genes

MDPI

Article

Whole-Genome Profiles of Malay Colorectal Cancer Patients with Intact MMR Proteins

Wan Khairunnisa Wan Juhari [1,2], Khairul Bariah Ahmad Amin Noordin [3], Andee Dzulkarnaen Zakaria [4], Wan Faiziah Wan Abdul Rahman [5], Wan Muhamad Mokhzani Wan Muhamad Mokhter [4], Muhammad Radzi Abu Hassan [6], Ahmad Shanwani Mohammed Sidek [7] and Bin Alwi Zilfalil [1,2,*]

1 Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; khairunnisa.juhari@gmail.com
2 Malaysian Node of the Human Variome Project, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia
3 School of Dental Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; kbariah@usm.my
4 Department of Surgery, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; andee@usm.my (A.D.Z.); mokhzani@usm.my (W.M.M.W.M.M.)
5 Department of Pathology, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian 16150, Kelantan, Malaysia; wfaiziah@usm.my
6 Clinical Research Centre, Hospital Sultanah Bahiyah, Alor Star 05460, Kedah, Malaysia; drradzi91@yahoo.co.uk
7 Surgery Department, Hospital Raja Perempuan Zainab II, Kota Bharu 15200, Kelantan, Malaysia; drahmad69@gmail.com
* Correspondence: zilfalil@usm.my; Tel.: +60-9-7676531

Citation: Juhari, W.K.W.; Ahmad Amin Noordin, K.B.; Zakaria, A.D.; Rahman, W.F.W.A.; Mokhter, W.M.M.W.M.; Hassan, M.R.A.; Sidek, A.S.M.; Zilfalil, B.A. Whole-Genome Profiles of Malay Colorectal Cancer Patients with Intact MMR Proteins. *Genes* **2021**, *12*, 1448. https://doi.org/10.3390/genes12091448

Academic Editor: Yannick D. Benoit

Received: 13 July 2021
Accepted: 18 September 2021
Published: 20 September 2021

Abstract: Background: This study aimed to identify new genes associated with CRC in patients with normal mismatch repair (MMR) protein expression. Method: Whole-genome sequencing (WGS) was performed in seven early-age-onset Malay CRC patients. Potential germline genetic variants, including single-nucleotide variations and insertions and deletions (indels), were prioritized using functional and predictive algorithms. Results: An average of 3.2 million single-nucleotide variations (SNVs) and over 800 indels were identified. Three potential candidate variants in three genes—*IFNE*, *PTCH2* and *SEMA3D*—which were predicted to affect protein function, were identified in three Malay CRC patients. In addition, 19 candidate genes—*ANKDD1B*, *CENPM*, *CLDN5*, *MAGEB16*, *MAP3K14*, *MOB3C*, *MS4A12*, *MUC19*, *OR2L8*, *OR51Q1*, *OR51AR1*, *PDE4DIP*, *PKD1L3*, *PRIM2*, *PRM3*, *SEC22B*, *TPTE*, *USP29* and *ZNF117*—harbouring nonsense variants were prioritised. These genes are suggested to play a role in cancer predisposition and to be associated with cancer risk. Pathway enrichment analysis indicated significant enrichment in the olfactory signalling pathway. Conclusion: This study provides a new spectrum of insights into the potential genes, variants and pathways associated with CRC in Malay patients.

Keywords: whole-genome sequencing; colorectal cancer; Malay; olfactory signalling pathway

1. Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer worldwide. It is the third most common cancer worldwide and the second most common cause of death [1]. Geographical and distribution differences influencing the incidence of CRC have been observed across the world, including an accelerated incidence rate in several Asian countries [2,3]. According to the Malaysia National Cancer Registry 2008–2013, CRC is one of the most common cancers in men and the third most common cancer in women [4].

Hereditary colorectal cancers are caused by highly penetrant mutations, such as those involved in tumour suppression or in the DNA mismatch repair system, including Hereditary Nonpolyposis Colorectal Cancer (HNPCC), Familial Adenomatous Polyposis (FAP),

MYH-associated polyposis, and the rare hamartomatous polyposis syndromes [5]. Hereditary nonpolyposis colorectal cancer (HNPCC), also known as Lynch syndrome (LS), is an autosomal dominant cancer syndrome that is known to be the most common hereditary cancer, accounting for 5–10% of total CRC [6]. Individuals with LS are characterized by a high tendency to develop cancers in the extracolonic organs, such as endometrium, stomach, ovary, small bowel, hepatobiliary tract, renal pelvis, ureter, skin, and brain [7,8]. In the context of familial colorectal cancer, the genetic causes of familial adenomatous polyposis and LS have been well documented, with over 30% of all CRC cases having been identified to carry underlying genetic factors [5]. Mismatch repair (MMR) genes, including *MLH1*, *MSH2*, *MSH6*, and *PMS2*, are the most common genes that cause germline mutations in LS, with almost 90% of the cases diagnosed being associated with mutations in the *MLH1* and *MSH2* genes [9,10].

Individuals may develop a hereditary cancer syndrome when they acquire an inherited mutation, thus having an increased risk of developing certain tumours, which can appear at a relatively early age. In most known hereditary malignant syndromes, the increased risk is due to the mutation of a single gene, making these pathologies monogenic hereditary diseases. The affected genes commonly control the cell cycle or are involved in the process of repairing DNA damage. Non-hereditary tumours (sporadic cases) are also caused by an increased incidence of mutations in these genes; however, in sporadic cases, the genetic changes have newly developed in the cells of a tissue, causing somatic mutations, and are absent in other body cells [11].

In addition, there is a number of common low-risk loci identified in other studies which are known to contribute to an increased risk of both sporadic and hereditary cases of CRC [12,13]. With recent advancements in human genetic research, the technological progress in sequencing linked to next-generation sequencing (NGS) has led to an increase in knowledge and a better understanding of genetic mutations in cancer cells and pathway alterations, serving to create new models and enhance findings in the biology of cancer [14]. Whole-genome sequencing (WGS), which is a part of NGS, can be utilised to identify additional possible mutations and/or variants associated with CRC. The NGS technology through whole-genome sequencing has also revealed numerous single-nucleotide polymorphisms and somatic mutations in cancer genomes which had not been previously reported [15]. Although most inherited variants common in human populations have been discovered and are listed in databases, there are myriad rare inherited single-nucleotide polymorphisms (SNPs) and structural variants yet to be found and, in most cancer genomes, these rare germline mutations are present in higher frequency than somatic mutations [14]. Hence, the use of WGS has led to the discovery of causative mutations for specific types of cancer [16,17].

The discernible difference between whole-exome sequencing (WES) and whole-genome sequencing (WGS) is that WES can capture or identify variants only in genes' coding regions, while WGS is more efficient in identifying variants in the entire genome [18]. WGS itself is able to accurately detect and identify a higher percentage of true positive single-nucleotide variants (SNVs) in the exome [19]. Therefore, in this study, WGS was performed to capture various types of genomic alterations, in order to discover and further determine high-impact variants and other mutations, including rare mutations in other genes, that may be associated with an increased risk of CRC, particularly, in our Malay patients who fulfilled the Bethesda criteria.

2. Materials and Methods

2.1. Selection of Patients

Ethical approval was obtained from the Research and Ethics Committee, Universiti Sains Malaysia (USMKK/PPP/JEPeM [259.3.(9)]), and the Medical Research and Ethics Committee (MREC), Ministry of Health (NMRR-12-856-11623). All patients were selected from three hospitals, i.e., the Hospital Universiti Sains Malaysia (USM) and two hospitals under the Ministry of Health of Malaysia: Hospital Raja Perempuan Zainab II, Kota Bharu,

Kelantan, and Hospital Sultanah Bahiyah, Alor Setar, Kedah. Sample recruitment was only focused on the Malay probands, due to the demographic pattern and to the fact that the majority of patients in these three hospitals were Malays. In addition, there is a scarcity of data for HNPCC in Malaysia, including mutations and/or polymorphisms specifically for the Malay population—the biggest ethnic group in Malaysia. In Malaysia, a multi-ethnic country with three different major ethnic groups (Malay, Chinese, and Indian), at present there are limited data on HNPCC variants in the Chinese population, but none have been reported for the Malay and Indian populations. Therefore, we decided to include and focus on patients of Malay ethnicity only. The CRC patients who fulfilled at least one of the revised Bethesda Criteria were enrolled into this study, according to the inclusion and exclusion criteria (see Table 1). Informed consent was obtained for each patient prior to sample collection. Seven patients were enrolled into this study, with five of them being unrelated patients (denoted as F1, F2, F12, F18, and F19), and two from the same family (denoted as F5 and F8).

Table 1. Inclusion and exclusion criteria for the selection of the patients.

Inclusion Criteria
Malay patients (at least three generations and no admixture in the parental heritage) with colorectal cancer who fulfilled at least one of the following Bethesda Criteria:
i. Colorectal Cancer (CRC) diagnosed in patient aged <50 years.
ii. Presence of synchronous, metachronous colorectal, or other Lynch syndrome-related tumours *, regardless of age.
iii. Patient with CRC and a first-degree relative with a Lynch syndrome-related tumour, with one of the cancers diagnosed at age <50 years.
* Lynch syndrome-related tumours include colorectal, endometrial, stomach, ovarian, pancreas, ureter, renal pelvis, biliary tract and brain tumours, sebaceous gland adenomas and keratoacanthomas and carcinomas of the small bowel.
Exclusion Criteria
(1) Non-Malay (Chinese and Indian ethnic groups).
(2) Patients with Familial Adenomatous Polyposis

2.2. Immunohistochemical Screening

Immunohistochemical staining was performed on formalin-fixed, paraffin-embedded (FFPE) tissue from a biopsy or resected bowel specimen. Immunohistochemical staining using four types of MMR antibodies—MLH1, MSH2, MSH6, and PMS2—and semi-quantitative scoring assessment were performed, as described previously by our groups [20].

2.3. Whole-Genome Sequencing

Genomic DNA was extracted from blood by a QIAamp DNA Blood Kit, following the manufacturer's protocol (Qiagen, Hilden, Germany). Library preparation was carried out using TruSeq Nano DNA HT (Illumina, San Diego, CA, USA) prior to library quantification. The DNA libraries were then clustered onto the HiSeqX flow cell and were sequenced using the HiSeqX platform of the Beijing Genome Institute (BGI). Base calling was processed by an Illumina pipeline with default parameters, and the sequences of each patient were generated as 150 base pair (bp) paired-end reads. The adapter sequences of unknown bases, low-quality reads, and reads with unknown bases corresponding to more than 10% were removed from the raw sequencing data, prior to sequence alignment to the reference genome. The filtered reads were aligned to a human genome reference (GRCh37/hg19) using the Burrows–Wheeler Aligner (BWA) (Supplementary Table S1) [21]. Duplicate reads caused by PCR were further marked by Picard tools prior to variant calling. SOAPsnp was used to call single-nucleotide variants (SNVs) [22], and small insertion/deletions (indels) were detected by Samtools [23]. Following sequencing, the predicted effects of each variant were annotated using Annovar [24]. A list of databases and additional prediction algorithms were used to estimate the allele frequencies of each variant, including dbSNP [25],

1000 Genomes Project [26], and 1000 Genomes East Asian Project [27]. The variant functional effects and pathogenicity were further predicted by Polymorphism Phenotyping v2 (PolyPhen2) [28], Sorting Intolerant From Tolerant (SIFT) [29], MutationAssessor [30] and Functional Analysis through Hidden Markov Models (FATHMM). Additional annotation to facilitate the characterization and interpretation of variants was carried out using a cancer-related database, the Catalogue of Somatic Mutations in Cancer (COSMIC) database [31]; a disease-related database, ClinVar [32]; the Human Gene Mutation Database (HGMD) [33]; a systematic review of the literature.

2.4. Variant Prioritization

In the present study, the variants were filtered to prioritize the causative variants to LS predisposition. For single-nucleotide variants (SNVs), variants in the coding region and high-impact variants were selected. The high-impact variants were identified to be functional variants, i.e., missense, nonsense, splice acceptor and splice donor variants [34]. Over 200 high-impact variants were identified in each patient, with an average of 76 high-impact variants classified as nonsense mutations. In addition, to fully characterize the plausible variants in our patients, the identification of rare SNVs was further carried out on the basis of the following criteria: synonymous and SNVs present in the coding region were excluded, as well as variants with no subsequent effect on amino acids. The SNVs present in the dbSNP141 (SNP database) and 1000 Genomes were also excluded, and we selected only variants that could actually be damaging and have an effect on protein function, through in silico prediction using SIFT and Polyphen2.

For prioritization of indels, only indels in coding regions were selected, including indels leading to disruptive in-frame insertion, disruptive in-frame deletions frameshift variants, frameshift mutations causing stoploss mutation, in-frame insertion, in-frame deletion, splice acceptor variants, splice region variants, and in-frame deletion causing stoploss mutation. Synonymous and indels that lay in non-coding regions, including indels with no functional protein annotation mapped to SIFT and PolyPhen2, were excluded for downstream interpretation. Rare indels, which were hypothetically considered as rare indels when those variants were not identified in dbSNP141 (SNP database) and 1000 Genomes database, were included in this study. The variants presented in the COSMIC database were included in order to ascertain the impact of the variants on human cancers.

Pathway enrichment analysis, using STRING (https://string-db.org/ (accessed on 26 August 2021)) [35] and Reactome (https://reactome.org/ (accessed on 28 August 2021)) [36], was performed to prioritize the non-synonymous variants harbouring nonsense single-nucleotide variants in the respective candidate genes. Interaction and correlation based on the knowledge-based pathway map was employed by String, based on Gene Ontology (http://www.geneontology.org/GO (accessed on 28 August 2021)) and KEGG (http://www.genome.jp/kegg/ (accessed on 28 August 2021)).

3. Results

3.1. Variant Identification in Patients with Intact MMR Protein Expression

Immunohistochemical staining of the seven studied patients showed no loss of expression in the four MMR antibodies MLH1, MSH2, MSH6 and PMS2. The resulting variants in MMR genes were mostly discovered in intronic regions, suggesting that the variants may have no effect on protein expression. In the present study, an average of 3.2 million single nucleotide variations were identified in each genome, when mapping against the human genome reference sequence assembly GRCh37, also known as hg19 (Supplementary Table S2). The genome sequences covered over 99% of the reference sequence, in an approximate range of 44- to 52-fold sequencing depth for each sample. The whole-genome sample data of the studied patients were filtered and prioritized to fully characterize the high- and low-risk loci that may be associated with CRC in Malay patients fulfilling the Bethesda criteria. We discovered a non-synonymous polymorphism in exon 3 of the *EPCAM* gene c.344T>C (p.Met115Thr) (rs1126497) in the seven studied patients.

In addition, three female patients were identified to harbour a nonsense variant in the *IFNE* gene, c.211C>T (p.Gln71*) (rs2039381). Regarding this variant, in silico prediction using SIFT found that the variant was predicted to abolish protein function. Among the three patients with the variant in the *IFNE* gene, two were first-degree relatives (F5 and F8). Two rare heterozygous missense mutations—c.1307C>T (p.Ala436Val) in exon 10 of *PTCH2* and a mutation in the *SEMA3D* gene, c.278T>A (p.Leu93His) located at exon 2—were also exclusively identified in these patients. F5 and F8 were diagnosed with colorectal cancer at the age of 43 and 56 years, respectively. In silico prediction to further assess the functional consequences of these mutations was performed using SIFT, Poyphen2, MutationAssessor, RadialSVM, and FATHMM. Both SIFT and Polyphen2 predicted that these mutations would be deleterious, with scores of 0 and 0.8–0.9, respectively. For the mutation in the *PTCH2* gene, MutationAssessor and RadialSVM resulted in scores of 2.515 (Medium) and 1.0647 (Damaging), respectively, and the mutation in the *SEMA3D* gene resulted in scores of 3.765 (High) and 0.0534 (Damaging), respectively. Based on FATHMM prediction, both mutations were predicted to be pathogenic.

The SNVs were classified into high-, moderate-, and low-impact, based on the annotation algorithms. SNVs that lead to protein truncation could have a highly disruptive effect on gene function, whereas SNVs that influence only protein effectiveness are most likely to have only a moderate effect, and synonymous SNVs that are unlikely to change protein behaviour probably have a minimal effect. In these studied patients, a total of more than 200 SNVs that could affect gene function (high impact) were identified, and those SNVs that caused stop-gain mutations (nonsense mutations) were selected for further pathway enrichment analysis. In patient F1, a total of 81 nonsense mutations were identified from 274 high-impact SNVs, whereas, in patient F2, 73 nonsense mutations were identified out of 265 high-impact SNVs. For patient F5, 82 nonsense mutations were discovered from 286 high-impact SNVs identified, while a total of 73 nonsense mutations were identified in patient F8 from 274 high-impact SNVs. Of 270 high-impact SNVs found in patient F12, 78 SNVs were identified to be nonsense mutations. A total of 73 nonsense mutations were identified in patients F18 and F19, from 262 and 267 high-impact SNVs, respectively. Twenty nonsense mutations in 19 genes were prioritized, taking into account only the shared mutations among these seven CRC patients (Table 2). We also identified 15 nonsense and 7 indels occurring in five or six samples. Two variants within these five or six samples were identified in two genes, *KRT10* (p.Gly490_Gly493del/c.1468_1479delGGCCACGGCGGC) and *MTSS1* (c.1417-37delT).

3.2. Pathway Analysis

Pathway enrichment analysis was then performed for all the candidate genes harbouring nonsense mutations. For this, Reactome, a curated database of pathways and reactions in human biology, was used. From the test, a probability score was produced, which was then corrected for false discovery rate (FDR) using the Benjamani–Hochberg method. Based on our submitted data, the olfactory signalling pathway was discovered to be the most significant pathway for each patient (Supplementary Table S3). In concordance, Gene Ontology (GO) enrichment analysis and KEGG pathway, performed using the String tool, revealed that the molecular function and pathway identified were primarily related to olfactory receptor activity and olfactory transduction pathway, respectively (Table 3). In addition to SNVs, over 800,000 insertions and deletions (indels) were identified in each patient. Eight indels of two in-frame insertions, two frameshift deletions, three frameshift insertions, and one disruptive in-frame deletion were considered as rare indels in the Malay CRC candidate genes *CDK11B*, *CCDC144NL*, *GOLGA8R*. *MAFA*, *MUC6*, and *PRIM2* (Table 4). Pathway enrichment analysis was also performed in all candidate genes harbouring rare indels in these seven patients. Several significant pathways were identified from the submitted data, including a pathway that caused colorectal cancer by defective GALNT12.

Table 2. Nonsense mutations identified in the seven whole-genome samples.

Gene	Transcript	Codon Change	Chr	Start	End	Ref	Obs	Frequency *
MOB3C	NM_145279.4:p.Arg24*/c.70C>T	Cga/Tga	chr1	47080679	47080679	G	A	0.642173
PDE4DIP	NM_001198834.4:p.Trp2351*/c.7053G>A	tgG/tgA	chr1	144852390	144852390	C	T	-
PDE4DIP	NM_001198834.4:p.Arg622*/c.1864C>T	Cga/Tga	chr1	144915561	144915561	G	A	-
OR2L8	NM_001001963.1:p.Tyr289*/c.867T>A	taT/taA	chr1	248113026	248113026	T	A	0.745807
SEC22B	NM_004892.5:p.Arg132*/c.394C>T	Cga/Tga	chr1	145112420	145112420	C	T	-
ZNF117	NM_015852.3:p.Arg428*/c.1282C>T	Cga/Tga	chr7	64438667	64438667	G	A	0.881789
ANKDD1B	NM_001276713.1:p.Trp480*/c.1439G>A	tGg/tAg	chr5	74965122	74965122	G	A	0.506989
PRIM2	NM_000947.4:p.Gln325*/c.973C>T	Cag/Tag	chr6	57398270	57398270	C	T	-
OR51Q1	NM_001004757.2:p.Arg236*/c.706C>T	Cga/Tga	chr11	5444136	5444136	C	T	0.453874
MUC19	NM_173600.2:p.Cys1238*/c.3714C>A	tgC/tgA	chr12	40834955	40834955	C	A	0.518371
USP29	NM_020903.2:p.Tyr913*/c.2739C>A	taC/taA	chr19	57642782	57642782	C	A	0.952077
PRM3	NM_021247.2:p.Arg104*/c.310C>T	Cga/Tga	chr16	11367143	11367143	G	A	1
MAP3K14	NM_003954.4:p.Ser902*/c.2705C>G	tCa/tGa	chr17	43342141	43342141	G	C	0.718251
MS4A12	NM_017716.2:p.Gln71*/c.211C>T	Caa/Taa	chr11	60265002	60265002	C	T	0.478235
PKD1L3	NM_181536.1:p.Arg789*/c.2365C>T	Cga/Tga	chr16	72001136	72001136	G	A	0.263379
TPTE	NM_199261.3:p.Arg229*/c.685C>T	Cga/Tga	chr21	10942756	10942756	G	A	-
OR5AR1	NM_001004730.1:p.Gln19*/c.55C>T	Cag/Tag	chr11	56431216	56431216	C	T	0.623003
CENPM	NM_001110215.1:p.Arg3*/c.7C>T	Cga/Tga	chr22	42336172	42336172	G	A	0.239217
MAGEB16	NM_001099921.1:p.Arg272*/c.814C>T	Cga/Tga	chrX	35821127	35821127	C	T	0.620397
CLDN5	NM_003277.3:p.Gln37*/c.109C>T	Cag/Tag	chr22	19511925	19511925	G	A	0.496805

* Frequency in 1000 Genomes Project.

Table 3. Significant pathways identified in the seven whole-genome samples of CRC patients.

Patient ID	Pathway ID	Pathway Description	FDR *	KEGG Pathway	FDR *
F1	GO.0004984	Olfactory receptor activity	0.0000169	Olfactory transduction	0.00000315
	GO.0004930	G-protein coupled receptor activity	0.000486		
	GO.0004888	Transmembrane signalling receptor activity	0.00497		
F2	GO.0004984	Olfactory receptor activity	0.00000228	Olfactory transduction	0.000000432
	GO.0004930	G-protein coupled receptor activity	0.0000529		
F5	GO.0004984	Olfactory receptor activity	0.00133	Olfactory transduction	0.000207
	GO.0004888	Transmembrane signalling receptor activity	0.00423		
	GO.0004930	G-protein coupled receptor activity	0.00602		
F8	GO.0004984	Olfactory receptor activity	0.00676	Olfactory transduction	0.00097
F12	GO.0004984	Olfactory receptor activity	0.000105	Olfactory transduction	0.0000179
	GO.0004930	G-protein coupled receptor activity	0.000229		
	GO.0004888	Transmembrane signalling receptor activity	0.000411		
F18	GO.0004984	Olfactory receptor activity	0.00064	Olfactory transduction	0.000101
	GO.0004930	G-protein coupled receptor activity	0.000768		
	GO.0004888	Transmembrane signalling receptor activity	0.00569		
F19	GO.0004984	Olfactory receptor activity	0.000548	Olfactory transduction	0.0000863
	GO.0004930	G-protein coupled receptor activity	0.00444		

* False discovery rate (FDR) ≤ 0.05.

Table 4. Insertions and deletions identified in the seven whole-genome samples of CRC patients.

Function	Gene	Transcript	Chr	Start	End	Ref	Obs
In-frame insertion	*CDK11B*	NM_001787.2:p.Arg127_Glu128insLysGluArg/ c.379_380insAAGAAA	1	1647893	1647893	C	CTTTCTT
Frameshift variant	*CCDC144NL*	NM_001004306.1:p.Lys213fs/ c.638delA	17	20768755	20768755	CT	C
Frameshift variant	*CCDC144NL*	NM_001004306.1:p.Lys211_Gly212fs/ c.631_632insG	17	20768762	20768762	T	TC
In-frame insertion	*GOLGA8R*	NM_001282484.1:p.Gln271_Asp272insGlnGln/ c.813_814insCAA	15	30700168	30700168	C	CTTG
Disruptive in-frame deletion	*MAFA*	NM_201589.3:p.His207_His208del/ c.621_623delCCA	8	144511953	144511953	ATGG	A
Frameshift variant	*MUC6*	NM_005961.2:p.Pro1571fs/c.4712delC	11	1018088	1018088	TG	T
Frameshift variant	*MUC6*	NM_005961.2:p.Pro1569_Pro1570fs/c.4707_4708insA	11	1018093	1018093	G	GT
Frameshift variant	*PRIM2*	NM_000947.4:p.Glu297_Asn298fs/c.890_891insA	6	57398186	57398186	G	GA

4. Discussion

Whole-genome sequencing (WGS) was employed to further discover the molecular basis of predisposition to CRC in patients fulfilling the Bethesda criteria, which demonstrated an intact protein expression of four common MMR genes: *MLH1*, *MSH2*, *MSH6* and *PMS2*. We hypothesized that other genes could contribute to the CRC predisposition and sought to further identify other pathways that may be associated with the candidate genes. In this study, WGS was employed over exon capture approaches, in order to fully discover whether the causal variants would reside in known coding regions or other non-coding regions, as exome sequencing or other targeted approaches are only tailored to capture limited regions of variants [37]. Approximately 11% of the variants discovered by WGS have been reported to be missed by WES [38], and, even though WGS may fail to identify WES-specific variants, the number of variants missed by WGS has been found to be less significant [39]. Due to the massive amount of NGS data, a constructive approach must be considered to thoroughly select, filter and extract the functional variants from the data. Potential variants should be prioritized, including functional variants of uncommon polymorphisms, in order to search for likely true candidates for the studied disease [40]. The region-based annotation that has been implemented by ANNOVAR to classify the variants into specific genomic regions, such as intronic, intergenic, exonic, untranslated, splicing and non-coding RNA, including downstream and upstream genomic regions as well as their classes, such as synonymous, missense and frameshift variants [41], provides a useful handling step, as gene-based annotations cannot fully predict the functional consequences of variants outside protein coding regions [24].

A non-synonymous polymorphism, c.344T>C (p.Met115Thr) (rs1126497), in exon 3 of the *EPCAM* gene was discovered in the seven considered patients. This non-synonymous polymorphism has been previously reported as having a significant association with an increased risk of developing breast cancer in a Chinese population [42]. The location of this non-synonymous polymorphism in the thyroglobulin (TY) domain of the *EPCAM* gene suggests its role in inhibiting cathepsins, a family of cysteine proteases that are frequently secreted by tumour cells during metastasis [43]. The amino acid change of methionine to threonine in this TY domain may suggest its role in *EPCAM* gene function [42]. In addition, the association between *EPCAM* and *MSH2* was due to the simultaneous loss of EPCAM and MSH2 protein expression in colorectal cancer cases in *EPCAM* deletion carriers [44]. MSH2 inactivation was predicted in several patients with no MMR germline mutations but, in the presence of heterozygous germline deletion at the polyadenylation site in the exon 8 and 9 of *EPCAM* gene [45], the deletions were shown to cause a transcriptional read-through that silenced and inactivated the promoter of the *MSH2* gene located downstream of EPCAM gene [46]. Germline deletions in the last exon of the *EPCAM* gene may silence its neighbouring gene, *MSH2*, which is located 17 kb downstream of *EPCAM*, via

promoter hypermethylation [47]. The germline deletion that causes *MSH2* inactivation was considered a novel mutation predisposing to HNPCC [48,49].

We discovered that the high-impact variants in this cohort of patients harboured nonsense variants, and these variants should be of primary interest in disease-related studies due to their potentially high cellular impact, as they include exonic missense, nonsense, stop-loss, frameshift, and splice site variants, which potentially affect protein function [41]. Nineteen candidate genes harbouring nonsense variants that may be involved in CRC in these Malay patients—*ANKDD1B*, *CENPM*, *CLDN5*, *MAGEB16*, *MAP3K14*, *MOB3C*, *MS4A12*, *MUC19*, *OR2L8*, *OR51Q1*, *OR51AR1*, *PDE4DIP*, *PKD1L3*, *PRIM2*, *PRM3*, *SEC22B*, *TPTE*, *USP29* and *ZNF117*—were further ascertained by consulting the literature and public databases for their possible clinical implications with respect to predisposition to other cancers. Claudin-5 is primarily expressed by the vascular endothelium and functions in the blood–brain barrier and pulmonary endothelial barrier, in addition to serving as a regulator of epithelial function [50]. Two different entries for the human claudin-5 protein with gene products of different size were found in the NCBI and Uniprot databases, which were 218 and 303 amino acids in length, respectively [51]. Considering the different types of gene products, a study was carried out to explore the coding sequence of *CLDN5*, and a reported SNP of rs885985 was found in the general population [51,52], as well as in our studied patients. The *CLDN5* allele introduces a stop codon and results in a claudin-5 open reading frame (ORF) with 218 amino acids. The presence of the G allele may introduce an overlapping ORF, which consequently encodes the two types of gene products. Immunoblotting of human lung tissue was, then, carried out to measure the size of the produced protein, which resulted in to be only 218-amino acid long [51]. A previous study discovered the role of MAP3K14 in the suppression of epithelial cell proliferation and its involvement in non-canonical NF-κB signalling during CRC development [53]. NF-κB-inducing kinase (NIK, also known as MAP3K14) signalling has been found to be essential in modulating the activity of the NF-κB pathway in pancreatic cancer [54], as the increased activity of the NF-κB pathway in pancreatic cancer caused cell proliferation and tumour development [55,56]. In addition, it has been reported that MS4A14 protein expression was detected in several cases of colorectal cancer, and retained expression was observed during the malignant transformation of tumours [57]. MS4A12 also functions in modulating EGFR signalling, the main tumour-promoting factor in colon cancer, causing tumour growth and survival, whereas loss of MS4A12 protein deteriorated EGFR-dependent cell functions [57]. Two other candidate genes harbouring indels were both identified to be associated with cancer predisposition. *MSS1* may be involved in the inhibition of CRC metastasis [58], and *KRT10* was reported to be highly expressed in hereditary skin cancer [59].

A majority of the identified variants should be validated and, in this case, biological knowledge including their molecular functions and interactions is essential to explore and confirm the role of these candidate variants and genes [41]. In addition, biological differences between mutations associated with human diseases and polymorphisms that cause stop codons should be further explored, as they may shed light into other molecular pathways at the basis of the disease [60]. Pathway enrichment analysis was further carried out by String and Reactome, using all genes of high-impact variants causing stop-gain mutations. This approach allowed us to identify genes enriched in several relevant biological and molecular processes. Given that the genes were selected based on their high-impact consequences on protein function, the olfactory signalling pathway was identified as the most significant pathway, based on the enriched candidate genes. Being the largest multigene family in the human genome, with approximately 3% of total human genes, the role of olfactory receptors (ORs) in cancer has been disregarded, due to their specific role in the olfactory epithelium [61]. Recent studies have also identified genes for potential ORs as alternative genes for the treatment of cancer, including *OR51E2*, which is involved in the regulation and proliferation of prostate cancer [62]. Another olfactory receptor (OR), OR51B4, has been observed to be highly expressed in the colon cancer cell line HCT116 [63]. Although, among the ORs enriched in the pathway analysis, OR51B4 was not discovered in

our cohorts, our findings suggest that ORs could be potential genes for further exploration in colorectal cancer research.

While dealing with a vast number of candidate genes and variants, rare variants should be primarily considered for the prediction of functional effects [64]. In this study, synonymous variants and variants in non-coding regions were presumed to have non-functional impacts and were excluded [41], as well as variants that were less likely to be identified in the dbSNP. Among 16 genes harbouring rare missense mutations in the two first-degree relative patients, the best candidate genes and variants were then prioritized, based on in-silico prediction. Considering the detrimental effect and pathogenicity of these two mutations—c.1307C>T(p.Ala436Val) in the *PTCH2* gene and c.278T>A (p.Leu93His) in the *SEMA3D* gene—the functions of these genes were further delineated. A novel mutation in *PTCH2* was identified in an autosomal dominant disorder of naevoid basal cell carcinoma syndrome (NBCCS) in a Chinese population [65]. High *PTCH2* expression has also been observed in familial and sporadic basal cell carcinoma [66], and *PTCH2* is considered to be an important gene in murine medulloblastoma tumorigenesis [67]. However, the potential roles of this gene, specifically with regard to colorectal cancer predisposition, are yet unknown. However, the mutation c.1307C>T(p.Ala436Val) presented in these patients was predicted to block the protein function and may likely contribute to tumorigenesis. Several studies have determined the role of SEMA3D in the predisposition to several type of cancers, including breast cancer [68], glioma [69], and thyroid cancer [70]. Intriguingly, higher mRNA expression of *SEMA3D* mRNA has been observed in normal colorectal mucosa, as compared to the CRC tissues, suggesting that *SEMA3D* may function as a tumour suppressor gene in CRC progression [71]. *IFNE* has been identified as an apoptosis regulatory gene that can suppress cell proliferation in human colorectal cancer cells [72], despite its role in protecting the female reproductive tract against viral and bacterial infection [73].

Insertions and deletions are responsible for most genomic divergence, also in mammalian genomes [74]. Rare indels were identified in six genes—*CDK11B*, *CCDC144NL*, *GOLGA8R*, *MAFA*, *MUC6* and *PRIM2*—which harboured a total of eight rare functional indels. Two frameshift indels in *MUC6* caused a deletion (c.4712delC) and an insertion (c.4707_4708insA). In the normal colon, *MUC5AC* is rarely expressed, and there have been conflicting reports concerning the expression of *MUC6* in the colon [75]. However, *MUC6* expression has been reported to be associated with favourable outcomes in intermediate-stage (II and III) CRC patients [76]. Meanwhile, two frameshift indels were identified in the *CCDC144NL* gene. However, the gene function, with respect to predisposition towards cancer, remains unclear. Variations in the *CCDC144NL* gene were associated with poor prognosis and may facilitate cancer metastatic progression [77]. The *MAFA* gene, identified as harbouring a disruptive in-frame deletion in the seven considered patients, is known to be involved in oncogenic activities and cancer progression [78]. MAFA is one of the large Maf proteins that have been implicated in carcinogenesis, as demonstrated in cell culture, animal models, and cancer tissues [78]. Maf proteins have been identified to be involved in oncogenesis by the discovery of *v-maf* oncogene, which codes for the Maf protein member that causes fibrosarcoma in chickens [79]. *CDK11B*, characterized by an in-frame insertion in the proband, was known to likely be linked to predisposition to various human cancers [80]. *CDK11B* and its homologue gene encoding CDK11, a protein kinase that has been shown to be involved in the proliferation of various cancer cells, is involved in modulating the Wnt/β-catenin pathway in colon cancer [81]. The *PRIM2* gene, which is involved in synthesizing the Okazaki fragments in DNA replication, has been discovered as having the highest mutation rate in prostate cancer [82]. Among the several pathways enriched on the basis of the rare indels in the seven patients, defective GALTN12 signalling was identified as a significant pathway associated with colorectal cancer [83]. *GALNT12* was not identified in our cohorts; however, *MUC6*, a gene that encodes the mucin protein in epithelial tissue was enriched in this pathway. The GALNT family is classified as CAZy family GT27, and abnormality in one of the GALNT family genes, including

GALNT12, may result in reduced glycosylation of mucins [84]. Mucin genes are mainly expressed in digestive organs such as stomach, small intestine and colon and may play a role in colorectal cancer [85].

5. Conclusions

This study provides new insight into the gene variants related to CRC predisposition in a Malay population. However, the small number of patients and family members recruited in this study resulted in a small number of samples available for the analysis; therefore, it was challenging to elucidate the role of the identified variants in the pathogenicity of CRC in our Malay cohorts. The analysis of whole-genome data allowed the discovery of a new spectrum of variants, including candidate genes, and pathways. It would thus be beneficial to verify these findings in a larger cohort of patients, so to further validate them, carry out a functional analysis and rule out variant segregation. The whole-genome sequencing approach used in this study has provided new molecular knowledge of CRC in the considered cohort of Malay patients.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12091448/s1, Table S1: Alignment statistics of the whole genome probands, Table S2: Summary of SNVs in seven whole genome samples from CRC patients, Table S3: The most significant pathway enriched by genes harboured the nonsense mutation by Reactome tool in a) F1 b) F2 c) F5 d) F8 e) F12 f) F18 g) F19.

Author Contributions: Conceptualization, B.A.Z.; data curation, W.K.W.J.; formal analysis, W.K.W.J. and K.B.A.A.N.; funding acquisition, A.D.Z., W.M.M.W.M.M. and M.R.A.H.; resources, A.D.Z., W.M.M.W.M.M., M.R.A.H. and A.S.M.S.; supervision, K.B.A.A.N., W.F.W.A.R. and B.A.Z.; writing—original draft, W.K.W.J., K.B.A.A.N., A.D.Z. and W.F.W.A.R.; writing—review and editing, K.B.A.A.N., W.F.W.A.R., M.R.A.H., A.D.Z., W.M.M.W.M.M. and B.A.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Universiti Sains Malaysia Research University grant (1001/PPSP/812112) and a Universiti Sains Malaysia Short-Term grant (304/PPSP/61313202).

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the Institutional Review Board (or Ethics Committee) of Universiti Sains Malaysia (FWA Reg. No: 00007718; IRB Reg. No: 00004494).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank all the patients involved in this study.

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

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
2. Ferlay, J.; Soerjomataram, I.; Dikshit, R.; Eser, S.; Mathers, C.; Rebelo, M.; Parkin, D.M.; Forman, D.; Bray, F. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **2015**, *136*, E359–E386. [CrossRef]
3. Veettil, S.K.; Lim, K.G.; Chaiyakunapruk, N.; Ching, S.M.; Hassan, M.R.A. Colorectal cancer in Malaysia: Its burden and implications for a multiethnic country. *Asian J. Surg.* **2017**, *40*, 481–489. [CrossRef]
4. Hassan, M.R.A.; Khazim, W.K.W.; Othman, Z.; Mustapha, N.R.N.; Said, R.M.; Leong, T.W.; Suan, M.A.M.; Soelar, S.A. *The Second Report of the National Cancer Patient Registry-Colorectal Cancer, 2008–2013*; National Cancer Patient Registry-Colorectal Cancer and Clinical Research Centre (CRC): Kuala Lumpur, Malaysia, 2014.
5. Kastrinos, F.; Syngal, S. Inherited colorectal cancer syndromes. *Cancer J.* **2011**, *17*, 405. [CrossRef]
6. Cai, Q.; Sun, M.-H.; Lu, H.-F.; Zhang, T.-M.; Mo, S.-J.; Xu, Y.; Cai, S.-J.; Zhu, X.-Z.; Shi, D.-R. Clinicopathological and molecular genetic analysis of 4 typical Chinese HNPCC families. *World J. Gastroenterol.* **2001**, *7*, 805. [CrossRef]
7. Lynch, H.T.; De la Chapelle, A. Genetic susceptibility to non-polyposis colorectal cancer. *J. Med. Genet.* **1999**, *36*, 801–818. [PubMed]

8. Lynch, H.T.; Lynch, J.F.; Shaw, T.G.; Lubiński, J. HNPCC (Lynch Syndrome): Differential Diagnosis, Molecular Genetics and Management—A Review. *Hered. Cancer Clin. Pract.* **2003**, *1*, 7. [CrossRef]
9. Gala, M.; Chung, D.C. Hereditary colon cancer syndromes. In *Seminars in Oncology*; Elsevier: Omaha, NE, USA, 2011; pp. 490–499.
10. Jasperson, K.W.; Tuohy, T.M.; Neklason, D.W.; Burt, R.W. Hereditary and familial colon cancer. *Gastroenterology* **2010**, *138*, 2044–2058. [CrossRef]
11. Rahner, N.; Steinke, V. Hereditary cancer syndromes. *Dtsch. Ärzteblatt Int.* **2008**, *105*, 706. [CrossRef] [PubMed]
12. Valle, L. Genetic predisposition to colorectal cancer: Where we stand and future perspectives. *World J. Gastroenterol. WJG* **2014**, *20*, 9828. [CrossRef]
13. Jiao, X.; Liu, W.; Mahdessian, H.; Bryant, P.; Ringdahl, J.; Timofeeva, M.; Farrington, S.M.; Dunlop, M.; Lindblom, A. Recurrent, low-frequency coding variants contributing to colorectal cancer in the Swedish population. *PLoS ONE* **2018**, *13*, e0193547. [CrossRef]
14. Mardis, E.R.; Wilson, R.K. Cancer genome sequencing: A review. *Hum. Mol. Genet.* **2009**, *18*, R163–R168. [CrossRef]
15. Pleasance, E.D.; Cheetham, R.K.; Stephens, P.J.; McBride, D.J.; Humphray, S.J.; Greenman, C.D.; Varela, I.; Lin, M.-L.; Ordóñez, G.R.; Bignell, G.R. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **2010**, *463*, 191. [CrossRef]
16. Shanmugam, V.; Ramanathan, R.K.; Lavender, N.A.; Sinari, S.; Chadha, M.; Liang, W.S.; Kurdoglu, A.; Izatt, T.; Christoforides, A.; Benson, H. Whole genome sequencing reveals potential targets for therapy in patients with refractory KRAS mutated metastatic colorectal cancer. *BMC Med. Genom.* **2014**, *7*, 36. [CrossRef]
17. Kan, Z.; Zheng, H.; Liu, X.; Li, S.; Barber, T.; Gong, Z.; Gao, H.; Hao, K.; Willard, M.D.; Xu, J. Whole genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res.* **2013**, *23*, 1422–1433. [CrossRef] [PubMed]
18. Suwinski, P.; Ong, C.; Ling, M.H.; Poh, Y.M.; Khan, A.M.; Ong, H.S. Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Front. Genet.* **2019**, *10*, 49. [CrossRef]
19. Meienberg, J.; Bruggmann, R.; Oexle, K.; Matyas, G. Clinical sequencing: Is WGS the better WES? *Hum. Genet.* **2016**, *135*, 359–362. [CrossRef] [PubMed]
20. Juhari, W.K.W.; Rahman, W.F.W.A.; Sidek, A.S.M.; Hassan, M.R.A.; Ahmad, K.B.; Noordin, A.; Zakaria, A.D.; Macrae, F.; Zilfalil, B.A. Analysis of Hereditary Nonpolyposis Colorectal Cancer in Malay Cohorts using Immunohistochemical Screening. *Asian Pac. J. Cancer Prev.* **2015**, *16*, 3767–3771. [CrossRef]
21. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]
22. Li, R.; Li, Y.; Fang, X.; Yang, H.; Wang, J.; Kristiansen, K.; Wang, J. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **2009**, *19*, 1124–1132. [CrossRef] [PubMed]
23. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
24. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef]
25. Sherry, S.T.; Ward, M.-H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E.M.; Sirotkin, K. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308–311. [CrossRef] [PubMed]
26. Consortium, G.P. A map of human genome variation from population-scale sequencing. *Nature* **2010**, *467*, 1061. [CrossRef]
27. Huang, G.-H.; Tseng, Y.-C. Genotype imputation accuracy with different reference panels in admixed populations. In *BMC Proceedings*; BioMed Central: London, UK, 2014; p. S64.
28. Adzhubei, I.; Jordan, D.M.; Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **2013**, *76*, 7–20. [CrossRef] [PubMed]
29. Sim, N.-L.; Kumar, P.; Hu, J.; Henikoff, S.; Schneider, G.; Ng, P.C. SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **2012**, *40*, W452–W457. [CrossRef]
30. Reva, B.; Antipin, Y.; Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **2011**, *39*, e118. [CrossRef]
31. Forbes, S.A.; Beare, D.; Gunasekaran, P.; Leung, K.; Bindal, N.; Boutselakis, H.; Ding, M.; Bamford, S.; Cole, C.; Ward, S. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **2014**, *43*, D805–D811. [CrossRef] [PubMed]
32. Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Hoover, J. ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **2015**, *44*, D862–D868. [CrossRef]
33. Stenson, P.D.; Mort, M.; Ball, E.V.; Howells, K.; Phillips, A.D.; Thomas, N.S.; Cooper, D.N. The human gene mutation database: 2008 update. *Genome Med.* **2009**, *1*, 13. [CrossRef]
34. McLaren, W.; Pritchard, B.; Rios, D.; Chen, Y.; Flicek, P.; Cunningham, F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **2010**, *26*, 2069–2070. [CrossRef]
35. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P. STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2014**, *43*, D447–D452. [CrossRef] [PubMed]

36. Croft, D.; Mundo, A.F.; Haw, R.; Milacic, M.; Weiser, J.; Wu, G.; Caudy, M.; Garapati, P.; Gillespie, M.; Kamdar, M.R. The Reactome pathway knowledgebase. *Nucleic Acids Res.* **2013**, *42*, D472–D477. [CrossRef] [PubMed]
37. Lupski, J.R.; Reid, J.G.; Gonzaga-Jauregui, C.; Rio Deiros, D.; Chen, D.C.; Nazareth, L.; Bainbridge, M.; Dinh, H.; Jing, C.; Wheeler, D.A. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* **2010**, *362*, 1181–1191. [CrossRef] [PubMed]
38. O'Rawe, J.; Jiang, T.; Sun, G.; Wu, Y.; Wang, W.; Hu, J.; Bodily, P.; Tian, L.; Hakonarson, H.; Johnson, W.E. Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. *Genome Med.* **2013**, *5*, 28. [CrossRef] [PubMed]
39. Clark, M.J.; Chen, R.; Lam, H.Y.; Karczewski, K.J.; Chen, R.; Euskirchen, G.; Butte, A.J.; Snyder, M. Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **2011**, *29*, 908. [CrossRef]
40. Zahir, F.R.; Mwenifumbo, J.C.; Chun, H.-J.E.; Lim, E.L.; Van Karnebeek, C.D.; Couse, M.; Mungall, K.L.; Lee, L.; Makela, N.; Armstrong, L. Comprehensive whole genome sequence analyses yields novel genetic and structural insights for Intellectual Disability. *BMC Genom.* **2017**, *18*, 403. [CrossRef]
41. Sefid Dashti, M.J.; Gamieldien, J. A practical guide to filtering and prioritizing genetic variants. *BioTechniques* **2017**, *62*, 18–30. [CrossRef] [PubMed]
42. Jiang, L.; Zhang, C.; Li, Y.; Yu, X.; Zheng, J.; Zou, P.; Li, Y.; Bin, X.; Lu, J.; Zhou, Y. A non-synonymous polymorphism Thr115Met in the *EpCAM* gene is associated with an increased risk of breast cancer in Chinese population. *Breast Cancer Res. Treat.* **2011**, *126*, 487–495. [CrossRef]
43. Baeuerle, P.; Gires, O. *EpCAM* (CD326) finding its role in cancer. *Br. J. Cancer* **2007**, *96*, 417–423. [CrossRef]
44. Kloor, M.; Voigt, A.Y.; Schackert, H.K.; Schirmacher, P.; von Knebel Doeberitz, M.; Bläker, H. Analysis of *EPCAM* protein expression in diagnostics of Lynch syndrome. *J. Clin. Oncol.* **2011**, *29*, 223–227. [CrossRef]
45. Ligtenberg, M.J.; Kuiper, R.P.; Chan, T.L.; Goossens, M.; Hebeda, K.M.; Voorendt, M.; Lee, T.Y.; Bodmer, D.; Hoenselaar, E.; Hendriks-Cornelissen, S.J. Heritable somatic methylation and inactivation of *MSH2* in families with Lynch syndrome due to deletion of the 3′ exons of *TACSTD1*. *Nat. Genet.* **2009**, *41*, 112. [CrossRef] [PubMed]
46. Kempers, M.J.; Kuiper, R.P.; Ockeloen, C.W.; Chappuis, P.O.; Hutter, P.; Rahner, N.; Schackert, H.K.; Steinke, V.; Holinski-Feder, E.; Morak, M. Risk of colorectal and endometrial cancers in *EPCAM* deletion-positive Lynch syndrome: A cohort study. *Lancet Oncol.* **2011**, *12*, 49–55. [CrossRef]
47. Guarinos, C.; Castillejo, A.; Barberá, V.-M.; Pérez-Carbonell, L.; Sánchez-Heras, A.-B.; Segura, Á.; Guillén-Ponce, C.; Martínez-Cantó, A.; Castillejo, M.-I.; Egoavil, C.-M. *EPCAM* germ line deletions as causes of Lynch syndrome in Spanish patients. *J. Mol. Diagn.* **2010**, *12*, 765–770. [CrossRef] [PubMed]
48. Ligtenberg, M.J.; Kuiper, R.P.; van Kessel, A.G.; Hoogerbrugge, N. *EPCAM* deletion carriers constitute a unique subgroup of Lynch syndrome patients. *Fam. Cancer* **2013**, *12*, 169–174. [CrossRef] [PubMed]
49. Tutlewska, K.; Lubinski, J.; Kurzawski, G. Germline deletions in the *EPCAM* gene as a cause of Lynch syndrome–literature review. *Hered. Cancer Clin. Pract.* **2013**, *11*, 9. [CrossRef]
50. Jia, W.; Lu, R.; Martin, T.A.; Jiang, W.G. The role of claudin-5 in blood-brain barrier (BBB) and brain metastases. *Mol. Med. Rep.* **2014**, *9*, 779–785. [CrossRef] [PubMed]
51. Cornely, R.M.; Schlingmann, B.; Shepherd, W.S.; Chandler, J.D.; Neujahr, D.C.; Koval, M. Two common human CLDN5 alleles encode different open reading frames but produce one protein isoform. *Ann. Acad. Sci.* **2017**, *1397*, 119–129. [CrossRef]
52. Sherry, S.T.; Ward, M.; Sirotkin, K. dbSNP—Database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **1999**, *9*, 677–679. [PubMed]
53. Allen, I.C.; Eden, K.; Heid, B.; Holl, E.K. Map3K14 signaling attenuates the development of colorectal cancer through activation of the non-canonical NF-κB signaling cascade. *Am. Assoc. Immnol.* **2017**, *198* (Suppl. 1), 197.6.
54. Storz, P. Targeting the alternative NF-κB pathway in pancreatic cancer: A new direction for therapy? *Expert Rev. Anticancer Ther.* **2013**, *13*, 501–504. [CrossRef]
55. Wharry, C.E.; Haines, K.M.; Carroll, R.G.; May, M.J. Constitutive noncanonical NFκB signaling in pancreatic cancer cells. *Cancer Biol. Ther.* **2009**, *8*, 1567–1576. [CrossRef]
56. Marchesi, F.; Monti, P.; Leone, B.E.; Zerbi, A.; Vecchi, A.; Piemonti, L.; Mantovani, A.; Allavena, P. Increased survival, proliferation, and migration in metastatic human pancreatic tumor cells expressing functional CXCR4. *Cancer Res.* **2004**, *64*, 8420–8427. [CrossRef]
57. Koslowski, M.; Sahin, U.; Dhaene, K.; Huber, C.; Türeci, Ö. MS4A12 is a colon-selective store-operated calcium channel promoting malignant cell processes. *Cancer Res.* **2008**, *68*, 3458–3466. [CrossRef] [PubMed]
58. Chen, L.; Chen, Q.; Wu, Y.; Zhu, M.; Hu, J.; Zhuang, Z. MTSS1 inhibits colorectal cancer metastasis by regulating the CXCR4/CXCL12 signaling axis. *Int. J. Mol. Med.* **2021**, *47*, 1–13. [CrossRef] [PubMed]
59. Han, W.; Hu, C.; Fan, Z.-J.; Shen, G.-L. Transcript levels of keratin 1/5/6/14/15/16/17 as potential prognostic indicators in melanoma patients. *Sci. Rep.* **2021**, *11*, 1–12.
60. Savas, S.; Tuzmen, S.; Ozcelik, H. Human SNPs resulting in premature stop codons and protein truncation. *Hum. Genom.* **2006**, *2*, 274. [CrossRef] [PubMed]
61. Ranzani, M.; Iyer, V.; Ibarra-Soria, X.; Velasco-Herrera, M.D.C.; Garnett, M.; Logan, D.; Adams, D.J. Revisiting olfactory receptors as putative drivers of cancer. *Wellcome Open Res.* **2017**, *2*. [CrossRef]

62. Xu, L.L.; Stackhouse, B.G.; Florence, K.; Zhang, W.; Shanmugam, N.; Sesterhenn, I.A.; Zou, Z.; Srikantan, V.; Augustus, M.; Roschke, V. PSGR, a novel prostate-specific gene with homology to a G protein-coupled receptor, is overexpressed in prostate cancer. *Cancer Res.* **2000**, *60*, 6568–6572. [PubMed]
63. Weber, L.; Al-Refae, K.; Ebbert, J.; Jägers, P.; Altmüller, J.; Becker, C.; Hahn, S.; Gisselmann, G.; Hatt, H. Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis. *PLoS ONE* **2017**, *12*, e0172491. [CrossRef] [PubMed]
64. Nelson, M.R.; Wegmann, D.; Ehm, M.G.; Kessner, D.; Jean, P.S.; Verzilli, C.; Shen, J.; Tang, Z.; Bacanu, S.-A.; Fraser, D. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **2012**, *337*, 100–104. [CrossRef]
65. Fan, Z.; Li, J.; Du, J.; Zhang, H.; Shen, Y.; Wang, C.-Y.; Wang, S. A missense mutation in PTCH2 underlies dominantly inherited NBCCS in a Chinese family. *J. Med. Genet.* **2008**, *45*, 303–308. [CrossRef]
66. Zaphiropoulos, P.G.; Undén, A.B.; Rahnama, F.; Hollingsworth, R.E.; Toftgård, R. PTCH2, a Novel Human Patched Gene, Undergoing Alternative Splicing andUp-regulated in Basal Cell Carcinomas. *Cancer Res.* **1999**, *59*, 787–792.
67. Lee, Y.; Miller, H.L.; Jensen, P.; Hernan, R.; Connelly, M.; Wetmore, C.; Zindy, F.; Roussel, M.F.; Curran, T.; Gilbertson, R.J. A molecular fingerprint for medulloblastoma. *Cancer Res.* **2003**, *63*, 5428–5437.
68. Kigel, B.; Varshavsky, A.; Kessler, O.; Neufeld, G. Successful inhibition of tumor development by specific class-3 semaphorins is associated with expression of appropriate semaphorin receptors by tumor cells. *PLoS ONE* **2008**, *3*, e3287. [CrossRef]
69. Karayan-Tapon, L.; Wager, M.; Guilhot, J.; Levillain, P.; Marquant, C.; Clarhaut, J.; Potiron, V.; Roche, J. Semaphorin, neuropilin and VEGF expression in glial tumours: SEMA3G, a prognostic marker? *Br. J. Cancer* **2008**, *99*, 1153. [CrossRef] [PubMed]
70. Gomez-Rueda, H.; Palacios-Corona, R.; Gutiérrez-Hermosillo, H.; Trevino, V. A robust biomarker of differential correlations improves the diagnosis of cytologically indeterminate thyroid cancers. *Int. J. Mol. Med.* **2016**, *37*, 1355–1362. [CrossRef] [PubMed]
71. Wang, Z.; Ding, M.; Qian, N.; Song, B.; Yu, J.; Tang, J.; Wang, J. Decreased expression of semaphorin 3D is associated with genesis and development in colorectal cancer. *World J. Surg. Oncol.* **2017**, *15*, 67. [CrossRef] [PubMed]
72. Paschall, A.V.; Liu, K. Epigenetic regulation of apoptosis and cell cycle regulatory genes in human colon carcinoma cells. *Genom. Data* **2015**, *5*, 189–191. [CrossRef]
73. Nickerson, M.L.; Witte, N.; Im, K.M.; Turan, S.; Owens, C.; Misner, K.; Tsang, S.X.; Cai, Z.; Wu, S.; Dean, M. Molecular analysis of urothelial cancer cell lines for modeling tumor biology and drug response. *Oncogene* **2017**, *36*, 35–46. [CrossRef] [PubMed]
74. Tao, S.; Fan, Y.; Wang, W.; Ma, G.; Liang, L.; Shi, Q. Patterns of insertion and deletion in mammalian genomes. *Curr. Genom.* **2007**, *8*, 370–378. [CrossRef] [PubMed]
75. De Bolós, C.; Garrido, M.; Real, F.X. MUC6 apomucin shows a distinct normal tissue distribution that correlates with Lewis antigen expression in the human stomach. *Gastroenterology* **1995**, *109*, 723–734. [CrossRef]
76. Betge, J.; Schneider, N.I.; Harbaum, L.; Pollheimer, M.J.; Lindtner, R.A.; Kornprat, P.; Ebert, M.P.; Langner, C. MUC1, MUC2, MUC5AC, and MUC6 in colorectal cancer: Expression profiles and clinical significance. *Virchows Arch.* **2016**, *469*, 255–265. [CrossRef] [PubMed]
77. Qu, L.-W.; Zhou, B.; Wang, G.-Z.; Chen, Y.; Zhou, G.-B. Genomic variations in paired normal controls for lung adenocarcinomas. *Oncotarget* **2017**, *8*, 104113. [CrossRef]
78. Eychène, A.; Rocques, N.; Pouponnot, C. A new MAFia in cancer. *Nat. Rev. Cancer* **2008**, *8*, 683. [CrossRef] [PubMed]
79. Kawai, S.; Goto, N.; Kataoka, K.; Saegusa, T.; Shinno-Kohno, H.; Nishizawa, M. Isolation of the avian transforming retrovirus, AS42, carrying the v-maf oncogene and initial characterization of its gene product. *Virology* **1992**, *188*, 778–784. [CrossRef]
80. Zhou, Y.; Shen, J.K.; Hornicek, F.J.; Kan, Q.; Duan, Z. The emerging roles and therapeutic potential of cyclin-dependent kinase 11 (CDK11) in human cancer. *Oncotarget* **2016**, *7*, 40846. [CrossRef]
81. Naik, S.; Dothager, R.S.; Marasa, J.; Lewis, C.L.; Piwnica-Worms, D. Vascular endothelial growth factor receptor-1 is synthetic lethal to aberrant β-catenin activation in colon cancer. *Clin. Cancer Res.* **2009**, *15*, 7529–7537. [CrossRef]
82. Lindquist, K.J.; Paris, P.L.; Hoffmann, T.J.; Cardin, N.J.; Kazma, R.; Mefford, J.A.; Simko, J.P.; Ngo, V.; Chen, Y.; Levin, A.M. Mutational landscape of aggressive prostate tumors in African American men. *Cancer Res.* **2016**, *76*, 1860–1868. [CrossRef]
83. Fabregat, A.; Jupe, S.; Matthews, L.; Sidiropoulos, K.; Gillespie, M.; Garapati, P.; Haw, R.; Jassal, B.; Korninger, F.; May, B. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2018**, *46*, D649–D655. [CrossRef]
84. Abulí, A.; Fernández-Rozadilla, C.; Alonso-Espinaco, V.; Muñoz, J.; Gonzalo, V.; Bessa, X.; González, D.; Clofent, J.; Cubiella, J.; Morillas, J.D. Case-control study for colorectal cancer genetic susceptibility in EPICOLON: Previously identified variants and mucins. *BMC Cancer* **2011**, *11*, 1–8. [CrossRef] [PubMed]
85. Fernandez, A.J.; Daniel, E.J.P.; Mahajan, S.P.; Gray, J.J.; Gerken, T.A.; Tabak, L.A.; Samara, N.L. The structure of the colorectal cancer-associated enzyme GalNAc-T12 reveals how nonconserved residues dictate its function. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 20404–20410. [CrossRef] [PubMed]