Original Research Article

Progress in Microbes and Molecular Biology

HH PUBLISHER

Whole exome sequencing identifies genomic alterations in proximal and distal colorectal cancer

Ryia-Illani Mohd Yunos[1], Nurul-Syakima Ab Mutalib[1*], Khor Sheau Sean[2], Sazuita Saidin[1], Mohd Ridhwan Abd Razak[1], Norshahidah Mahamad Nadzir[1], Zuraini Abd. Razak[1], Isa Mohamed Rose[3], Ismail Sagap[4], Luqman Mazlan[4], Nadiah Abu[1], Rahman Jamal[1*]

[1]UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, 56000, Kuala Lumpur, Malaysia
[2]Thermo Fisher Scientific, Blk 33 #07-06, Marsiling Industrial Estate Road 3, 739256 Singapore
[3]Department of Pathology, Faculty of Medicine, Universiti Kebangsaan Malaysia, 56000, Kuala Lumpur, Malaysia
[4]Department of Surgery, Faculty of Medicine, Universiti Kebangsaan Malaysia, 56000, Kuala Lumpur, Malaysia

Abstract: Majority of colorectal cancer (CRC) patients are presented with advanced disease at diagnosis, particularly in cases of proximal CRCs. Little is known about the relationship between the genetic landscape and the anatomical location of the tumour; as well as the prognostication in CRC patients. The objectives of this study were to determine the somatic single nucleotide variants (SNV) and the cellular pathways between the proximal and distal CRCs. Whole exome sequencing was performed on the Ion Proton platform on 10 pairs of normal and CRC samples. The sequencing results were analysed using the Torrent Suite Software and the variants were annotated using ANNOVAR; followed by validation with Sanger sequencing. *APC* is the most frequently altered gene in both proximal and distal CRCs. *KRAS* and *ATM* genes were particularly altered in the proximal CRCs with a frequency of 60% and 40%, respectively. On the other hand, *TP53* mutations did not show any CRC anatomical predominance. There were five recurrent novel variants in proximal CRCs and no recurrent variants identified in distal CRC. Wnt signalling pathway was the most frequently altered pathway in both proximal and distal CRCs whereas TGF-β and PI3K signalling pathways were predominantly altered in the proximal CRCs. We found that proximal CRCs presented with more variants and different altered pathways as compared to distal CRCs. However, further study in a larger series of samples coupled with functional studies will be required to confirm the identified variants and determine their roles in the pathogenesis of proximal and distal CRCs.

Keywords: Colorectal; proximal; distal; whole-exome sequencing; insertion; deletion

***Correspondence:** Nurul-Syakima Ab Mutalib, UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, 56000, Kuala Lumpur, Malaysia; syakima@ppukm.ukm.edu.my. Rahman Jamal, UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia, 56000, Kuala Lumpur, Malaysia; rahmanj@ppukm.ukm.edu.my.

Received: 12th August 2019
Accepted: 13th September 2019
Publish Online: 30th September 2019

Citation: Mohd Yunos R-I, Ab Mutalib N-S, Khor SS, *et al*. Whole exome sequencing identifies genomic alterations in proximal and distal colorectal cancer. Prog Microb Mol Biol, 2019; 1(1): a0000036

Introduction

Colorectal cancer (CRC) is the third most commonly diagnosed cancer worldwide[1] and is ranked as the second most common cancer in Malaysia[2]. According to the National Cancer Registry Report, CRC is the most common cancer among men and the third most common among women in Malaysia[3]. In the year 2018, it was estimated that there were 1.8 million cases and 881,000 deaths from CRC worldwide[1].

Anatomically, CRCs are classified into three subsets named, proximal, distal and rectum. Proximal CRC is located on the right side of the colon, which includes cecum, ascending colon, hepatic flexure, transverse colon and splenic flexure; while distal CRC is located on the left side of the colon, including sigmoid and descending colon[4]. It is postulated that both the proximal and distal CRCs are anatomically different and arose from different biological pathways, suggesting different molecular mechanisms involved[5]. Biological differences between the normal proximal and distal colon suggest that the carcinogenesis in these locations may be mediated via different molecular pathways[6,7]. This may have profound prognostic and therapeutic implications. For instance, it has been reported that the gene expression profile between adenocarcinoma of the proximal and distal CRC is different, and therefore, the information

should be taken into consideration when investigating new predictive and prognostic biomarkers[8]. Among the molecular characteristics that distinguish between proximal and distal CRCs is the low frequency of *TP53* and *KRAS* gene mutations, lower c-MYC expression and DNA mismatch repair (MMR) deficiency in proximal CRC. Distal CRC, on the other hand, shows a higher frequency of *TP53* and *KRAS* gene mutations and c-MYC expression[9].

Several large cohort studies in the last 20 years have demonstrated that the proximal and distal CRCs differed in their susceptibility to screening tests, the stage at which they were diagnosed, patient characteristics, pathology and prognosis[10,11]. A few studies have discovered that the distribution of distal and proximal CRCs varies according to ethnicity[12]. Individuals with African ancestry are more likely to develop proximal CRCs rather than distal CRCs[13,14], whereas Asians and Pacific Islanders are more prone to distal colon and rectal cancers[13]. A study by Goh and colleagues (2005) supported this fact, whereby 77% of CRC cases in Malaysia were diagnosed as distal CRCs[15]. However, despite the different distributions, the actual causative molecular events that lead to the different prognosis between proximal and distal CRCs remains poorly understood.

Next-generation sequencing technologies have revolutionized cancer research and management, particularly in diagnostic and treatment strategies[16,17]. Using whole-exome sequencing approach, we examined the exomes of Malaysian patients diagnosed with proximal and distal CRCs in order to identify the differences and the distinct signatures based on their anatomical distribution. We hope that this can contribute to better management and treatment of CRC patients in the future.

Materials and Methods

Clinical material

A total of ten paired colorectal carcinoma and their corresponding adjacent normal tissues (five proximal and five distal), were collected from patients at Universiti Kebangsaan Malaysia Medical Centre, Kuala Lumpur. The written informed consents were provided by patients to participate in the study. This study was approved by the UKM Research Ethics Committee (Reference no: UKM 1.5.3.5/244/UMBI-004-2012). The tissues were subjected to cryosectioning; followed by haematoxylin and eosin (H&E) staining. Our pathologist confirmed the presence of at least 80% of cancer cells in the tumour specimens and less than 20% necrosis in paired unaffected colorectal tissue adjacent to the tumour site. The DNA extraction was performed using the QIAamp DNA mini kit (QIAGEN, Valencia, CA) according to the manufacturer's protocol. Quality and quantity of the extracted DNA were assessed using Qubit Fluorometer (Invitrogen, Carlsbad, CA, USA), NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE); as well as agarose gel electrophoresis. To confirm the identity of each tumour and normal paired samples, genotyping was performed by multiplex PCR based on microsatellite polymorphisms using Coriell Identity Mapping Kit (Coriell Institute, New Jersey, USA).

Exome capture, library construction and next-generation sequencing

One microgram (1 µg) of genomic DNA from each pair of tumour and normal sample was mechanically sheared by ultrasonic fragmentation using the Covaris® System to achieve fragments of about 50–500 bp. The fragmentation profile was assessed by the Bioanalyzer High Sensitivity DNA Analysis Kit (Agilent Technologies, Carlsbad, CA, USA). The fragmented DNA was used to construct a fragment library using the Ion Plus Fragment Library Kit (Life Technologies, Guilford, CT) according to the manufacturer's instructions for ligation, end repair, purification, size selection and final amplification prior to exome capture. For multiplexing the samples, adapters with short stretches of index sequences from Ion Xpress™ Barcode Adapters 1–16 kit was used and thus allowing the sequencing of two samples in a single Ion PI™ chip run. Five hundred nanograms (500 ng) of the amplified, size-selected library (~285 bp) from each sample was subjected to exome capture procedure using the Ion TargetSeq™ Exome kit (Life Technologies, Guilford, CT) according to manufacturer's protocol. Exomes were captured using ~2 million TargetSeq™ capture probes with biotinylated oligos that range from ~50 bases to ~120 bases. The captured DNA fragments were isolated using streptavidin-coated Dynabeads paramagnetic beads and they were amplified and purified. Finally, the samples were quantitated and qualitatively assessed on the Bioanalyzer High Sensitivity DNA chip (Agilent Technologies, Carlsbad, CA, USA). The purified, 10 pM of exome-enriched libraries were used for template preparation on Ion PI™ Ion Sphere™ Particles (ISPs) for sequencing on an Ion PI™ Chip using Ion Proton sequencer (Life Technologies, Guilford, CT).

Bioinformatics analyses

The data was processed using the Torrent Suite software v4.2.1. The Torrent Suite software automated the generation of sequence reads, trimming of adapter sequences, removal of poor quality reads; as well as sequence alignment to the hg19 human genome reference. Variants were called using the Torrent Variant Caller plugin, configured for somatic mutation detection with low stringency setting. Variant filtering and calculation of transition to transversion ratio (Ti/Tv) were performed using SnpSift tool in SnpEff (Version 4.0, 2014-11)[18]. The variants were then subjected to annotation using ANNOVAR, Version 2013May09[19]. Gene-based annotation was performed against RefSeq Gene[20], UCSC Known Gene[21] and ENSEMBL Gene[22]. The variants were further annotated against the conserved region (phastConsElements46way)[23], alternative allele frequency in all subjects in the National Heart Lung and Blood Institute Exome Sequencing Project (NHLBI-ESP) project with 6500 exomes (esp6500si_all)[24], alternative allele frequency data in 1000 Genomes Project (1000g2014oct_all) (The 1000 Genomes Project Consortium, 2015), the Exome Aggregation Consortium (ExAC 01)[25], dbSNP version 138 (snp138)[26], CLINVAR (clinvar_20140929)[27] and COS-

MIC version 70 (cosmic70)[28]. Protein impact prediction was also performed on ANNOVAR (Version 2013May09) using command ljb26_all[19] which include SIFT[29] and PolyPhen2[30]. In order to identify potentially druggable variants, the drug-gene interaction database, DGIdb 2.0 was utilized to annotate the variants against drugs genes interaction dataset (http://dgidb.genome.wustl.edu/)[31].

Variants prioritization

Variants exclusion criteria included those with base quality less than Q30, not resulting in amino acid changes, identified in unannotated genes (unknown), called in both tumour and normal exomes, representing probable mapping ambiguities, and have minimal allele frequency (MAF) of more than 5% in the 1000 Genomes Project[32], ExAC and ESP6500 database[25]. A variant in a tumour was considered to be a novel true candidate somatic mutation if the corresponding normal sample has at least 10 reads covering this position, zero variant reads and has not been reported in dbSNP138[26] or the 1000 Genomes data set (October 2014)[32]. For the resulting candidate of somatic mutations, the alignment of each sample was manually examined to check for sequencing artefacts and alignment errors using the Integrated Genomic Viewer (IGV)[33,34]. We were then assessed for potential protein impact prediction of each genetic variant identified in both proximal and distal cohorts based on variant prediction algorithms, SIFT[29,35] and PolyPhen2[30].

Statistical Analysis

We utilized the Fisher's exact test to define significance values in a number of protein-altering mutations and number of affected pathways between proximal and distal CRC using the 2 × 2 contingency tables and the GraphPad QuickCalcs Online Calculator for Scientists (http://www.graphpad.com/quickcalcs/index.cfm). All p values are two-sided and statistical significance is denoted by $p < 0.05$.

Gene pathway analysis

Altered genes identified in both proximal and distal cohorts were pooled and run through the Ingenuity Pathway Analysis (IPA) (Qiagen, Valencia, CA) software at the same time to identify the affected canonical pathways.

Variant confirmation

Variants were validated using the Sanger sequencing method on tumour and normal samples. Primers corresponding to the selected locations were designed using IDT-DNA Primer Quest (Coralville, IA). PCR products were generated and cycle sequencing was performed using the Big Dye Terminator V3.1 reagent (Life Technologies, Guilford, CT). The cycle sequencing products were then processed using ethanol precipitation and sequencing was carried out using the ABI 3130xl capillary electrophoresis (Life Technologies, Guilford, CT). The results were analysed using the Basic Local Alignment System Tool (BLAST)[36] and Sequence Scanner (Applied Biosystem, Foster City, CA).

Results

Clinicopathological characteristic of patients

The characteristics of all ten patients were listed in Table 1. With regards to cancer stage, 20% (n=2) of the patients were of Dukes' A, 40% (n=4) were Dukes' B and the remaining 40% (n=4) were Dukes' C. The average age of patients was approximately 69 years old (range 58–75 years). The samples comprised of three well-differentiated adenocarcinomas, three moderate differentiated adenocarcinomas and four poorly differentiated adenocarcinomas. From these ten samples, four patients were positives for lymph nodes metastasis and six were negative.

Exome sequencing analysis and coverage

The capture regions covered by Ion TargetSeq was about 37.3 Mb and we managed to obtain an average of 39.6 million reads. Average coverage of about 70X for each sample was obtained and the coverage of the target region at 20X was more than 85%. This was comparatively higher than what was obtained by a study on pancreatic cancer[37] and a study on colon and rectal cancers[38]. On average, the number of variants detected at Q30 for each sample was 35, 713 (32, 804 - 37,487 variants) (Table 2). To assess the quality of variants, the ratio of the number of transitions to the number of transversions was determined. The expected Ti/Tv ratio for exome target regions is 2.8[39]. However, the target regions of exome capture kits often covered more than just exons. For SNPs resided in these target regions, Ti/Tv ratios between 2.0 and 3.0 were observed[40]. The target regions of Ion TargetSeq kit covered both exonic and non-exonic regions, such as 3' UTR and 5' UTR, therefore, we obtained an average Ti/Tv ratio of 2.7.

Upon variants prioritizations, we obtained a total of 4,835 and 4,177 variants in proximal and distal CRC, respectively. Among all the variants found in proximal CRC, 539 were protein-altering variants in 508 genes located in the conserved regions. On the other hand, the distal CRCs had 245 protein-altering mutations in 180 genes located in the conserved regions. The proximal CRC showed significantly more protein-altering variants as compared to distal CRC (p-value = 0.0001) (Table 3).

Based on mutation rates, we stratified the cases into two groups: hypermutated (>30 per 10^6 bases) and non-hypermutated (<20 per 10^6 bases). The average mutation rate for proximal and distal CRC is 22 per 10^6 bases and 26 per 10^6 bases respectively (median mutation rate in both proximal and distal is 15 per 10^6 bases). One case of each proximal (sample R4T, 72 per 10^6 bases) and distal (sample L2T, 58 per 10^6 bases) were classified as hypermutated. We then analysed the MMR gene mutations (*MSH2, MSH3, MSH6* and *MLH1*) in all samples. Interestingly, we identified at least one somatic missense mutation in MMR gene presence in sample R4T (*MSH2* g.47656969C>T) and L2T (*EPCAM* g.47604176C>T), that possibly explain the hypermutation status of the samples. On the other hand, none of the rest of the samples was having a mutation in MMR genes.

Table 1. Clinicopathological characteristics of CRC patients

Sample ID	Gender/ Age	Ethnic	Histological Subtype	Stage	
				Dukes'	TNM Staging
L1	M/71	Chinese	Well Differentiated, Adenocarcinoma	B2	T3 N0 Mx
L2	M/66	Chinese	Moderately Differentiated, Adenocarcinoma	A	T2 N0 Mx
L3	M/75	Malay	Poorly differentiated, Adenocarcinoma	C	T3 N1a Mx
L4	F/73	Malay	Well Differentiated, Adenocarcinoma	C1	T2 N1b Mx
L5	F/75	Malay	Well Differentiated, Adenocarcinoma	A	T2 N0 Mx
R1	F/65	Chinese	Poorly Differentiated, Mucinous Adenocarcinoma	B	T3 N0 Mx
R2	M/70	Chinese	Poorly Differentiated, Adenocarcinoma	B2	T3 N0 Mx
R3	M/58	Malay	Moderately Differentiated, Adenocarcinoma	C	T3 N1a Mx
R4	M72	Chinese	Poorly Differentiated, Adenocarcinoma	B	T3 N0 Mx
R5	M/64	Chinese	Moderately Differentiated, Adenocarcinoma	C	T2 N1 Mx

L= Left, R= Right, M= Male, F= Female, TNM= Tumor-Node-Metastasis

Table 2. Exome sequencing coverage for normal and tumour samples

Sample	Total Aligned Output (G)	# of mapped reads	% Reads On Target	Average Coverage	Uniformity (%)	Target Base Coverage at 1x (%)	Target Base Coverage at 20x (%)	Target Base Coverage at 100x (%)	Target Base Coverage at 500x (%)	Variants
L1N	13.7	38,327,122	80.75	71.39	93.48	98.03	90.72	20.66	0.11	34,706
L1T	13.7	43,581,666	84.79	84.68	93.78	98.17	92.64	29.79	0.21	35,675
L2N	13.6	37,248,504	79.6	67.21	94.07	98.5	90.28	16.84	0.14	37,472
L2T	13.6	50,527,425	76.29	88.39	94.40	98.71	93.55	30.66	0.35	37,487
L3N	14.1	45,586,546	83.14	86.35	93.33	98.08	92.29	31.86	0.21	36,077
L3T	14.1	41,217,844	82.66	77.71	93.21	98.01	91.12	25.17	0.19	35,358
L4N	10.7	38,273,507	77.96	62.24	92.97	97.85	88.15	14.38	0.1	34,197
L4T	10.7	33,604,520	78.61	55.4	92.77	97.79	85.53	10.65	0.07	33,519
L5N	12.3	41,524,901	78.49	71.81	93.48	98.04	90.67	20.48	0.16	35,113
L5T	12.3	37,647,763	79.99	66.22	93.04	97.98	88.90	17.12	0.13	34,563
R1N	12.5	40,760,456	81.92	79.26	93.65	97.95	92.03	26.01	0.17	35,213
R1T	12.5	32,407,582	82.14	62.67	92.85	97.76	87.69	15.10	0.12	32,804
R2N	12.7	36,590,333	86.11	73.47	93.83	98.12	91.43	21.70	0.15	34,752
R2T	12.7	40,262,039	86.99	80.74	93.87	98.22	92.34	26.17	0.21	34,862
R3N	11.9	38,279,157	73.01	65.69	93.76	98.06	90.18	16.24	0.09	34,870
R3T	11.9	31,589,214	79.78	58.71	93.51	97.91	88.09	11.93	0.07	35,274
R4N	14.6	40,855,511	77.97	72.79	93.01	98.02	90.25	22.26	0.12	34,516
R4T	14.6	47,748,666	78.07	85.35	93.35	98.11	92.27	31.51	0.18	39,337
R5N	11.6	37,243,895	75.27	60.64	93.54	98.04	88.53	13.48	0.06	34,497
R5T	11.6	37,455,362	75.79	62.11	92.59	97.82	87.48	15.11	0.07	33,175
Mean				**71.64**	**93.42**	**98.06**	**90.21**	**20.86**	**0.15**	**35,173**

L= Left (Distal), R= Right (Proximal), N= Normal, T= Tumour

Table 3. Number of protein altering mutation and number of affected pathway

	Proximal CRC	Distal CRC	P Value
Number of Protein Altering Mutations	539	245	0.0001
Number of Affected Pathways	5	3	0.66

Profile of mutated CRC-related genes and altered pathways

We profiled mutated genes in our set of discovery patients based on thirty-three (33) genes that have been reported in the tumorigenesis of CRCs, which is a compilation from ten publications[41–50]. We plotted each altered gene which was detected in our patients in Figure 1A. Altered genes were defined as any gene that has at least one or more protein-altering mutations. Fifteen (15) out of the 33 CRC-related genes were altered in proximal cancers (29 mutations) and five CRC-related genes were altered in distal cancers (eight mutations) as listed in Table 4. We discovered that some of the nucleotide changes may lead to multiple protein amino acid changes. This is corresponding to different transcript isoforms. To explore the affected pathways and the differences between the two subsites of CRC, we focused on major pathways involved in CRC tumorigenesis[42,43,46–50].

This approach was adapted from Ashktorab *et al*.[51]. An affected pathway is defined when at least one or more genes are altered in any pathway (Figure 1B). We observed that 90% (n=9) of the CRC patients shared an affected Wnt signalling pathway with five genes being altered (12 mutations). RTK-RAS and TP53 signalling pathways were also found to be altered in both proximal and distal CRCs with six mutations in both pathways. We also discovered that the TGF-Beta signalling (four mutations) and PI3K signalling (two mutations) pathways were exclusively altered in proximal CRCs. Overall, there were more altered pathways in proximal as compared to distal CRCs. However, the number of different affected pathways between these two groups were not significant (p=0.66) (Table 3). Since this is discovery research, the lack of significance could be due to our small sample size. Analysis using Ingenuity Pathway Analysis (IPA) software identified the Wnt signalling and growth factor signalling pathways as the most commonly affected (Figure 3).

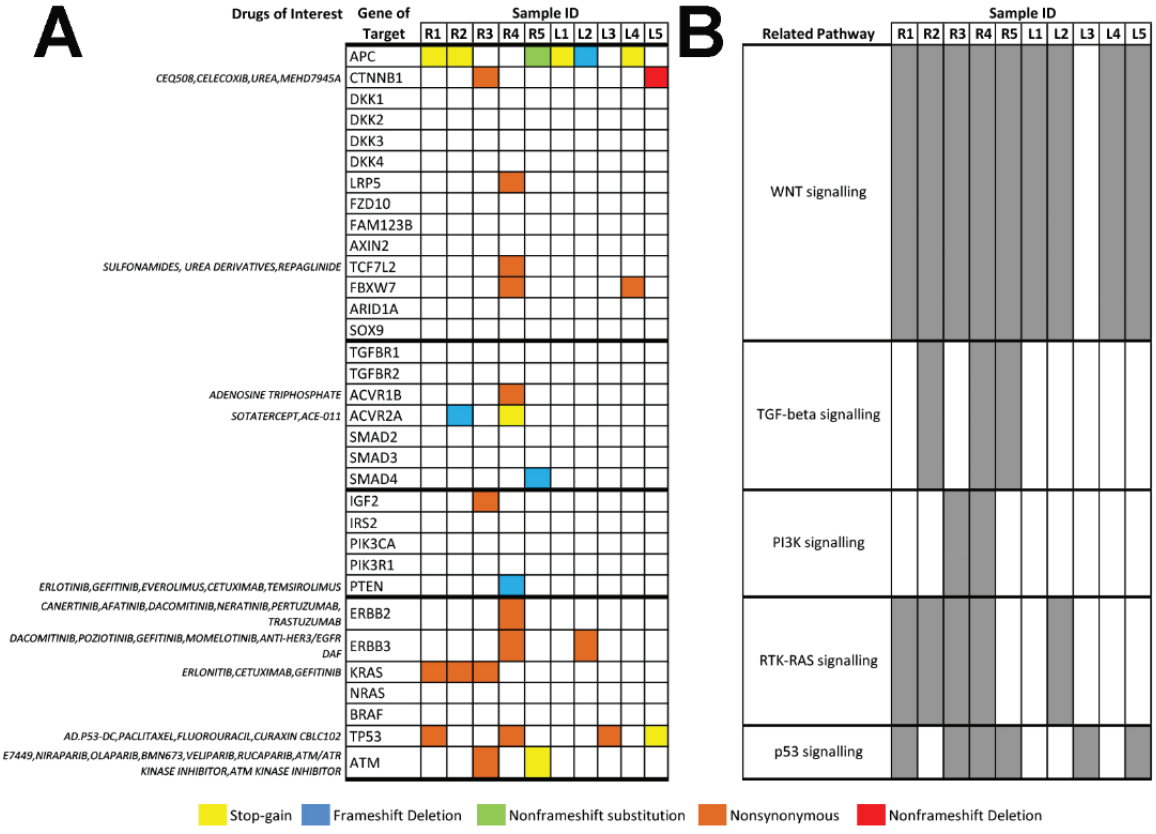

Figure 1. Altered genes and pathway implicated in ten CRC patients (five patients of each proximal and distal CRCs). (A) The genes associated with existing treatment option or novel targeted therapies currently being investigated in clinical trials. (B) Major signalling pathways altered in CRCs.

Table 4. Novel and known mutations in CRC-related genes in proximal and distal CRCs.

Sample ID	Gene	Start	End	Ref	Alt	Protein Change	Exonic Function	dbSNP ID	COSMIC ID	Known / Novel
R1T	*APC*	112175255	112175255	G	T	E1304X E1322X	Stop gain	NA	COSM18702	Known
	TP53	7578550	7578550	G	A	S88F S127F	Nonsynonymous	NA	COSM216414 COSM3378368 COSM44226 COSM216412 COSM216413 COSM1637542	Known
R2T	*TP53*	7577022	7577022	G	A	R174X R267X R306X	Stop gain	rs121913344	COSM3388168 COSM10663 COSM1640820 COSM99947	Known
	KRAS	25398284	25398284	C	T	G12D	Nonsynonymous	rs121913529	COSM521 COSM1135366	Known
	KRAS	25398284	25398284	C	T	G12D	Nonsynonymous	rs121913529	COSM521 COSM1135366	Known
	ACVR2A	148683686	148683686	A	-	K327fs K435fs	Frameshift deletion	NA	COSM252949	Known
	APC	112175639	112175639	C	T	R1432X R1450X	Stop gain	rs121913332	COSM13127	Known
R3T	*ATM*	108141828	108141828	A	G	Y959C	Nonsynonymous	NA	NA	Novel
	CTNNB1	41266136	41266136	T	C	S45P	Nonsynonymous	rs121913407	COSM5663	Known
	CTNNB1	41277302	41277302	A	G	R591G	Nonsynonymous	NA	NA	Novel
	IGF2-AS	2167618	2167618	A	C	T150P	Nonsynonymous	NA	NA	Novel
	KRAS	25380283	25380283	C	T	A59T	Nonsynonymous	rs121913528	COSM546 COSM1562187	Known
R4T	*ACVR1B*	52387841	52387841	C	T	R489C R437C R530C	Nonsynonymous	NA	NA	Novel
	ACVR1B	52379132	52379132	G	A	R379Q R327Q R420Q	Nonsynonymous	NA	NA	Novel

R4T	*ACVR2A*	148683718	148683718	G	A	W337X W445X	Stop gain	NA	NA	Novel
	ERBB2	37879658	37879658	G	A	R678 R648Q R663Q	Nonsynonymous	NA	COSM436498	Known
	ERBB2	37864598	37864598	G	A	V84M V54M V69M	Nonsynonymous	rs376524324	NA	Known
	ERBB3	56495023	56495023	G	A	R1127H	Nonsynonymous	rs2271188	COSM1363018 COSM1363017	Known
	ERBB3	56478854	56478854	G	A	V104M	Nonsynonymous	NA	COSM172423 COSM20710 COSM1152549	Known
	FBXW7	153249456	153249456	C	T	R323Q R361Q R441Q	Nonsynonymous	NA	COSM1052092 COSM1052093 COSM1052091 COSM1052094	Known
	LRP5	68201247	68201247	G	A	R733H R1314H	Nonsynonymous	NA	NA	Novel
	PTEN	89720812	89720812	A	-	T321fs	Frameshift deletion	NA	COSM5823	Known
	TCF7L2	114925436	114925436	G	A	R482Q R505Q R499Q	Nonsynonymous	NA	NA	Known
	TP53	7577138	7577138	C	T	R135Q R108Q R228Q R267Q	Nonsynonymous	NA	COSM43923 COSM3691863 COSM1290766 COSM3691864	Known
	TP53	7578458	7578458	G	A	R26C R119C R158C	Nonsynonymous	NA	COSM1750371 COSM984954 COSM984957 COSM3932746 COSM984958 COSM984956 COSM43848	Known
R5T	*APC*	112176559	112176559	TT	GC	5214_5215GC 5268_5269GC 5268_5269GC	Nonframeshift Substitution	NA	NA	Novel
	ATM	108106477	108106477	G	T	G138X	Stop gain	NA	NA	Novel
	SMAD4	48573529	48573536	GAG-CAATT	-	38_40del	Frameshift deletion	NA	NA	Novel
L1T	*APC*	112162896	112162896	T	G	Y482X Y500X	Stop gain	NA	NA	Novel
L2T	*APC*	112175212	112175216	AAAAG	-	1289_1291del 1307_1309del	Frameshift deletion	NA	COSM18764	Known
	ERBB3	56478854	56478854	G	C	V104L	Nonsynonymous	NA	COSM160824	Known

L3T	*TP53*	7577121	7577121	G	A	R141C R114C R234C R273C	Nonsynonymous	rs121913343	COSM3355991 COSM10659 COSM1645518 COSM99933	Known
L4T	*APC*	112175322	112175322	C	G	S1326X S1344X	Stop gain	NA	NA	Novel
	FBXW7	153249385	153249385	G	A	R347C R385C R465C	Nonsynonymous	NA	COSM1154293 COSM170727 COSM170726 COSM170725 COSM22932	Known
L5T	*CTNNB1*	41266071	41266100	GT-CACTG-GCAG-CAA-	-	23_33del	Nonframeshift Deletion	NA	NA	Novel
	TP53	7578263	7578263	G	A	R64X R37X R157X R196X	Stop gain	NA	COSM1640847 COSM99667 COSM99666 COSM99668 COSM3378446 COSM99665 COSM10705	Known

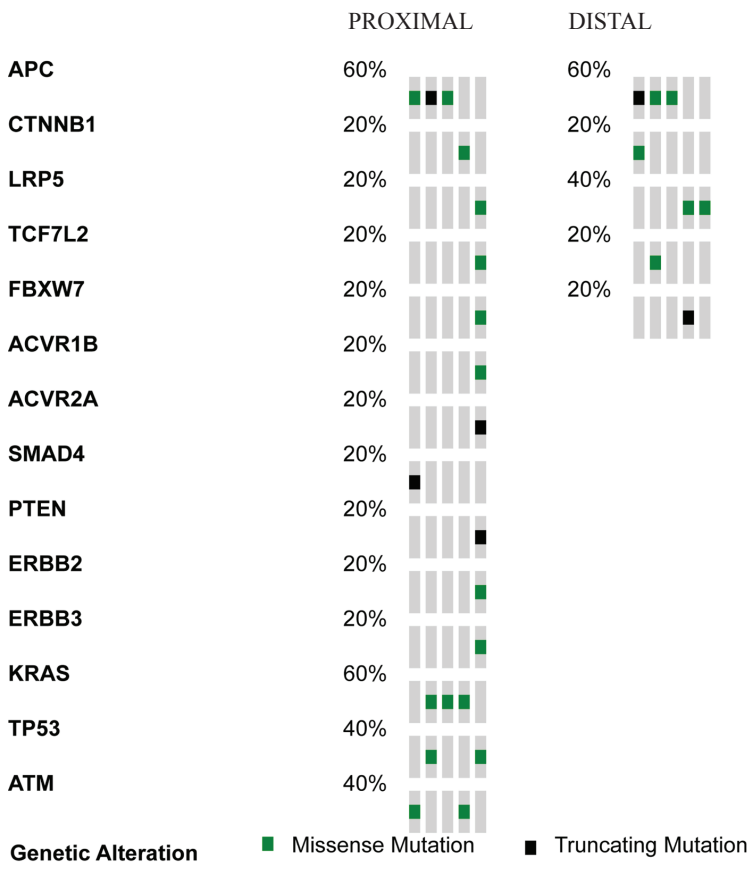

Figure 2. Diagram illustrating the number of patients with alteration in the CRC-associated genes.

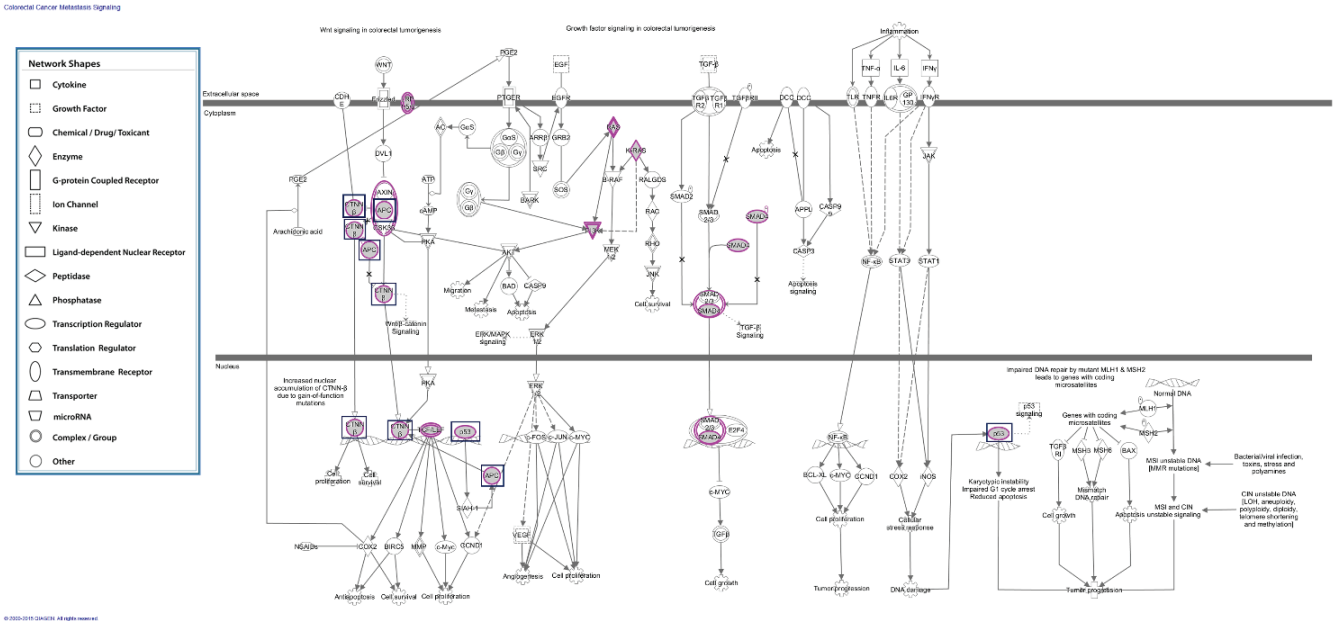

Figure 3. The most commonly mutated canonical pathway. The key mutated genes detected in proximal CRC are highlighted in pink while mutated genes detected in distal CRCs are highlighted in blue.

Most frequently altered genes in proximal and distal CRCs

We discovered that the *APC* gene (six mutations) was the most frequently mutated gene with a frequency of 60% (n=3) in both proximal and distal CRCs. The second most frequently mutated gene in both proximal and distal CRC was *TP53* gene (six mutations) with the frequency of 40% (n=2). This suggested that *APC* and *TP53* did not have predominance for either side of the colon. Interestingly, we found that the *KRAS* gene (three mutations) and *ATM* gene (two mutations) were uniquely altered in proximal CRCs with the frequency of 60% and 40% respectively (Figure 2). Among all of the frequently altered genes in this set of analysis, we identified three novel mutations in *APC* gene, namely, a non-frameshift substitution (*APC* g.112176559TT>GC) in tumour R5T, stop-gain (*APC* g.112162896T>G) in tumour L1T, stop-gain (*APC* g. 112175322C>G) in tumour L4T and two *novel* mutations in *ATM* gene of tumour R3T (*ATM* g.108141828A>G) and R5T (*ATM* g.108106477G>T) (Table 5). Sanger sequencing successfully confirmed each of the 17 somatic mutations we had detected in these four genes.

Recurrent variants and mutated genes

In this discovery set of ten Malaysian CRC patients, five genes with one mutation were demonstrated in at least two individuals of proximal CRC patients. On the other hand, there is no recurrent mutation identified in individuals of distal CRCs (Table 6). Detailed analysis revealed that the minimum reads covering the mutated allele were 35 reads with a minimum of 1% of the total reads containing the alternate base. This was found in the variant presented in *C9orf50* gene. The variant in the *KRAS* gene (*KRAS* g.25398284C>T) had the highest variant coverage with 121 reads and 52 reads (63%) contained the variant in sample R2T. The same variant was detected in sample R5T with 96 reads and 48 of the reads (50%) contained the alternate base.

Actionable alterations in proximal and distal CRC

We looked at the drug-gene interaction by annotating against the Drug Gene Interaction Database (DGIdb) and identified ten genes implicated in CRC tumorigenesis that are clinically relevant, including targets of new and existing therapies and genes. Twenty-one (21) variants in ten genes and eight variants in three genes were identified in proximal and distal CRCs, respectively. Notably, 80% (8/10) of CRC patients harboured at least one actionable alteration (range one to seven alterations) that has been linked to a clinical treatment option or is currently being investigated in clinical trials for novel targeted therapies. For example, in the present study, 5FU-based chemotherapy is considered to target patients with wild type *TP53*, which includes patients R2, R3, R5, L1, L2 and L4 (Figure 1A).

Validation against TCGA Data

We performed external validation using somatic mutation data from 618 patients (consisting of proximal and distal CRC patients) in the TCGA data set[38]. We successfully validate that *ZNF337* and *c9orf50* genes, both are recurrently mutated genes identified from our discovery data set, were exclusively mutated in proximal patients in TCGA patients. While *GPR6* gene was found to be less frequently mutated in TCGA patients (about 5%), we discovered that two out of five patients in our proximal patients (40%) harboured one known *GPR6* mutations (*GPR6* g. 110301081G>A). We performed a comparison between mutation frequency in proximal versus distal CRC of TCGA patients. Subsequently, we profiled our patients based on the significantly mutated genes from the TCGA data set. We discovered that at least eight genes were found to be predominantly mutated in proximal CRC (p=0.0032) (Table 7).

Table 5. Novel and known mutations in most frequently mutated genes in proximal and distal

Gene/ Chr	Sample ID	Start	End	Ref	Alt	Protein Change	Exonic Function	dbSNP ID	COSMIC ID	Known / Novel
***APC*/ Chr5**	R1T	112175255	112175255	G	T	R1432X R1450X	Stop gain	NA	COSM18702	Known
	R2T	112175639	112175639	C	T	5214_5215GC 5268_5269GC	Stop gain	rs121913332	COSM13127	Known
	R5T	112176559	112176560	TT	GC	E1304X E1322X	Nonframeshift substitution	NA	NA	Novel
	L1T	112162896	112162896	T	G	Y482X Y500X	Stop gain	NA	NA	Novel
	L2T	112175212	112175216	AAAAG	-	1289_1291del 1307_1309del	Frameshift Deletion	NA	COSM18764	Known
	L4T	112175322	112175322	C	G	S1326X S1344X	Stop gain	NA	NA	Novel
***TP53*/ Chr17**	R1T	7578550	7578550	G	A	S127F S88F	Nonsynonymous	NA	COSM216414	Known
	R1T	7577022	7577022	G	A	R174X R147X R306X R267X	Stop gain	rs121913344		Known
	R4T	7577138	7577138	C	T	R135Q G323A R108Q R267Q R228Q	Nonsynonymous	NA	COSM43923	Known
	R4T	7578458	7578458	G	A	R26C R158C R119C	Nonsynonymous	NA		Known
	L3T	7577121	7577121	G	A	R141C R273C R234C	Nonsynonymous	rs121913343		Known
	L5T	7578263	7578263	G	A	R64X R37X R196X R157X	Stop gain	NA		Known
***KRAS*/ Chr 12**	R1T	25398284	25398284	C	T	G12D	Non synonymous	rs121913529	COSM521	Known
	R2T	25398284	25398284	C	T	G12D	Non synonymous	rs121913529	COSM521	Known
	R3T	25380283	25380283	C	T	A59T	Non synonymous	rs121913528	COSM546	Known
***ATM*/ Chr11**	R3T	108141828	108141828	A	G	Y959C	Nonsynonymous	NA	NA	Novel
	R5T	108106477	108106477	G	T	G138X	Stop gain	NA	NA	Novel

NA = Not Available

Table 6. Recurrent variants and mutated genes in proximal CRC

Sample ID	Gene/ Chr	Start	End	Ref	Alt	Protein Change	Exonic Function	dbSNP ID	COSMIC ID	Known / Novel
R2T, R5T	*C9orf50*/ Chr9	132377900	132377900	C	-	R248fs	Frameshift Deletion	NA	NA	Novel
R1T, R4T	*GPR6*/ Chr6	110301081	110301081	G	A	A256T, A271T	Nonsynonymous	NA	COSM3429854 COSM3429853	Known
R1T, R2T	*KRAS*/ Chr12	25398284	25398284	C	T	G12D	Nonsynonymous	rs121913529	COSM521 COSM1135366	Known
R3T, R4T	*ZNF337*/ Chr20	25657029	25657029	G	A	R299X	Stop gain	NA	NA	Novel
R3T, R4T	*ZNF783*/ Chr7	148963763	148963763	G	A	R121H	Nonsynonymous	NA	NA	Novel

NA = Not Available

Table 7. Validation of mutation frequency in proximal and distal CRC of TCGA patients

	This study			TCGA CRC		
Gene	**Proximal CRC (%)**	**Distal CRC (%)**	**p-value**	**Proximal CRC (%)**	**Distal CRC (%)**	**p-value**
KRAS	60	0	0.0001	29	26	n.s
ATM	40	0	0.0001	10	4	n.s
ZNF337	40	0	0.0001	9	0	0.003
C9ORF50	40	0	0.0001	3	0	n.s
GPR6	40	0	0.0001	5	2	n.s
ZNF783	40	0	0.0001	4	1	n.s
PIK3CA	20	0	0.0001	28	14	0.02
ACVR2A	40	20	0.0032	15	3	0.005
TP53	40	60	0.0071	40	61	0.004

n.s = not significant

Discussion

CRC is a heterogeneous disease with the genetic landscape and clinical outcomes depend on the anatomic location of cancer. Many efforts have been made to unveil the genetic alterations and molecular features of colorectal cancer[38,51]. Studies show that these alterations would determine the prognosis and response to treatment[8,52,53]. Nevertheless, how the anatomical location could have an impact on the molecular features and more importantly, the prognosis, is unknown.

In this study, we analyzed somatic alterations between proximal and distal CRC in Malaysian patients. It has provided an insight into the identification of known and novel somatic mutations that suggest a relationship between the genomic alterations, cellular pathways, actionable genes and anatomical location of the tumour. Overall, we discovered that proximal CRCs exhibited a higher number of somatic mutations and altered pathways as opposed to distal CRCs. Proximal CRCs has been proven to have increased mutational burden, with higher rates of microsatellite instability as compared to distal colon and rectal cancers[54]. These results may account for the poor prognosis of proximal CRC patients.

Based on CRC TCGA data published in 2012 (accessed through cBioPortal), 16.1% of distal CRC and 47.1% proximal CRC were microsatellite instable (MSI), while 83.2% of distal and 52.7% of proximal CRC were microsatellite stable (MSS)[55]. Even though generally, MSI CRC accounted for approximately 15% of sporadic CRCs, the frequency of MSI was higher in proximal CRCs[55]. MSI cancers were shown to have eight times more somatic nonsynonymous variants than MSS cancers[56]. However, we are unable to perform MSI determination status due to the lack of tissue for staining.

The Wnt signalling and EGFR pathways are among the commonly affected pathways in CRC tumorigenesis[57,58]. To identify possible pathway differences in proximal and distal, we performed pathway analysis using IPA. We discovered that the most enriched pathways are the Wnt and the growth factor signalling pathways. Ninety per cent (90%) of our patients in this discovery set, irrespective of their anatomical location, had a mutation in one or more members of the Wnt signalling pathway, predominantly in *APC*. The frequency of mutated APC gene across proximal and distal CRC patients in our study is 60% (Figure 2). This is in concordance with recent findings where they discovered over 50% of the recruited patients in the study exhibited altered Wnt signalling pathway with APC being the most significantly mutated gene[38,59].

Wnt signalling pathway activation is required for maintenance of colorectal tumours harbouring *APC* mutations[60]. Inactivating *APC* mutations occur in about 85% of CRCs, resulting in β-catenin stabilization and increased signalling through the Tcf/Lef transcription factors. Mutant β-catenin is free to enter the nucleus and constitutively activates transcription through Tcfs[61]. β-catenin inhibition *in vivo* strongly inhibited the growth of established *APC*-mutant colorectal tumour xenografts[60]. In our small set of patients, those with wild type APC gene (R3T and L5T) harboured at least one mutation in the *CTNNB1* gene (Figure 1A), resulting in altered Wnt signalling pathway. It has been shown that up to 50% of CRC with wild type *APC* gene were found to have *CTNNB1* mutations[62].

From our analyses, we also found that the TP53 signalling pathway was altered in both proximal and distal CRCs. With a frequency of 40%, *TP53* is the second commonest altered gene in our study. Similarly, TP53 gene mutations in an Iranian cohort of CRC patients occurred as frequent as in other studies with equal distribution, suggesting no differences across the anatomic location[63,64]. However, our findings were in contrast with another study which different mutation spectra of *TP53* was observed depending on proximal or distally located tumour[65]. Alteration of *TP53* may have different prognostic significance depending on the ethnic group[66], an anatomic subsite of the colon[65,66] and stage of disease[67,68]. There is convincing evidence that patients with wild-type *TP53* gained survival benefit from the use of 5-fluorouracil (5FU)-based chemotherapy[69]. Conversely, patients with mutant p53 do not gain this survival benefit[70]. We identified six patients with wild type *TP53* and these patients could potentially benefit

from 5FU based chemotherapy. However, we postulated that four *TP53* mutated patients, in our study, might not gain a survival benefit from 5FU treatment.

In addition, the *ATM* gene which plays a significant role in the TP53 signalling pathway was found to be mutated in 40% of the proximal CRC patients. For the patients with mutated *ATM* and wild type *TP53* (R3 and R5) they have the additional option of the ATM/ATR Kinase inhibitor as the alternative treatment. A study by Batey and colleague demonstrated that ATM is a valid target for the development of drugs designed to improve the activity of certain cytotoxic anticancer therapies[71]. Small molecule inhibitors of *ATM* are currently in preclinical and clinical development. KU59403 is the first ATM inhibitor to show good tissue distribution and significant chemosensitization in *in vivo* models of human cancer, without major toxicity. This preclinical data of the ATM inhibitor was utilised to support the future clinical development of ATM inhibitors[71]. The tumour progression and response towards treatment are believed to be dependent on both *ATM* and *TP53* status[72]. For instance, ATM signalling is necessary for the survival of TP53-deficient cells after DNA damage; whereas in cancers with wild type *TP53*, inactivation of *ATM* allows the survival of genomically unstable cells and induces chemoresistance[73]. In *TP53*-deficient settings, inhibition of *ATM* dramatically sensitizes tumours to DNA-damaging chemotherapy, whereas, conversely, in the presence of functional *TP53*, inhibition of *ATM* actually promotes resistance effect towards chemotherapy[73]. Thus, the specific set of alterations induced during tumour development play an important role in determining both the tumour response towards chemotherapy and specific susceptibilities to targeted therapies in a given cancer type.

We highlighted here five genes which were of particular interest due to being recurrently mutated exclusively in proximal CRCs. Out of these six recurrently mutated genes, *KRAS* and *GPR6* are known to be involved in cancer, as reported in COSMIC. In our small set of data, we found three patients of proximal CRC (R1, R2 and R3) harbouring at least one of *KRAS* druggable targeted variants in *KRAS* Exon 2 (G12D) and *KRAS* Exon 3 (A59T) and none were found in distal CRC patients. Previous studies also identified that *KRAS*-mutated carcinomas are more frequently found in the proximal colon[74,75]. Activating *KRAS* mutations have been proven to predict a lack of response to anti-EGFR therapy in patients with metastatic colorectal cancer (mCRC). The combination regimen of Panitumumab and Oxaliplatin have no value in metastatic colorectal cancer patients with mutated *KRAS*[76]. A substantial group of mCRC patients with mutated *KRAS* acquired resistance to anti-EGFR treatment cetuximab and a study has suggested an early initiation of MEK inhibitor to delay or reverse the drug resistance[77]. Testing for *KRAS* exon 2 mutation is currently recommended to guide decisions regarding the eligibility for anti-EGFR therapy in mCRC. Profiling of tumour-specific genetic marker will help to guide the selection of patients who are likely to have a response to a particular treatment and prevent adverse effects on those who are unlikely to benefit.

The *GPR6* gene was mutated uniquely in two patients with proximal CRC. *GPR6* protein plays an important role in signal transmission and regulates many cellular functions. There are pieces of evidence implicating the *GPR6* protein and its downstream signalling targets are involved in cancer initiation and progression, where it can influence cell growth and survival through the activation of AKT/mTOR and MAPK pathways[78]. Cancer cells may exploit this pathway which can result in the promotion of tumour growth, angiogenesis and metastasis to distant sites[79]. Findings of recurrently altered *GPR6* gene unique to the proximal CRC patients may explain the poor prognosis and low survival rate in these patients. By directly targeting *GPR6* or its downstream signalling components, it may help to identify novel therapeutic opportunities for cancer prevention and treatment. However, further investigations will be warranted to examine the potential impact of this mutation.

Exome sequencing may lead to the discovery of novel targets, driver mutations as well as novel colorectal cancer-predisposing mutations. This application is getting more common in clinical practice and represents a cost-effective approach to characterize somatic mutations. This discovery study in our own local CRC patients provides a number of insights into the differences in genetic landscape of proximal and distal and identifies potential therapeutic targets in particular to the anatomic subsites of CRC, specifically in Malaysian patients. Nevertheless, further study in a larger series of samples coupled with functional studies will be needed to confirm the identified variants and determine their role in the genesis of proximal and distal CRCs.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

RIMY, NSAM and RJ conceived and designed the experiments. RIMY, SS, NSMN, MRAR and ZAR performed the experiments. RIMY, NSAM and JKSS analysed the data. IS and LM provide the tissue samples. IMR is the pathologist who assessed the tumour percentage of the samples. RIMY and NSAM wrote the manuscript and prepare the figures and tables. NA, NSAM and RJ provide critical reviews.

Funding

The study was funded by a grant from the Ministry of Science, Technology and Innovation (07-05-MGI-GMB016).

Acknowledgements

The authors thanked Ms Chia Chiu Lim for her extensive help in the validation process of the identified variants. This manuscript has been released as a Pre-Print at[80].

Reference

1. Bray F, Ferlay J, Soerjomataram I, *et al.* GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2018; 68(6): 394–424.
2. Zainal Ariffin O. and Nor Saleha IT. National Cancer Registry Report: Malaysia Cancer Statistics- Data and Figure 2007. 2011. Ministry of Health, Malaysia.
3. Lim GCC, Rampal S, and Halimah Y (Eds). Cancer incidence in Peninsular Malaysia 2003–2005. 2008. National Cancer Registry, Ministry of Health Malaysia.
4. Gonzalez EC, Roetzheim RG, Ferrante JM, *et al.* Predictors of proximal vs. distal colorectal cancers. Dis Colon Rectum, 2001; 44(2):251–258.
5. Bufill JA. Colorectal cancer: Evidence for distinct genetic categories based on proximal or distal location. Ann Intern Med, 1990; 113(10):779–788.
6. Minoo P, Zlobec I, Peterson M, *et al.* Characterization of rectal, proximal and distal colon cancers based on clinicopathological, molecular and protein profiles. Int J Oncol, 2010; 37(3):707–718.
7. Missiaglia E, Jacobs B, D'Ario1 G, *et al.* Distal and proximal colon cancers differ in terms of molecular, pathological, and clinical features. Ann Oncol, 2014; 00: 1–7.
8. Maus MKH, Hanna DL, Stephen CL, *et al.* Distinct gene expression profiles of proximal and distal colorectal cancer: Implications for cytotoxic and targeted therapy. Pharmacogenomics J, 2015; 15: 354–362.
9. Bogaert J. and Prenen H. Molecular genetics of colorectal cancer. Ann of Gastroenterol, 2014; 27(1): 9–14.
10. Benedix F, Kube R, Meyer F, *et al.* Comparison of 17,641 patients with right and left-sided colon cancer: Differences in epidemiology, preoperative course, histology, and survival. Dis. Colon Rectum, 2010; 53(1): 57–64.
11. Myer PA, Mannalithara A, Singh G, *et al.* Proximal and distal colorectal cancer resection rates in the United States since widespread screening by colonoscopy. Gastroenterology, 2012; 5: 1227–1236.
12. Li FY. and Lai MD. Colorectal cancer, one entity or three. J Zhejiang Univ Sci B, 2009; 10(3): 219–229.
13. Wu X, Chen VW, Martin J, *et al.* Comparative analysis of incidence rates subcommittee, data evaluation and publication committee, North American Association of Central Cancer Registries. Subsite-specific colorectal cancer incidence rates and stage distributions among Asians and Pacific Islanders in the United States, 1995 to 1999. Cancer Epidemiol Biomarkers Prev. 2004; 13(7): 1215–1222.
14. Irby K, Anderson WF, Henson DE, *et al.* Emerging and widening colorectal carcinoma disparities between Blacks and Whites in the United States (1975–2002). Cancer Epidemiol Biomarkers Prev, 2006; 15(4): 792–797.
15. Goh KL, Quek KF, Yeo GT, *et al.* Colorectal cancer in Asians: A demographic and anatomic survey in Malaysian patients undergoing colonoscopy. Aliment Pharmacol Ther, 2005; 22(9): 859–864.
16. Meldrum C, Doyle MA. and Tothill RW. Next generation sequencing for cancer diagnostics: A practical perspective. Clin Biochem Rev, 2011; 32: 177–195.
17. Roychowdhury S, Iyer MK, Robinson DR, *et al.* Personalized oncology through integrative high-throughput sequencing: A pilot study. Sci Transl Med, 2011; 3: 111–121.
18. Cingolani P, Platts A, Wang le L, *et al.* Program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin), 2012; 6(2): 80–92.
19. Wang K, Li M, and Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010; 38(16): e164.
20. O'Leary NA, Wright MW, Brister JR, *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res, 2016; 44(D1): D733–745.
21. Rosenbloom KR, Armstrong J, Barber GP, *et al.* The UCSC genome browser database: 2015 update. Nucleic Acids Res, 2015; 43: D670–681.
22. Flicek P, Amode MR, Barrell D, *et al.* Ensembl 2014. Nucleic Acids Res, 2014; 42 Database issue: D749-D755; doi:10.1093/nar/gkp972.
23. Siepel A, Bejerano G, Pedersen JS, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res, 2005; 15, 1034–1050.
24. Wenqing F, Timothy DO, Goo J, *et al.* NHLBI Exome Sequencing Project, Joshua MA. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature, 2013; 493: 216–220.
25. Monkol L, Konrad JK, Eric VM, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. Nature, 2016; 536: 285–291.
26. Sherry ST, Ward MH, Kholodov M, *et al.* dbSNP: The NCBI database of genetic variation. Nucleic Acids Res, 2001; 29(1): 308–311.
27. Landrum MJ, Lee JM, Benson M, *et al.* ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Res, 2015; 44(D1): D862–868.
28. Forbes SA, Beare D, Gunasekaran P, *et al.* COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res, 2015; 43(Database issue): D805–811.
29. Ng PC and Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucl Acids Res, 2003; 31 (13): 3812–3814.
30. Adzhubei IA, Jordan DM, and Sunyae SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet, 2013; 7: Unit 7.20.
31. Griffith M, Griffith OL, Coffman AC, *et al.* Mining the druggable genome. Nat Methods, 2013; 10(12): 1209–1210.
32. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 2015; 526: 68–74.
33. Robinson JT, Thorvaldsdóttir H, Winckler W, *et al.* Integrative genomics viewer. Nat Biotechnol, 2011; 29(1), 24–26.
34. Thorvaldsdóttir H, Robinson James T, and Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinformatics, 2013; 14, 178–192.
35. Kumar P, Henikoff S, and Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc, 2009; 4(7): 1073–1081.
36. Altschul SF, Gish W, Miller W, *et al.* Basic local alignment search tool. J Mol Biol, 1990; 215: 403–410.
37. Wang L, Tsutsumi S, Kawaguchi T, *et al.* Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Res, 2012; 22(2): 208–219.
38. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature, 2012; 487: 330–337.
39. DePristo MA, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 2011; 43(5): 491–498; doi: 10.1038/ng.806.
40. Guo Y, Ye F, Sheng Q, *et al.* Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinformatics. 2014; 15(6):879–889.
41. Perreault N, Katz JP, Sackett SD, *et al.* Foxl1 controls the Wnt/beta-catenin pathway by modulating the expression of proteoglycans in the gut. J Biol Chem. 2001; 276: 43328–43333.
42. Moreno-Bueno G, Hardisson D, Sanchez C, *et al.* Abnormalities of the APC/beta-catenin pathway in endometrial cancer. Oncogene, 2002; 21: 7981–7990.
43. Segditsas S, Rowan AJ, Howarth K, *et al.* APC and the three-hit hypothesis. Oncogene, 2009; 28: 146–155.
44. Reya T, and Clevers H. Wnt signalling in stem cells and cancer. Nature, 2005; 434: 843–850.
45. Van Es JH, Jay P, Gregorieff A, *et al.* Wnt signalling induces maturation of Paneth cells in intestinal crypts. Nat Cell Biol, 2005; 7: 381386.
46. Femia AP, Dolara P, Giannini A, *et al.* Frequent mutation of Apc gene in rat colon tumours and mucindepleted foci, preneoplastic lesions in experimental colon carcinogenesis. Cancer Res, 2007; 67: 445–449.
47. Nagel R, le Sage C, Diosdado B, *et al.* Regulation of the adenomatous polyposis coli gene by the miR-135 family in colorectal cancer. Cancer Res, 2008; 68: 5795–5802.
48. Clevers H and Nusse R. Wnt/beta-catenin signalling and disease. Cell, 2012; 149: 1192–1205.
49. Diaz LA Jr, Williams RT, Wu J, *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. Nature, 2012; 486: 537–540.
50. Vogelstein B, Papadopoulos N, Velculescu VE, *et al.* Cancer genome landscapes. Science, 2013; 339: 1546–1558.
51. Ashktorab H, Daremipouran M, Devaney P, *et al.* Identification of novel mutations by exome sequencing in African American colorectal cancer patients. Cancer, 2014; 121(1): 34–42.
52. Nikhil W, Michael FB, Matthew JD, *et al.* High-throughput detection of actionable genomic alterations in clinical tumour samples by targeted, massively parallel sequencing. Cancer Discov, 2012; 2(1): 82–93.
53. Fang W, Radovich M, Zheng Y, *et al.* Druggable alterations detected by Ion Torrent in metastatic colorectal cancer patients. Oncol Lett, 2014; 7: 1761–1766.
54. Salem ME, Weinberg BA, Xiu J, *et al.* Comparative molecular analyses of left-sided colon, right-sided colon and rectal cancers. Oncotargets, 2017; 8(49): 86356–86368.
55. Gao J, Aksoy BA, Dogrusoz U, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal, 2013; 6(269): 11.
56. Birgisson H, Edlund K, Wallin U, *et al.* Microsatellite instability and mutations in BRAF and KRAS are significant predictors of disseminated disease in colon cancer. BMC Cancer, 2015; 15: 125.
57. Timmermann B, Kerick M, Roehr C, *et al.* Somatic mutation profiles of MSI and MSS colorectal cancer identified by Whole Exome Next Generation Sequencing and Bioinformatics Analysis. PLoS One. 2010; 5(12): 15661.
58. Kandoth C, McLellan MD, Vandin F, *et al.* Mutational landscape and significance across 12 major cancer types. Nature, 2013; 502: 333–339.
59. Vogelstein B, Papadopoulos N, Velculescu VE, *et al.* Cancer genome landscapes. Science, 2013; 339: 1546–1558.
60. Pamplona RS, Doriga AL, Brunet LP, *et al.* Exome sequencing reveals AMER1 as a frequently mutated gene in colorectal cancer. Clin Cancer Res, 2015; 21(20): 4709–4718.
61. Scholer-Dahirel A, Schlabach MR, Loo A, *et al.* Maintenance of adenomatous polyposis coli (APC)-mutant colorectal cancer is dependent on Wnt/β-catenin signalling. Proc Nat Acad Sci U.S.A., 2011; 108(41): 17135–17140.
62. Pino MS, and Chung DC. The chromosomal instability pathway in colon cancer. Gastroenterology, 2010; 138(6): 2059–2072. doi:10.1053/j.gastro.2009.12.065.
63. Malekzadeh R, Bishehsari F, Mahdavinia M, Ansari R. Epidemiology and molecular genetics of colorectal cancer in Iran: A review. Arch Iran Med, 2009; 12 (2): 161–169.
64. Mahdavinia M, Bishehsari F, Verginelli F, *et al.* P53 mutations in

colorectal cancer from northern Iran: Relationships with site of tumour origin, microsatellite instability and K-ras mutations. J Cell Physio, 2008; 216(2): 543–550.
65. Manne U, Weiss HL, Myers RB, *et al.* Nuclear accumulation of p53 in colorectal adenocarcinoma: Prognostic importance differs with race and location of the tumour. Cancer, 1998; 83: 2456–2467.
66. Samowitz WS, Curtin K, Ma KN, *et al.* Prognostic significance of p53 mutations in colon cancer at the population level. Int J Cancer. 2002; 99: 597–602.
67. Soong R, Grieu F, Robbins P, *et al.* p53 alterations are associated with improved prognosis in distal colonic carcinomas. Clin Cancer Res, 1997; 3: 1405–1411.
68. Adrover E, Maestro ML, Sanz-Casla MT, *et al.* Expression of high p53 levels in colorectal cancer: a favorable prognostic factor. Br J Cancer, 1999; 81: 122–126.
69. Iacopetta B. TP53 Mutation in Colorectal Cancer. Hum Mutat, 2003; 21: 271–276.
70. Daniel BL, Paul DH and Patrick GJ. 5-Fluorouracil: Mechanisms of action and clinical strategies. Nat Rev Cancer, 2003; 3: 330–338.
71. Batey MA, Zhao Y, Kyle S, *et al.* Preclinical evaluation of a novel ATM inhibitor, KU59403, in vitro and in vivo in p53 functional and dysfunctional models of human cancer. Mol Cancer Ther, 2013; 12(6): 959–967.
72. Song H, Hollstein M, and Xu Y. p53 gain-of-function cancer mutants induce genetic instability by inactivating ATM. Nat Cell Biol, 2007; 9: 573–580.
73. Reinhardt HC, Aslanian AS, Lees JA, *et al.* p53-deficient cells rely on ATM and ATR-mediated checkpoint signalling through the p38MAPK/MK2 pathway for survival after DNA damage. Cancer Cell, 2007; 11: 175–189.
74. Rosty C, Young JP, Walsh MD, *et al.* Colorectal carcinomas with KRAS mutation are associated with distinctive morphological and molecular features. Mod Pathol, 2013; 26(6): 825–834.
75. Wenbin L, Tian Q, Wenxue Z, *et al.* Colorectal carcinomas with KRAS codon 12 mutations are associated with more advanced tumor stages. BMC Cancer, 2015; 15: 340.
76. Douillard JY, Oliner KS, Siena S, *et al.* Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer. N Engl J Med, 2013; 369(11), 1023–1034.
77. Misale S, Yaeger R, Hobor S, *et al.* Emergance of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. Nature, 2012; 486: 532–536.
78. Morgan OH, Maria SD, and Silvio JG. Novel insights into G protein and G protein-coupled receptor signalling in cancer. Curr Opin Cell Biol, 2014; 27: 126–135.
79. Gaorav PG and Joan M. Cancer metastasis: Building a framework. Cell Press, 2006; 127(4): 679–695.
80. Mohd Yunos R, Ab Mutalib N, Khor SS, *et al.* Characterisation of genomic alterations in proximal and distal colorectal cancer patients. PeerJ Preprints, 2016; 4: e2109v1.

Availability of data and material

The aligned whole-exome sequencing data in .bam format were deposited at the NCBI Sequence Read Archive (SRA) at http://www.ncbi.nlm.nih.gov/sra with accession number PRJNA382764.