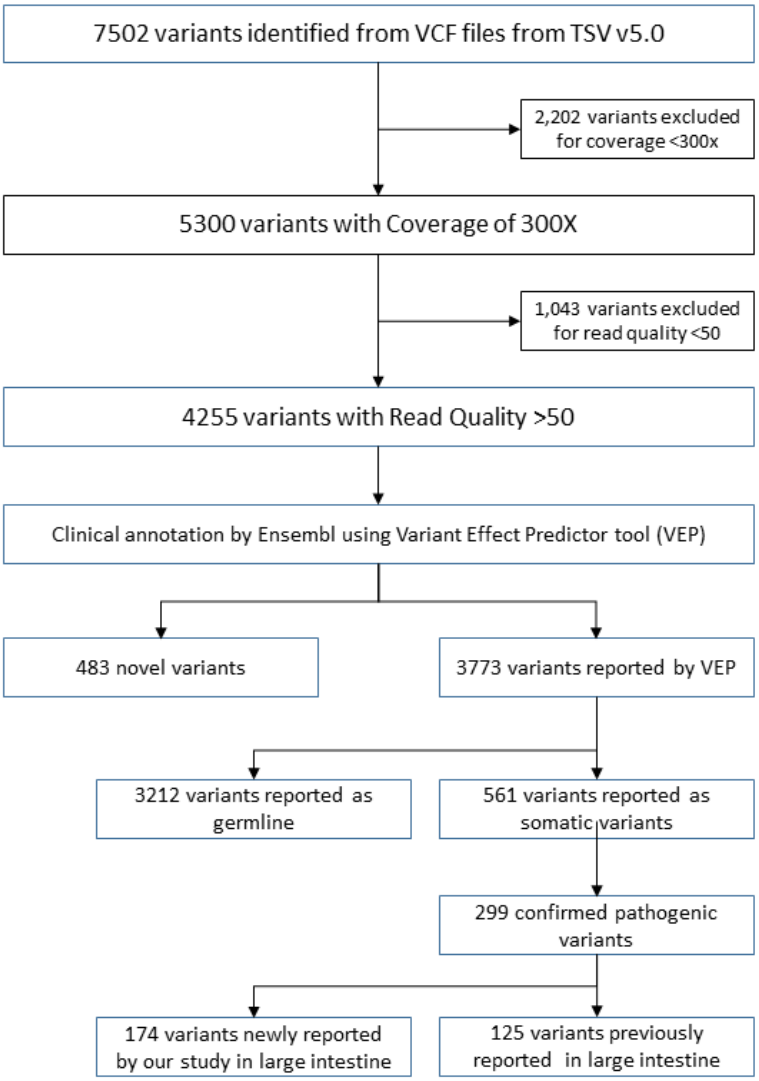


Figure 1

Figure 1: Variant filtration analysis workflow.

Schematic illustration of variants identified in this study. 483 novel variants were identified and 561 somatic variants were observed. This study focused on pathogenic variants that were identified as novel in large intestine

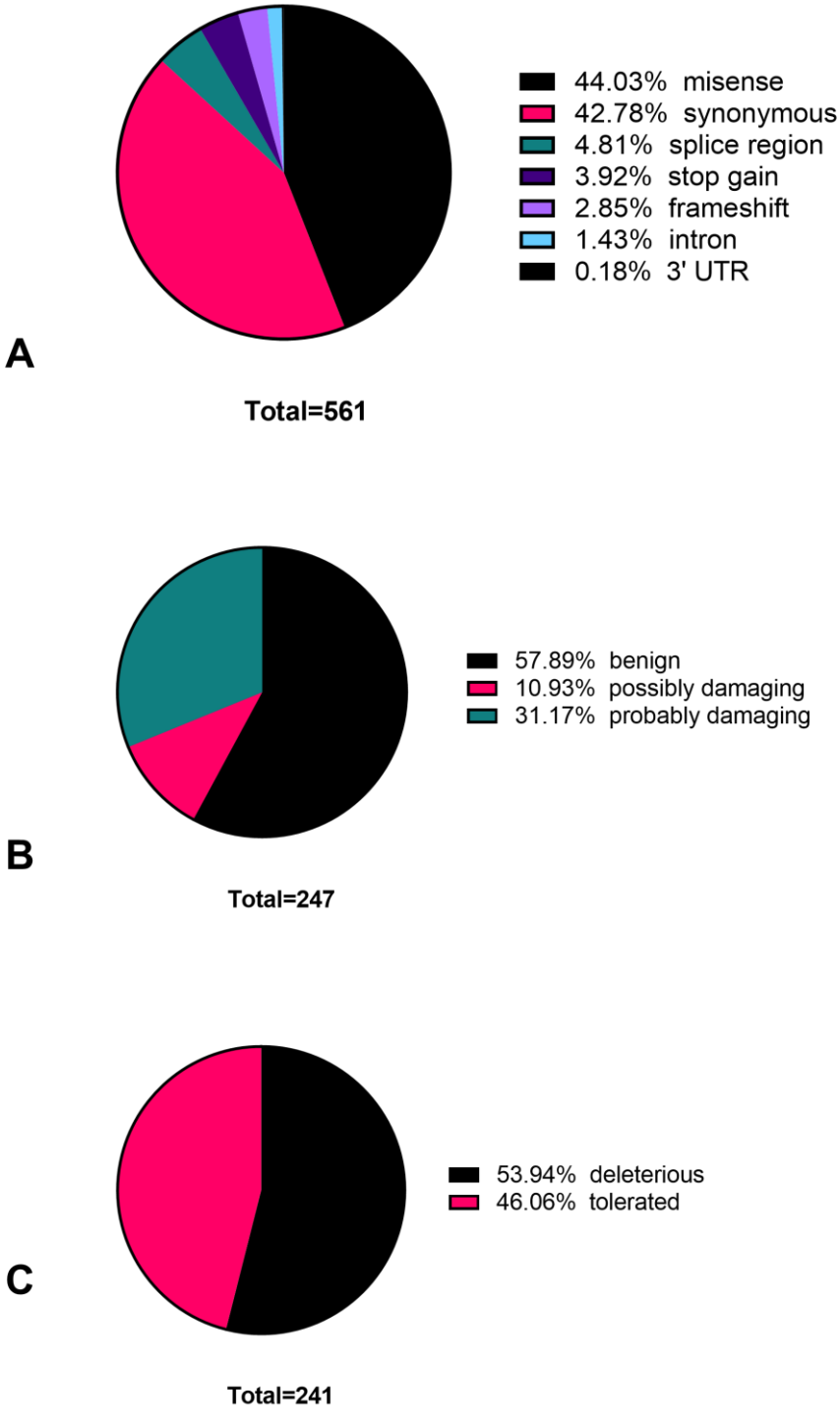


Figure 2: Classification of somatic variants

Total 561 variants were classified by consequence (A), polyphen score (B) SIFT score (C).

Figure 3

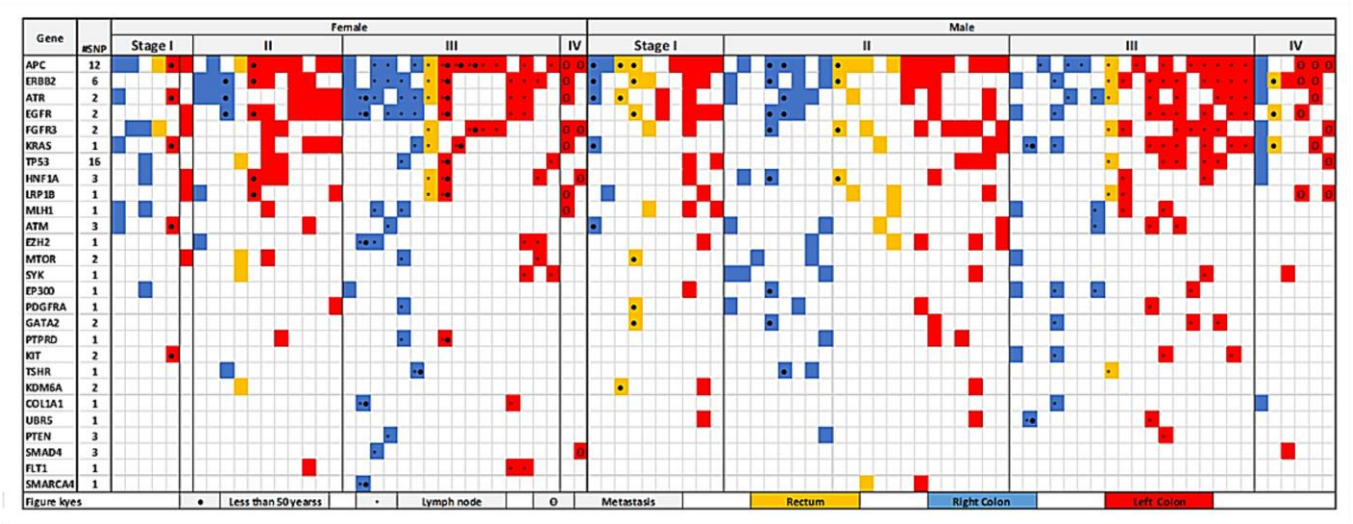
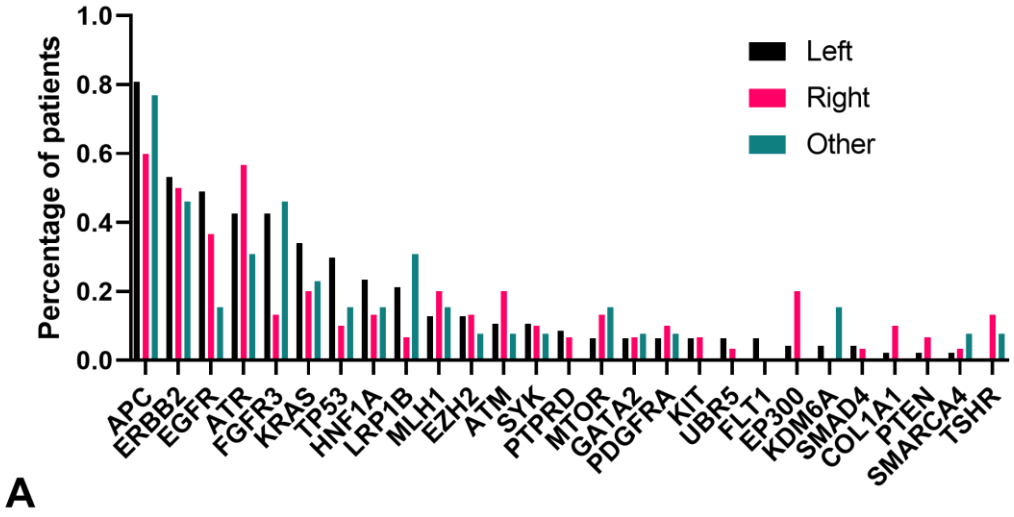
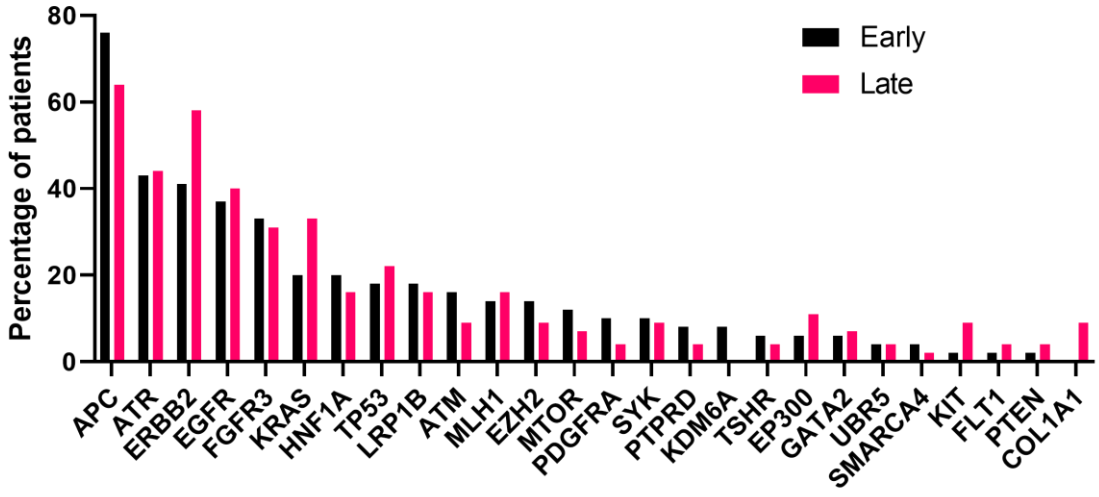


Figure 3: Mutation distribution based on gender, age, stage, site, lymph node and metastasis.

This figure shows only variants reported as pathogenic and located in large intestine in COSMIC database, and found in three individuals or more.



A



B

Figure 4

Figure 4: Frequency of variants in 27 genes among samples with at least one confirmed pathogenic variant for large intestine. Frequency of variants based on tumor location (A) and stage (B). Y-axis denotes the number of samples with at least one confirmed pathogenic variant for large intestine for that particular gene. For location, each bar is divided into left, right and other categories whereas for stage they were grouped into early (stage I & II) and late (stage III &IV).

Figure 5

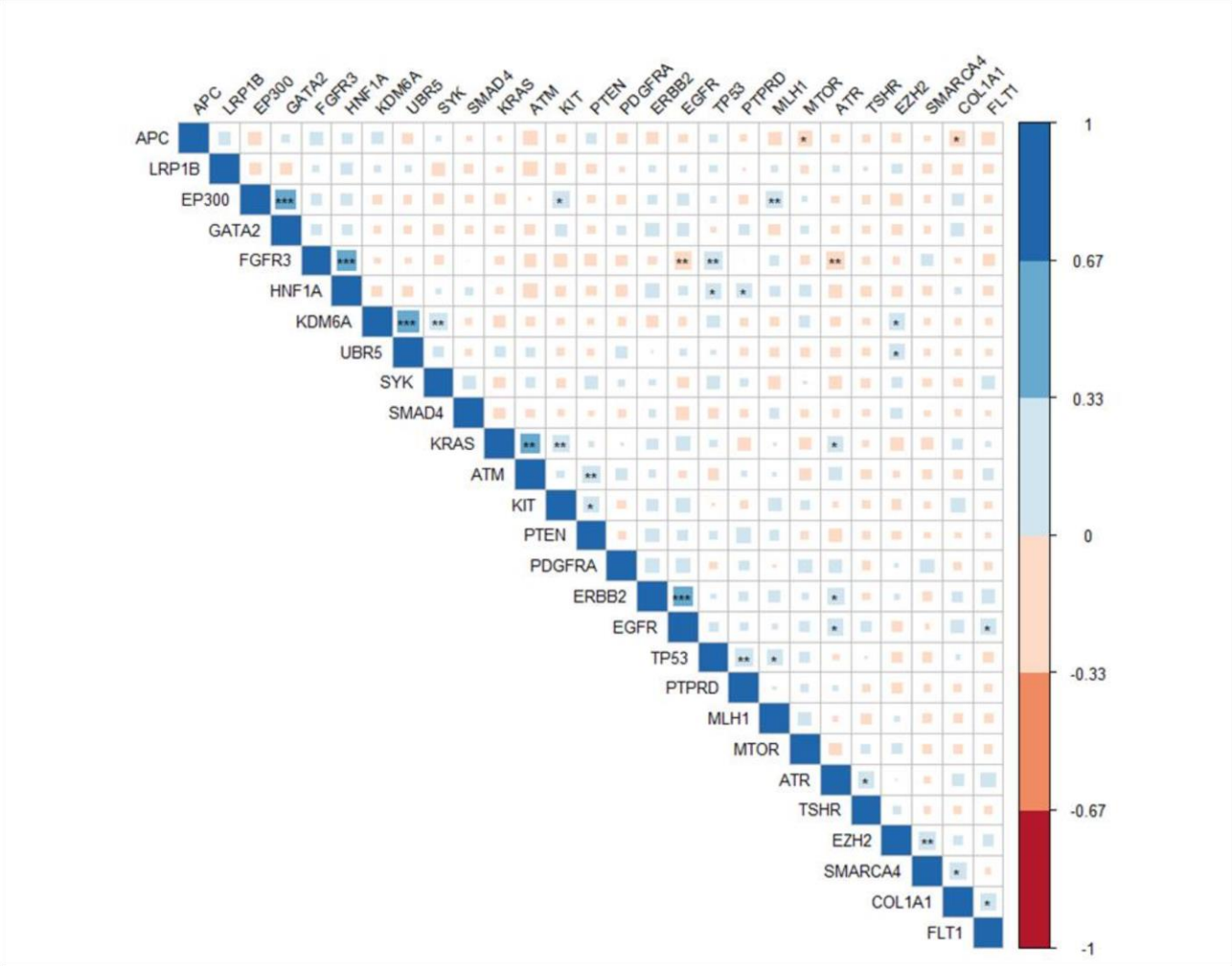


Figure 5: Correlation between mutated genes.

The Pearson correlation between presence of a tissue-specific pathogenic variant between genes. Significant correlation are noted with * for p-values < 0.05, ** for p-values < 0.01, and *** for p-values less than 0.001. Color and size of the square denotes the value of correlation as indicated in the bar legend.

Figure 6

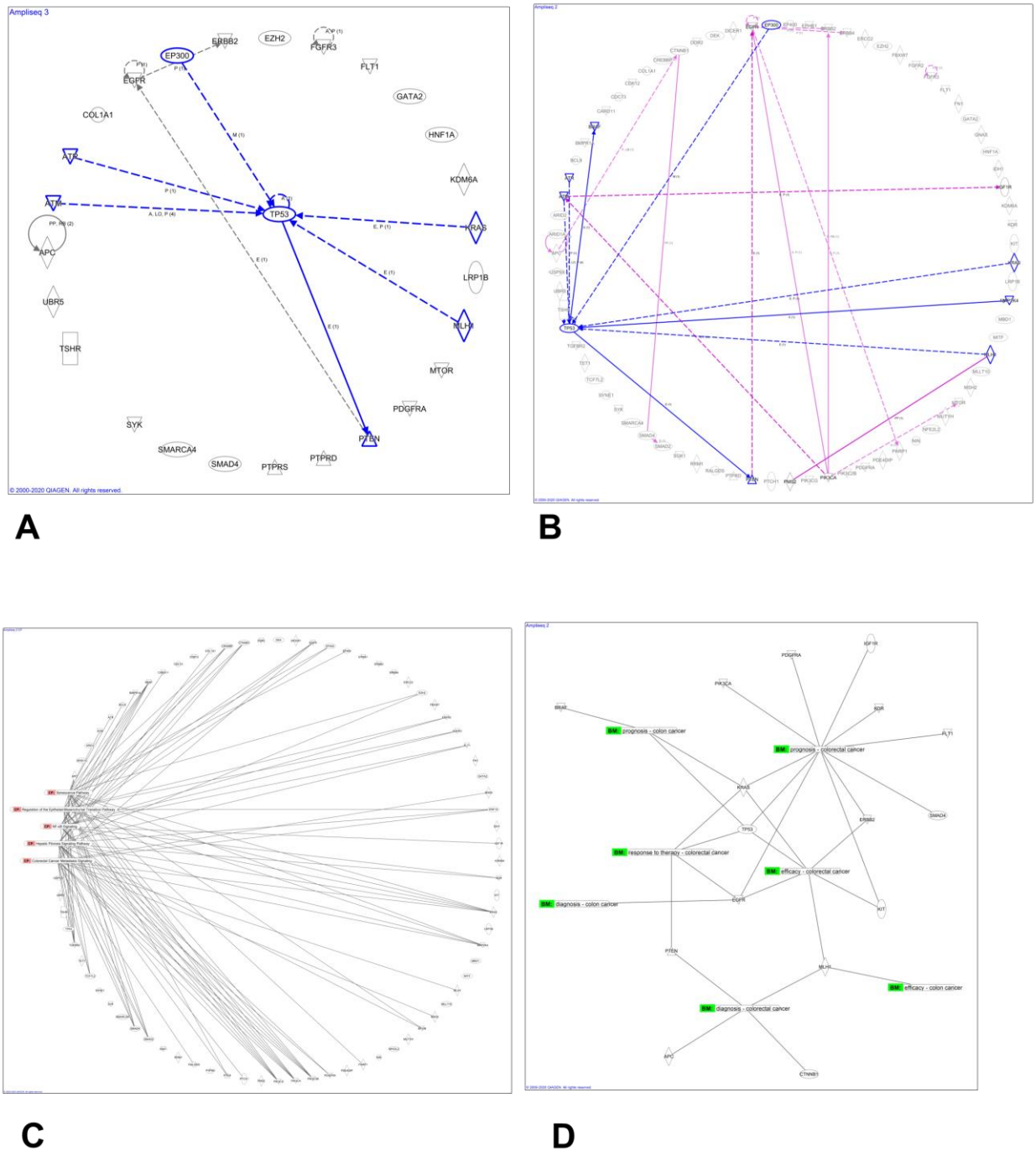


Figure 6: Network analyses of genes with reported variants.

Network of 27 most frequently mutated genes (A). Network of 75 genes harboring pathogenic mutations reported in large intestine (B). Association of 75 genes with canonical signaling pathways (C). Possible biomarker candidates for diagnosis, prognosis, efficacy and response to drugs for colon and colorectal cancer (D)

4. Discussion

Tumorigenesis and progression of cancer is suggested to be driven and supported by gene mutations [15-20]. Somatic mutations that are observed in cancer cells help to understand the cause and severity of the disease. Colorectal cancer is well known to have specific gene mutations associated with particular stages of the disease. In the present study, we aim to provide comprehensive analysis of gene variants as studied in a cohort of patients in Riyadh, Saudi Arabia. We employed NGS on Ampliseq comprehensive cancer panel to unravel the information locked in FFPE samples. This study provides successful evidence to support the use of archived samples and sequencing technology to generate information that is relevant for the local population. While we aim to understand the mutational profile in the local population, we found results that confirmed existing evidence supporting the initiation and progression of CRC. We also report novel variants in our population which is suggestive of unique genomic landscape of patients and supports the idea of precision and personalized medicine [5,6,21].

As determined by two separate prediction methods (Polyphen [11] and SIFT [22]), most of the detected mutations were missense and synonymous. This is in conformity with a recent pan cancer analysis [23] and opens up avenues to further study the effect of point mutations in CRC. These point mutations could be responsible for changes in gene expression and mRNA secondary structures. Similar studies from other populations have also reported the predominance of synonymous and missense mutations [2]. However, the challenge to separately identify driver mutations from passenger mutations with precision and accuracy is still an ongoing area of intense research [24-26].

We identified APC gene as highly mutated in our cohort with less common mutation frequency for RET, EGFR, LRP1B, and ERBB2 genes. EGFR has earlier been identified as one of the highest mutated genes in a cohort of patients from Jeddah, Saudi Arabia and confirms our results [7]. This study adds to the evidence of detected variants in a previous similar study from a different geographical location. More studies from different regions of the country are needed as there is an observed disparity in incidence and mortality of CRC in other regions within Saudi Arabia [4]. Identification of novel mutations in APC, RET and EGFR mutations may lead us to develop predictive and or prognostic biomarkers for CRC. Mutations in these genes have earlier been studied in detail for their use as biomarkers [27-30]. Most of the previous studies are associated with common mutation of APC gene except two studies on Arab cohort which showed APC mutation frequency was the second highest (34%) after P53 gene. Another study from gulf region patients showed 27.3% mutation frequency and higher mutation of TP53 (52.5%).

We found more CRC cases are localized at the left site comparing to right or rectum site. This is suggestive of the involvement of CIN pathway and is evident from our results which show APC, KRAS and P53 as highly mutated genes. However, PIK3CA do not appear among the highly mutated genes but 3 pathogenic mutation were identified among PIK3CA (p.R88Q, p.I102F and p.PI04L), all are reported in large intestine except (p.I102F mutation). This could be due to the population specific nature of the mutations and suggest further study to understand the mechanism of CRC progression in these patients. Left sided colorectal cancers have better prognosis and response to 5-Fluorouracil based and targeted therapies[31]. Our results are therefore very significant in understanding and predicting the prognosis of local patients which primarily exhibited mutations suggesting left sided CRC.

Increasing incidence of CRC cases in early ages has caused the guidelines for screening to be revised [32]. Our observation regarding young patients suggests presence of mutations in left colon of female patients in stage 1 whereas young male patients did not show any mutations in left colon in stage 1/2. This can be an important finding that can be studied further in larger cohorts to develop early diagnostic tests. Our catalog of reported variants have enriched the database for CRC and would be useful in building up larger studies for finding actionable targets and biomarkers. These variant information will need to be complemented with further levels of evidence to prove their role in CRC or identify them as drug targets. Multiomics approach is therefore recommended to be carried out on same samples for further proof of evidence [33,34].

The gene correlations observed in our cohort and network analysis would provide clues for the possible mechanism of CRC development. These networks and correlation analyses should be done at gene expression level to further understand the mechanistic details and effect of variants [35]. Network analyses confirms the probable effect of the detected variants through well-known pathways. We report mutations that can be associated with senescent pathway and points towards development of

therapeutic strategies. Targeting senescent pathways has been suggested as anticancer therapy and point towards their role in senescence and metastasis. Biomarker candidate molecules need to be further validated and tested for advancing into clinical setting.

Though our study is limited with smaller number of patient samples, it does exhibit the heterogeneous nature of CRC [36]. Another major limitation of our study is lack of matched normal samples to account for possible germline mutations. This is one of the bargains for utilizing the treasure of formalin fixed samples. Using a matched normal sample is a requirement for accurately classifying somatic mutations and ruling out germline mutations. However, the availability of matched normal tissue has been a limitation with archived and fresh samples [37]. Computational methods have been developed that are arguably better than matched normal tissue [38]. Most of the studies have relied on increasingly rich databases to identify novel mutations in absence of matched normal samples [39]. In order to address this issue, we used public databases and also employed an earlier reported method where the VAF corresponding to 50 or 100 % may indicate their probability to be germline mutation [40].

This study provides evidence that can be useful for developing biomarker based precision medicine as well as allow us to appreciate the heterogeneity in CRC and hence develop strategies accordingly.

Author Contributions: BM (Bader Al Muzaini): funding acquisition and project management . BM, JA (Jahad Alghamdi), SG (Saleh AlGhamdi),AH (Abdulmohsen G. Alhejaily),FJ (Feda S Aljaser),MAA (Mohammad Azhar Aziz), FM (Faisal Almajed) ,AS (Abdullah AlSharm),SA (Saeed Alshieban) ,AS (Ahood Sayed), MD (Manal AbuDawood) and AO (Alhanouf Alomani): contributed on study concept and study design and patient data collection. BM, AS (Abdulhadi Samman), MB (Mohammed AlBalwi) and AO: design experiment, sample collection , sample processing , performed library preparation , sequencing .JA: statistical analysis . BM , JA and MAA : analyzing data and original manuscript draft preparation, figures and tables preparation . All authors read and approved the final manuscript.

Funding: This study was supported by King Abdullah International Medical Research Center grant (RC 13/249/R) awarded to BM.

Institutional Review Board Statement: Full Institutional Review Board (IRB) approval was given by King Abdullah International Medical Research center (KAIMRC), Ministry of National Guard, Health Affairs (IRB protocol #RC13/249/R). All patients' data were secured and accessed only by research investigators.

Acknowledgments: We would like to acknowledge help of Dr. Mamoon Rashid in submitting the data to SRA repository.

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Rashid, M.; Vishwakarma, R.K.; Deeb, A.M.; Hussein, M.A.; Aziz, M.A. Molecular classification of colorectal cancer using the gene expression profile of tumor samples. Exp Biol Med (Maywood) **2019**, 244, 1005-1016, doi:10.1177/1535370219850788.
2. Zhunussova, G.; Afonin, G.; Abdikerim, S.; Jumanov, A.; Perfilyeva, A.; Kaidarova, D.; Djansugurova, L. Mutation Spectrum of Cancer-Associated Genes in Patients With Early Onset of Colorectal Cancer. Front Oncol **2019**, 9, 673, doi:10.3389/fonc.2019.00673.
3. Dos Santos, W.; Sobanski, T.; de Carvalho, A.C.; Evangelista, A.F.; Matsushita, M.; Berardinelli, G.N.; de Oliveira, M.A.; Reis, R.M.; Guimaraes, D.P. Mutation profiling of cancer drivers in Brazilian colorectal cancer. Sci Rep **2019**, 9, 13687, doi:10.1038/s41598-019-49611-1.
4. Alyabsi, M.; Alhumaid, A.; Allah-Bakhsh, H.; Alkelya, M.; Aziz, M.A. Colorectal cancer in Saudi Arabia as the proof-of-principle model for implementing strategies of predictive, preventive, and personalized medicine in healthcare. EPMA J **2020**, 11, 119-131, doi:10.1007/s13167-019-00186-x.

5. Aziz, M.A.; Allah-Bakhsh, H. Colorectal cancer: A looming threat, opportunities, and challenges for the Saudi population and its healthcare system. Saudi J Gastroenterol **2018**, 24, 196-197, doi:10.4103/sjg.SJG_164_18.
6. Aziz, M.A. Precision medicine in colorectal cancer. Saudi J Gastroenterol **2019**, 25, 139-140, doi:10.4103/sjg.SJG_24_19.
7. Dallol, A.; Buhmeida, A.; Al-Ahwal, M.S.; Al-Maghrabi, J.; Bajouh, O.; Al-Khayyat, S.; Alam, R.; Abusanad, A.; Turki, R.; Elaimi, A., et al. Clinical significance of frequent somatic mutations detected by high-throughput targeted sequencing in archived colorectal cancer samples. J Transl Med **2016**, 14, 118, doi:10.1186/s12967-016-0878-9.
8. Torrent suite version 5.8.
9. McLaren, W.; Gil, L.; Hunt, S.E.; Riat, H.S.; Ritchie, G.R.; Thormann, A.; Flicek, P.; Cunningham, F. The Ensembl Variant Effect Predictor. Genome Biol **2016**, 17, 122, doi:10.1186/s13059-016-0974-4.
10. Ng, P.C.; Henikoff, S. Predicting deleterious amino acid substitutions. Genome Res **2001**, 11, 863-874, doi:10.1101/gr.176601.
11. Adzhubei, I.; Jordan, D.M.; Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet **2013**, Chapter 7, Unit7 20, doi:10.1002/0471142905.hg0720s76.
12. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **2007**, 81, 559-575, doi:10.1086/519795.
13. Lee, S.; Emond, M.J.; Bamshad, M.J.; Barnes, K.C.; Rieder, M.J.; Nickerson, D.A.; Team, N.G.E.S.P.-E.L.P.; Christiani, D.C.; Wurfel, M.M.; Lin, X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet **2012**, 91, 224-237, doi:10.1016/j.ajhg.2012.06.007.
14. Kramer, A.; Green, J.; Pollard, J., Jr.; Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics **2014**, 30, 523-530, doi:10.1093/bioinformatics/btt703.
15. Baker, S.J.; Preisinger, A.C.; Jessup, J.M.; Paraskeva, C.; Markowitz, S.; Willson, J.K.; Hamilton, S.; Vogelstein, B. p53 gene mutations occur in combination with 17p allelic deletions as late events in colorectal tumorigenesis. Cancer Research **1990**, 50, 7717-7722.
16. Fearon, E.R.; Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell **1990**, 61, 759-767.
17. Huang, J.; Papadopoulos, N.; McKinley, A.J.; Farrington, S.M.; Curtis, L.J.; Wyllie, A.H.; Zheng, S.; Willson, J.K.; Markowitz, S.D.; Morin, P., et al. APC mutations in colorectal tumors with mismatch repair deficiency. Proc Natl Acad Sci U S A **1996**, 93, 9049-9054.
18. Jen, J.; Kim, H.; Piantadosi, S.; Liu, Z.F.; Levitt, R.C.; Sistonen, P.; Kinzler, K.W.; Vogelstein, B.; Hamilton, S.R. Allelic loss of chromosome 18q and prognosis in colorectal cancer. N Engl J Med **1994**, 331, 213-221, doi:10.1056/NEJM199407283310401.
19. Morin, P.J.; Vogelstein, B.; Kinzler, K.W. Apoptosis and APC in colorectal tumorigenesis. Proc Natl Acad Sci U S A **1996**, 93, 7950-7954.

20. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A., Jr.; Kinzler, K.W. Cancer genome landscapes. Science **2013**, 339, 1546-1558, doi:10.1126/science.1235122.
21. Aziz, M.A.; Yousef, Z.; Saleh, A.M.; Mohammad, S.; Al Knawy, B. Towards personalized medicine of colorectal cancer. Crit Rev Oncol Hematol **2017**, 118, 70-78, doi:10.1016/j.critrevonc.2017.08.007.
22. Kumar, P.; Henikoff, S.; Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc **2009**, 4, 1073-1081, doi:10.1038/nprot.2009.86.
23. Sharma, Y.; Miladi, M.; Dukare, S.; Boulay, K.; Caudron-Herger, M.; Gross, M.; Backofen, R.; Diederichs, S. A pan-cancer analysis of synonymous mutations. Nat Commun **2019**, 10, 2569, doi:10.1038/s41467-019-10489-2.
24. Foo, J.; Liu, L.L.; Leder, K.; Riester, M.; Iwasa, Y.; Lengauer, C.; Michor, F. An Evolutionary Approach for Identifying Driver Mutations in Colorectal Cancer. PLoS Comput Biol **2015**, 11, e1004350, doi:10.1371/journal.pcbi.1004350.
25. Huang, D.; Sun, W.; Zhou, Y.; Li, P.; Chen, F.; Chen, H.; Xia, D.; Xu, E.; Lai, M.; Wu, Y., et al. Mutations of key driver genes in colorectal cancer progression and metastasis. Cancer Metastasis Rev **2018**, 37, 173-187, doi:10.1007/s10555-017-9726-5.
26. Carethers, J.M.; Jung, B.H. Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer. Gastroenterology **2015**, 149, 1177-1190 e1173, doi:10.1053/j.gastro.2015.06.047.
27. Aghagolzadeh, P.; Radpour, R. New trends in molecular and cellular biomarker discovery for colorectal cancer. World J Gastroenterol **2016**, 22, 5678-5693, doi:10.3748/wjg.v22.i25.5678.
28. Jauhri, M.; Bhatnagar, A.; Gupta, S.; Shokeen, Y.; Minhas, S.; Aggarwal, S. Targeted molecular profiling of rare genetic alterations in colorectal cancer using next-generation sequencing. Med Oncol **2016**, 33, 106, doi:10.1007/s12032-016-0820-2.
29. Song, H.N.; Lee, C.; Kim, S.T.; Kim, S.Y.; Kim, N.K.; Jang, J.; Kang, M.; Jang, H.; Ahn, S.; Kim, S.H., et al. Molecular characterization of colorectal cancer patients and concomitant patient-derived tumor cell establishment. Oncotarget **2016**, 7, 19610-19619, doi:10.18632/oncotarget.7526.
30. Song, E.K.; Tai, W.M.; Messersmith, W.A.; Bagby, S.; Purkey, A.; Quackenbush, K.S.; Pitts, T.M.; Wang, G.; Blatchford, P.; Yahn, R., et al. Potent antitumor activity of cabozantinib, a c-MET and VEGFR2 inhibitor, in a colorectal cancer patient-derived tumor explant model. Int J Cancer **2015**, 136, 1967-1975, doi:10.1002/ijc.29225.
31. Baran, B.; Mert Ozupek, N.; Yerli Tetik, N.; Acar, E.; Bekcioglu, O.; Baskin, Y. Difference Between Left-Sided and Right-Sided Colorectal Cancer: A Focused Review of Literature. Gastroenterology Res **2018**, 11, 264-273, doi:10.14740/gr1062w.
32. Young, G.P.; Rabeneck, L.; Winawer, S.J. The Global Paradigm Shift in Screening for Colorectal Cancer. Gastroenterology **2019**, 156, 843-851 e842, doi:10.1053/j.gastro.2019.02.006.
33. Bian, S.; Hou, Y.; Zhou, X.; Li, X.; Yong, J.; Wang, Y.; Wang, W.; Yan, J.; Hu, B.; Guo, H., et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science **2018**, 362, 1060-1063, doi:10.1126/science.aao3791.

34. Hu, W.; Yang, Y.; Li, X.; Huang, M.; Xu, F.; Ge, W.; Zhang, S.; Zheng, S. Multi-omics Approach Reveals Distinct Differences in Left- and Right-Sided Colon Cancer. Mol Cancer Res **2018**, 16, 476-485, doi:10.1158/1541-7786.MCR-17-0483.
35. Jia, P.; Zhao, Z. Impacts of somatic mutations on gene expression: an association perspective. Brief Bioinform **2017**, 18, 413-425, doi:10.1093/bib/bbw037.
36. Molinari, C.; Marisi, G.; Passardi, A.; Matteucci, L.; De Maio, G.; Ulivi, P. Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine? Int J Mol Sci **2018**, 19, doi:10.3390/ijms19123733.
37. Grizzle, W.E.; Bell, W.C.; Sexton, K.C. Issues in collecting, processing and storing human tissues and associated information to support biomedical research. Cancer Biomark **2010**, 9, 531-549, doi:10.3233/CBM-2011-0183.
38. Hiltemann, S.; Jenster, G.; Trapman, J.; van der Spek, P.; Stubbs, A. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. Genome Res **2015**, 25, 1382-1390, doi:10.1101/gr.183053.114.
39. Kumar, A.; White, T.A.; MacKenzie, A.P.; Clegg, N.; Lee, C.; Dumpit, R.F.; Coleman, I.; Ng, S.B.; Salipante, S.J.; Rieder, M.J., et al. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci U S A **2011**, 108, 17087-17092, doi:10.1073/pnas.1108745108.
40. He, M.M.; Li, Q.; Yan, M.; Cao, H.; Hu, Y.; He, K.Y.; Cao, K.; Li, M.M.; Wang, K. Variant Interpretation for Cancer (VIC): a computational tool for assessing clinical impacts of somatic variants. Genome Med **2019**, 11, 53, doi:10.1186/s13073-019-0664-4.