 doi:10.20944/preprints202102.0040.v1

Article

Whole exome sequencing identifies APCDD1 and HDAC5 genes as potentially cancer predisposing in familial colorectal cancer

Diamanto Skopelitou [1,2,3,4], Beiping Miao [1,2,3], Aayushi Srivastava [1,2,3,4], Abhishek Kumar [1,5,6], Magdalena Kuswick [7], Dagmara Dymerska [7], Nagarajan Paramasivam [8], Matthias Schlesner [9], Jan Lubinski [7], Kari Hemminki [1,10,11], Asta Försti [1,2,3,] and Obul Reddy Bandapalli [1,2,3,4,*]

1 Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
2 Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany
3 Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany
4 Medical Faculty, Heidelberg University, Heidelberg, Germany
5 Institute of Bioinformatics, International Technology Park, Bangalore, India
6 Manipal Academy of Higher Education (MAHE), Manipal 576104, Karnataka, India
7 Department of Genetics and Pathology, Pomeranian Medical University in Szczecin, Poland
8 Computational Oncology, Molecular Diagnostics Program, National Center for Tumor Diseases (NCT), Germany
9 Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany
10 Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
11 Faculty of Medicine and Biomedical Center in Pilsen, Charles University in Prague, 30605 Pilsen, Czech Republic

* Correspondence: o.bandapalli@kitz-heidelberg.de; Tel.: +49-6221-421809

Abstract: Germline mutations in predisposition genes account for only 20% of all familial colorectal cancers (CRC) and the remaining genetic burden may be due to rare high-to-moderate-penetrance germline variants that are not explored. With the aim of identifying such potential cancer predisposing variants, we performed whole exome sequencing on three CRC cases and three unaffected members of a Polish family and identified two novel heterozygous variants; a coding variant in APC down-regulated 1 gene (*APCDD1*, p.R299H) and a non-coding variant in the 5′ untranslated region (UTR) of histone deacetylase 5 gene (*HDAC5*). Sanger sequencing confirmed the variants segregating with the disease and Taqman assays revealed 8 additional *APCDD1* variants in a cohort of 1705 familial CRC patients and no further *HDAC5* variants. Proliferation assays indicated an insignificant proliferative impact for the *APCDD1* variant. Luciferase reporter assays using the *HDAC5* variant resulted in an enhanced promoter activity. Targeting of transcription factor binding sites of SNAI-2 and TCF4 interrupted by the *HDAC5* variant showed a significant impact of TCF4 on promoter activity of mutated *HDAC5*. Our findings contribute not only to the identification of unrecognized genetic causes of familial CRC but also underline the importance of 5′UTR variants affecting transcriptional regulation and the pathogenesis of complex disorders.

Keywords: APCDD1; HDAC5; 5′UTR; germline variant; familial colorectal cancer; whole exome sequencing; promoter activity

1. Introduction

Whole exome sequencing (WES) is gaining relevance for molecular genetic research of familial cancer and the identification of new cancer predisposing variants. Among hereditary malignancies, colorectal cancer (CRC) shows one of the highest proportions of

 doi:10.20944/preprints202102.0040.v1

familial cases and heritable factors have been estimated to account for about 35% of the CRC risk, according to twin studies [1]. Cancer-predisposing germline mutations of *APC, MUTYH* and mismatch repair genes are already known to be associated with familial CRC and to lead to phenotypes of well-defined Mendelian CRC syndromes, familial adenomatous polyposis (FAP), resulting from *APC* gene mutations; *MUTYH*-associated polyposis (MAP); and Lynch syndrome, a hereditary non-polyposis colon cancer (HNPCC) syndrome caused by mismatch repair gene mutations (*MLH1, MSH2, MSH6, PMS2* and *EPCAM*) [2]. CRC fulfilling the diagnostic criteria of Lynch syndrome but not linked to pathogenic mismatch repair gene mutations or the resulting microsatellite instability has been classified as Familial Colorectal Cancer Type X (FCCTX). Although FCCTX is considered as a heterogeneous group with unknown genetic etiology, candidate genes such as *CENPE, KIF24, GALNT12, ZNF367, GABBR2* and *BMP4* have been suggested as cancer predisposing in these patients [3]. Sequencing studies have further identified germline variants in *HNRNPA0* and *WIF1* genes in a family with susceptibility to multiple early onset cancers including CRC [4] as well as a germline mutation in *NTHL1* gene in three unrelated families with adenomatous polyposis and various cancer types including CRC [5, 6]. Moreover, germline as well as somatic mutations in *POLE* and *POLD1* genes have been associated with both sporadic and familial CRC contributing to the genetic understanding of CRC inheritance [7, 8]. Nevertheless, germline mutations in these or other established predisposition genes account for only 5 to 10% of all CRC [9]. Since the genetic background of most familial CRC cases has still not been sufficiently explored, the application of WES on these patients within pedigree-based studies bears great potential for the exploration of the remaining genetic burden.
As 98% of the genome is non-coding, a high proportion of variants is identified in this region [10] and gaining relevance in the understanding of inherited cancer predisposition [11, 12]. The great potential impact of variants within the 5′ untranslated region (UTR) of a gene or up to 1 kb upstream of transcription start sites can be attributed to possible changes in transcriptional regulatory elements, such as binding motifs in promoters, enhancers or super-enhancers. On the other hand, the 3′UTR of a gene and the flanking region downstream of transcription end sites can carry essential miRNA binding sites where small RNAs containing the respective complementary sequence can post-transcriptionally attach; hence mRNA translation can be suppressed leading to the inhibition of gene expression [13]. By this and many other means, non-coding regions can play an important role in transcriptional and post-transcriptional regulation of gene expression, which is why genetic variation of non-coding DNA has to be considered in the analysis and prioritization of potentially cancer-causing variants.

We have developed the Familial Cancer Variant Prioritization Pipeline (FCVPPv2) for evaluation of both coding and non-coding variants and implemented it in the prioritization of novel missense variants in the tumor suppressor genes *DICER1* in Hodgkin Lymphoma and *CPXM1* in papillary thyroid cancer and in the pathways enriched in these entities [14-17]. In the present study, tools such as the Combined Annotation Dependent Depletion v1.4 (CADD) tool [18], SNPnexus [19] and the Bedtools intersect function were applied as part of the non-coding analysis of our pipeline in order to identify important regulatory elements. Using FCVPPv2 and literature mining, we were able to prioritize two novel heterozygous variants in a family affected by CRC, a coding variant in the APC down-regulated 1 (*APCDD1*) gene and a non-coding variant in the histone deacetylase 5 (*HDAC5*) gene. Whereas the *APCDD1* variant was identified in 8 additional cases among 1705 CRC families, cell proliferation assays indicated an insignificant proliferative impact for the variant. We did not find any other familial CRC cases with the *HDAC5* variant, but functional experiments showed a significant impact of the 5′UTR variant on expression of *HDAC5*, involved in cellular processes such as proliferation, differentiation, apoptosis and cell cycle progression. Luciferase reporter assays resulted in enhanced promoter activity of the *HDAC5* gene carrying this variant compared

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

to the wild type sequence and targeting of transcription factor binding sites interrupted by this variant showed an impact of TCF4 on promoter activity of mutated *HDAC5*.

2. Results

2.1. FCVPPv2 analysis of coding variants prioritized a missense variant in APCDD1 gene

Application of WES on the studied CRC-affected family identified 13,733 variants with MAF ≤ 0.1%. Filtering according to the probability for each family member of being a Mendelian case (Figure 1, Supplementary Table S1) narrowed down the number of identified variants to 783. For analysis of coding variants (*n = 101*), synonymous variants (*n = 35*) were filtered out as they are generally considered to play a minor role in the development of diseases and cancer. The remaining 66 nonsynonymous variants, frameshift deletions/insertions or variants of unknown significance were further evaluated. Application of the PHRED-like CADD score cut-off of ≥ 10 reduced the number of variants to 51 and screening according to the three conservational scores GERP, PhastCons and PhyloP further narrowed down the number to 38 variants. Eighteen variants were annotated with at least 3 out of 4 intolerance scores as being favorable and were further considered for deleteriousness screening. Since 12 of the variants did not fulfill the criterion of being annotated as deleterious by at least 60% of all deleteriousness scores, they were excluded. Last, 7 nonsynonymous variants remained passing all the criteria and considering a MAF of 0.1% in the non-Finnish European population of gnomAD database: *APCDD1* (p.R299H), *FLNC* (p.G553S), *KCNH6* (p.L403V), *LSR* (p.A139D), MTX1 (p.Y228C), *SDS* (p.T185I), *ZW10* (p.A732P) (Table 1).

SNAP² analysis indicated a functional effect of the amino acid substitutions induced by the variants in the *APCDD1* (p.R299H), *ZW10* (p.A732P) and *LSR* (p.A139D) genes by predicting scores of > 50, whereas application of CGI did not identify any cancer drivers among the variants. The lipolysis stimulated lipoprotein receptor encoded by the *LSR*

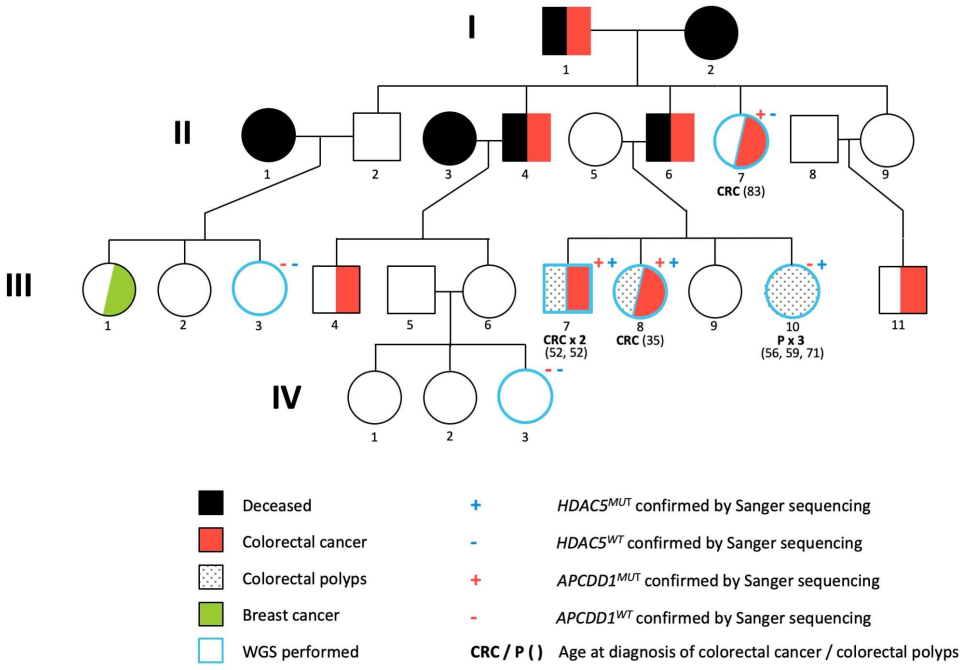

Figure 1. Pedigree of the CRC family with *APCDD1* and *HADC5* variants analyzed in this study.

Table 1. Overview of the top exonic missense variants prioritized in the studied CRC family. Chromosomal positions, pedigree segregation, allele frequencies in the Non-Finnish European population (*NFE*), PHRED-like CADD scores, conservational scores and the percentage of reached intolerance and deleteriousness scores are summarized for each variant. Snap² results for the predicted amino acid changes with calculated effect scores and accuracies given in % as well as CGI predictions are included. Respective protein functions of the encoded gene products are derived from Genecards [20]. *ND - no driver mutation; . – not annotated*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

* Following predictions given by deleteriousness scores were considered as favorable in our analysis: SIFT – Damaging (D); Polyphen2_HumDiv, Polyphen2_HumVar – Probably damaging (D) & Possibly damaging (P); LRT – Deleterious (D); MutationTaster – Disease causing (D) & disease causing automatic (A); MutationAssesor – High (H) & medium (M); FATHMM – Damaging (D); MetaSVM – Damaging (D); MetaLR – Damaging (D); Reliability Index ≥ 5; VEST3 ≥ 0.5; PROVEAN – Damaging (D).

gene is generally known to play a role in metabolism by inducing the uptake of triglyc-

Gene name	variant	Pedigree segregation	Allele frequency in NFE		CADD SCORE	Conservational scores			Intolerance scores (%)	Deleteriousness scores* (%)	Amino acid change	Snap²		CGI	Protein function
			ExAC	gnomAD		GERP	PhyloP	PhastCons				Score	Accuracy (%)		
APCDD1	18_10485580_G_A	III7, III8, III10	7.37 E-05	1.4 E-04	27.5	4.94	9.26	1	100	66.67	R299H	54	75	ND	Inhibition of Wnt signaling, controversial function
FLNC	7_128480709_G_A	III7, III8, III10	5.25 E-04	2.57 E-04	29.3	5.02	9.82	1	100	100	G553S	33	66	ND	Anchoring of membrane proteins for actin cytoskeleton
KCNH6	17_61613135_T_G	III7, III8, III10	4.51 E-05	6.16 E-05	23.4	3.39	3.96	1	100	91.67	L403V	6	53	ND	Regulation of neurotransmitter release, neuronal excitability, epithelial electrolyte transport
LSR	19_35741380_C_A	II7, III7, III8, III10	.	.	27.4	4.88	7.21	1	75	75	A139D	80	91	ND	Lipoprotein metabolism
MTX1	1_155181922_A_G	III7, III8, III10	9.25 E-05	1.52 E-04	23.3	3.64	2.89	1	66.67	66.67	Y228C	32	66	ND	Mitochondrial protein import
SDS	12_113836573_G_A	III7, III8, III10	4.59 E-05	5.44 E-05	23.4	4	5.69	1	75	100	T91I	46	71	ND	Serine and glycine metabolism, gluconeogenesis
ZW10	11_113607367_C_G	III7, III8, III10	5.99 E-04	6.43 E-04	34	6.17	7.49	1	75	75	A732P	70	85	ND	Chromosome segregation, mitotic checkpoint

eride-rich lipoproteins like chylomicrons, LDL and VLDL from blood into cells. Since further literature mining did not reveal any association with colorectal carcinogenesis, the identified variant in *LSR* gene was considered to be of minor impact on the development of CRC in the studied family. *ZW10* gene is encoding a protein of the mitotic checkpoint controlling chromosome segregation during cell division. On the background of causing chromosomal instability when mutated in a model system, Wang et al. have identified two somatic missense variants in *ZW10* gene (p.N123T, p.S623G) in a panel of CRCs [21]. As the prioritized *ZW10* variant identified in the studied family (p.A732P) is not located in the adjacent regions and is moreover located close to the end of the protein (779aa), its functional and potentially cancer-predisposing impact was considered as improbable.

The *APCDD1* variant is located in the second of two functional APCDDC domains (51-283, 284-490), according to Interpro, Pfam and SMART [22-24], required for interaction with Wnt ligands and their receptors. Since *APCDD1* gene has been linked to CRC by being involved in the Wnt signaling pathway as a direct target of the beta-Catenin/TCF4 complex, the associated variant (p.R299H) was prioritized as the top cancer-predisposing candidate of all identified missense variants (Figure 2) [25]. Pedigree segregation of the *APCDD1* variant was checked by IGV and further confirmed by targeted Sanger sequencing showing the heterozygous variant (p.R299H) for the two CRC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

cases (III7, III8) and the individual with polyps (III10) and the wild type sequence for II7 (CRC at the age of 83 years) and the two controls of the family (III3 and IV3), respectively (Supplementary Figure S1a).

2.2. FCVPPv2 analysis of non-coding variants prioritized a 5'UTR variant in HDAC5 gene

In agreement with the high proportion of non-coding DNA in the human genome, 674 of the 783 pedigree-filtered variants (86%) were located in the non-coding sequence such as intronic and intergenic regions, 1kb downstream and upstream regions, the 5'UTR, 3'UTR and non-coding RNAs (ncRNAs), respectively (Figure 2). Screening of the 174 upstream/5'UTR as well as downstream/3'UTR variants, excluding intronic, intergenic and ncRNAs variants, for an updated PHRED-like CADD score ≥ 10 resulted in 38 variants. Filtering with conservational scores narrowed down this number to 8 5'UTR and 18 3'UTR variants. Application of non-coding scores derived from SNPnexus revealed that all 8 5'UTR variants reached at least 50% of the cut-off values, whereas 2 out of 18 3'UTR variants were excluded due to insufficient non-coding scores. The remaining 24 non-coding candidates were further evaluated for the presence of specific regulatory elements such as TFBS and CpG islands for 5'UTR and miRNA binding sites for 3'UTR variants. Since all 5'UTR candidates showed either a CpG island or a TFBS identified by SNPnexus, CADD v1.4 or bedtools intersect function, our analysis resulted in 8 top 5'UTR variants, summarized in Table 2.

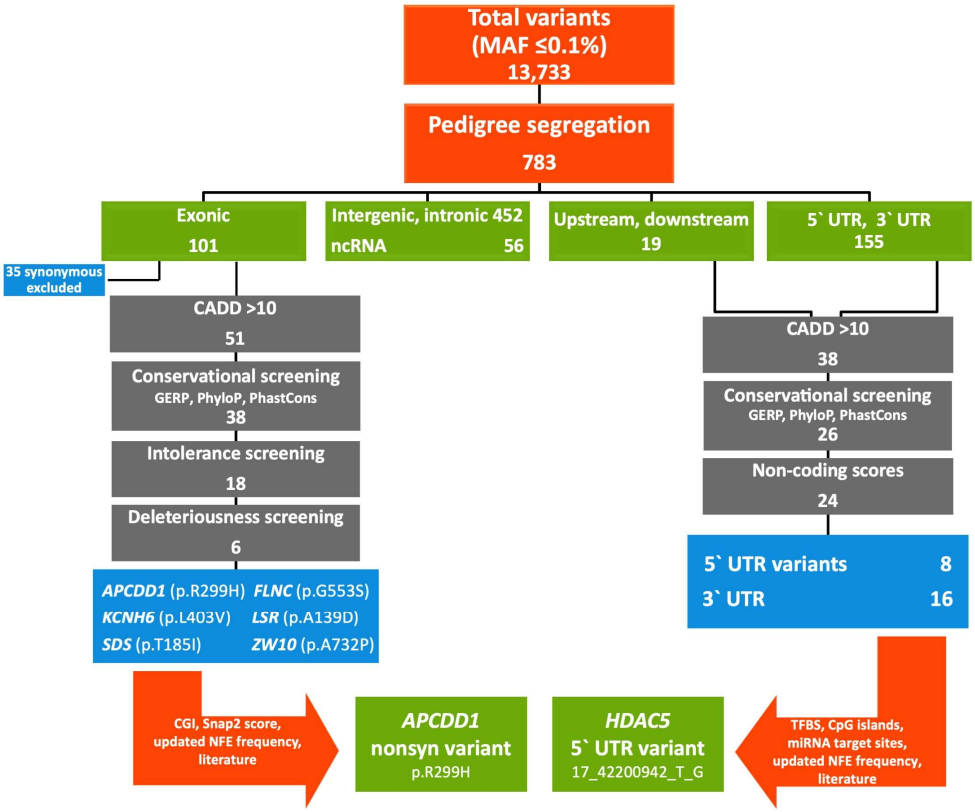

Figure 2 Graphical representation of the filtering process according to the Familial Cancer Variant Prioritization Pipeline version 2 (FCVPPv2).

Table 2. 5'UTR germline variants prioritized in the CRC family. Chromosomal position, pedigree segregation and allele frequencies in the Non-Finnish European population (*NFE*) are listed for each variant. Identified promoters, enhancers, super-enhancers according to Bedtools intersect function and SEA, FANTOM5 databases are included. Overall deleteriousness, genomic conservation, chromatin state and location within transcription factor binding sites (TFBS) are annotated by CADD v1.4. Further information on TFBS, CpG islands and the general functional impact in form of a summarizing percentage of positive non-coding scores are derived from SNPnexus. . – *not annotated*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

Gene name	Variant	Pedigree segregation	Allele frequency in NFE		Bedtools intersect[I]	CADD v1.4									SNPnexus			
			ExAC	gnomAD NFE		CADD SCORE	Conservational scores			Chromatin State			TFBS	TFBS peaks[IV]	Non coding scores[V] (%)	CpG island	CpG ratio[VI]	TFBS
							PhastCons	PhyloP	GERP	ChromHMM[II] state	ChromHMM[II] score	Segway[III]						
CA4	17_58227298_G_C	III7, III8, III10	.	.	P:58227287..58227313,+	12.6	0.44	0.72	2.77	Enh Biv	0.24	TF0	1	4	50	61	0.89	.
CALM3	19_47105342_A_G	II7, III7, III8, III10	.	.	.	19.8	0.87	1.94	2.49	Tss A	0.87	GS	1	1	50	.	.	.
HDAC5	17_42200942_T_G	II7, III7, III8	.	6.64 E-05	.	21.9	1.00	0.87	3.99	Tss A	0.98	TSS	17	42	50	92	0.97	
PLAA	9_26947165_G_A	III7, III8, III10	.	6.49 E-05	P:26947129..26947212,-	22	0.99	0.92	4.8	Tss A	0.94	TSS	41	61	83.3	72	0.86	NRF 2
PPTC7	12_111021082_G_C	III7, III8, III10	.	2.69 E-04	SE: hg19_A549_212_111015565	16.5	1.00	3.85	4.29	Tss A	0.98	TSS	26	50	50	83	1.13	.
TMEM 115	3_50396814_C_G	III7, III8, III10	.	1.69 E-03	.	18.1	1.00	1.89	4.82	Tss A	0.82	GS	18	27	66.7	64	0.67	.
TPM2	9_35690678_C_T	III7, III8, III10	.	1.36 E-03	.	17.2	1.00	0.61	3.76	Tss A	0.95	GE 2	6	6	50	109	0.76	.
UBE2K	4_39699921_G_C	III7, III8, III10	.	1.16 E-04	.	17.0	1.00	1.39	4.47	Tss A	0.93	TSS	14	25	66.7	89	1.01	.

*I Bedtools intersect:*Promoters (P) are listed with their specific genomic position (P:start..end,strand). Super-enhancers (SE) are shown including information about the used reference genome, cell line and genomic position.
II ChromHMM: The ChromHmm score shows the proportion of 127 cell types of the Roadmap Epigenomics project in a particular chromatin state with scores closer to 1 indicating more cell types in the particular chromatin state. The 15 chromatin states are defined as follows: TssA – Active transcription start site (TSS), TssAFlnk – Flanking active TSS, TxFlnk – Transcribed at gene 5' and 3', Tx – Strong transcription, TxWk – Weak transcription, EnhG – Genic enhancers, Enh – Enhancers, ZNF/Rpts – ZNF genes & repeats, Het – Heterochromatin, TssBiv – Bivalent/Poised TSS/Enhancers, BivFlnk – Flanking bivalent TSS/Enhancer, EnhBiv – Bivalent enhancers, ReprPC – Repressed PolyComb, ReprPCWk – Weak Repressed PolyComb, Quies – Quiescent/low [26, 27].
III Segway: Segway uses a genomic segmentation method to annotate the chromatin state based on multiple datasets of ChIP-seq experiments. The chromatin states can be annotated as follows: D – dead, F0/1 – FAIRE, R0/1/2/4/5 – Repressed Region, H3K9me1 – histone 3 lysine 9 monomethylation, L0/1 – Low zone, GE0/1/2 – Gene body (end),TF0/1/2 – Transcription factor activity, C0 – CTCF, GS – Gene body (start), E/GM – Enhancer/gene middle, GM0/1 – Gene body (middle), TSS – Transcription start site, ZnfRpts – zinc finger repeats [28].
IV TFBS peaks: The number of overlapping ChIP TFBS peaks summed over different cell types/tissue.
V Non-coding scores: Following cut-offs were used for the interpretation of the non-coding scores derived from SNPnexus: FitCons Score ≥ 0.2 with a P-Value ≤ 0.05; EIGEN > 0 (at least 1 of 2 scores); FatHMM > 0.5; GWAVA > 0.4 (at least 2 of 3 scores); DeepSEA > 0.5 (at least 2 of 3 scores); FunSeq2 > 3, ReMM > 0.5.
VI CpG Ratio: The ratio of observed to expected CpG in island.

Table 3. 3'UTR germline variants prioritized in the CRC family. Chromosomal position, pedigree segregation and allele frequencies in the Non-Finnish European population (*NFE*) are listed for each variant. Identified miRNA target sites according to Bedtools intersect function and Targetscan 7.0 databases are included. Overall deleteriousness, genomic conservation, chromatin state and the likelihood of target mRNA down-regulation predicted by mirSVR score are annotated by CADD v1.4. The general functional impact of each 3'UTR variant is summarized by the percentage of positive non-coding scores derived from SNPnexus. . – *not annotated*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

Gene name	Variant	Pedigree segregation	Allele frequency in NFE		Bedtools intersect[I]			CADD v1.4								SNPnexus Non-coding scores[IV] (%)
									Conservational scores			Chromatin state				
			ExAC	gnomAD	miRNA Target Sites	Context Score ++ Percentile	Site type	PHRED	verPhCons	verPhyloP	GerpN	ChromHMM[II] state	ChromHMM[II] score	Segway[III]	mirSVR-Score	
ACTN2	1_236927298_T_A	III7, III8, III10	.	6.48 E-05	miR-450b-5p miR-891b	36 71	7mer-m8 7mer-1a	13.84	0.84	0.60	5.70	Quies	0.488	R2	-0.90	66.7
CCNT2	2_135714351_C_T	III7, III8, III10	.	.	miR-4670-3p miR-3606-5p	93 83	7mer-1a 8mer	15.46	1.00	3.28	5.65	Tx Wk	0.551	GE0	-1.11	83.3
CLK1	2_201717953_C_T	III7, III8	.	2.01 E-03	miR-4718 let-7c-3p	98 96	7mer-m8 7mer-m8	15.57	0.68	0.38	5.34	Tx	0.921	GE0	-1.21	85.7
DDX17	22_38881603_ACT_A	II7, III7, III8	.	2.84 E-03	miR-550b-3p	91	7mer-m8	14.91	0.66	0.97	4.36	Tx	0.858	F1	-0.35	100
FH	1_241661076_T_C	III7, III8, III10	.	3.89 E-03	miR-4294	94	7mer-m8	10.4	0.47	0.11	4.56	Tx	0.709	GE0	-0.96	71.4
LRRC8C	1_90180607_T_A	III7, III8, III10	.	2.59 E-04	miR-499a-5p miR-208-3p miR-4432 miR-8087	96 91 81 90	8mer 7mer-1a 7mer-1a 7mer-1a	21.8	1.00	3.21	6.17	Tx Wk	0.583	F1	-0.76	83.3
SEMA4B	15_90772811_G_A	III7, III8, III10	.	.	miR-3918 miR-3127-5p miR-506-5p miR-10a-3p	99 99 92 94	8mer 8mer 7mer-m8 7mer-1a	13.65	0.53	3.01	4.95	Tx	0.606	GE1	-1.27	83.3
TJP1	15_29993152_G_A	III7, III8	.	6.49 E-05	miR-654-3p	89	8mer	15.03	1.00	1.48	5.79	Tx	0.441	GE1	-0.80	83.3

I Bedtools intersect: Using Bedtools intersect function and Targetscan 7.0 database, miRNA target sites were predicted for each variant. Multiple predicted miRNA target sites of one variant were listed separated with commas. The Context Score ++ Percentile shows the percentage of sites for the miRNA with a less favorable Context Score ++ and thus less repression capacity. Hence, a higher Context Score ++ Percentile indicates greater repression at a specific site by a miRNA compared to all sites of this miRNA family. The annotation of the site type gives information about the site efficacy in the seed region and thus about the different numbers of targets identified per miRNA. The order from the strictest to the least strict is as follows: 8mer > 7mer-m8 > 7mer-A1 > 6mer.

II ChromHMM: ChromHmm shows the proportion of 127 cell types of the Roadmap Epigenomics project in a particular chromatin state with scores closer to 1 indicating more cell types in the particular chromatin state. The 15 chromatin states are defined as follows: TssA – Active transcription start site (TSS), TssAFlnk – Flanking active TSS, TxFlnk – Transcribed at gene 5' and 3', Tx – Strong transcription, TxWk – Weak transcription, EnhG – Genic enhancers, Enh – Enhancers, ZNF/Rpts – ZNF genes & repeats, Het – Heterochromatin, TssBiv – Bivalent/Poised TSS/Enhancers, BivFlnk – Flanking bivalent TSS/Enhancer, EnhBiv – Bivalent enhancers, ReprPC – Repressed PolyComb, ReprPCWk – Weak Repressed PolyComb, Quies – Quiescent/low [26, 27].

III Segway: Segway uses a genomic segmentation method to annotate the chromatin state based on multiple datasets of ChIP-seq experiments. The chromatin states can be annotated as follows: D – dead, F0/1 – FAIRE, R0/1/2/4/5 – Repressed Region, H3K9me1 – histone 3 lysine 9 monomethylation, L0/1 – Low zone, GE0/1/2 – Gene body (end),TF0/1/2 – Transcription factor activity, C0 – CTCF, GS – Gene body (start), E/GM – Enhancer/gene middle, GM0/1 – Gene body (middle), TSS – Transcription start site, ZnfRpts – zinc finger repeats [28].

IV Non-coding scores: Following cut-offs were used for the interpretation of the non-coding scores derived from SNPnexus: FitCons Score ≥ 0.2 with a P-Value ≤ 0.05; EIGEN > 0 (at least 1 of 2 scores); FatHMM > 0.5; GWAVA > 0.4 (at least 2 of 3 scores); DeepSEA > 0.5 (at least 2 of 3 scores); FunSeq2 > 3, ReMM > 0.5.

Out of the remaining 16 3'UTR variants only 8 were annotated with a predicted miRNA target site by the Bedtools intersect function and with a mirSVR score ≤ -0.1, short-listed as the top 3'UTR variants in Table 3.

Checking the short-listed variants for their involvement in molecular mechanisms of colorectal carcinogenesis such as Wnt and Notch signaling pathways that are generally known to play a crucial role in CRC initiation [29], revealed that the *HDAC5* gene has been implicated in colorectal carcinogenesis by upregulating the Delta-like 4 ligand

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

(DLL4), a vascular specific Notch ligand essential for tumor angiogenesis [30, 31]. The potentially pathogenic role of HDAC5 in CRC was clinically confirmed by a further study showing the upregulation of HDAC5 protein in patients with early colon field carcinogenesis [32]. Therefore, the short-listed 5'UTR variant of *HDAC5* gene (17_42200942_T_G) was considered as a promising cancer-predisposing candidate. As the variant was annotated to be located at an active transcription start site according to ChromHmm (*cHmmTssA, Score = 0.984*) and Segway (*TSS*) (Table 2), an activating impact of the 5'UTR variant on *HDAC5* gene expression was hypothesized. The location of the variant in a CpG island and multiple TFBSs as well as the high PHRED-like CADD score of 21.9 supported the potential functional role of the identified *HDAC5* variant in cancer predisposition, leading to its final prioritization (Figure 2). Pedigree segregation of the prioritized variant was checked by IGV and further confirmed by targeted Sanger sequencing showing the wild type sequence for family members III3, III10 and IV3 and the heterozygous variant (T → G) for II7, III7 and III8, respectively (Supplementary Figure S1b).

2.3. Allele frequency in a large familial CRC cohort

Custom-made Taqman assays for screening of the *APCDD1* and *HDAC5* variants in 1705 familial CRC cases and 1674 healthy elderly individuals from Poland confirmed the variants in the family. Screening of *APCDD1* resulted in identifying the variant in additional 8 familial CRC cases and 2 healthy individuals (*odds ratio (OR) = 4.44, 95% confidence interval (CI) = [0.96; 20.56], p = 0.06*). Additionally, one individual, who originally was in the healthy control group but developed CRC at the age of 55 years, was heterozygous for the variant. That increased the OR to 4.93 (*95%CI = [1.08; 22.53], p = 0.04*). The existence of the heterozygous *APCDD1* variant was confirmed by Sanger sequencing in all positive samples. All CRC patients were diagnosed at ages between 30 and 64 years and had at least one family member diagnosed with CRC, in some cases also with other cancers such as breast, cervical, female genital tract, kidney and lung cancer and leukemia. The sampling ages of the two healthy individuals were 55 years and 80 years and they had no family history of any cancer. No other familial CRC cases showing the *HDAC5* variant were identified.

2.4. APCDD1 variant did not show a significant effect on proliferation of HEK293T and HT-29 cells

In order to investigate the functional impact of the identified *APCDD1* variant, CCK-8 proliferation assays were conducted for $pAPCDD1^{WT}$ and $pAPCDD1^{MUT}$ using HEK293T and HT-29 cell lines. We did not find any significant difference of viable cell numbers between $pAPCDD1^{WT}$ and $pAPCDD1^{MUT}$ transfected cells at any measured time point (*p = 0.05*, Supplementary Figure S2). These results indicate an improbable proliferative impact of the variant in HEK293T cells as well as in colon cancer cells HT-29, excluding the *APCDD1* variant as a sole potentially cancer predisposing candidate in the studied family.

2.5. 5'UTR variant of HDAC5 gene enhances promoter activity

To test our hypothesis that the identified 5'UTR variant contributes to increased *HDAC5* expression at transcriptional level, we performed luciferase reporter assays in HEK293T

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

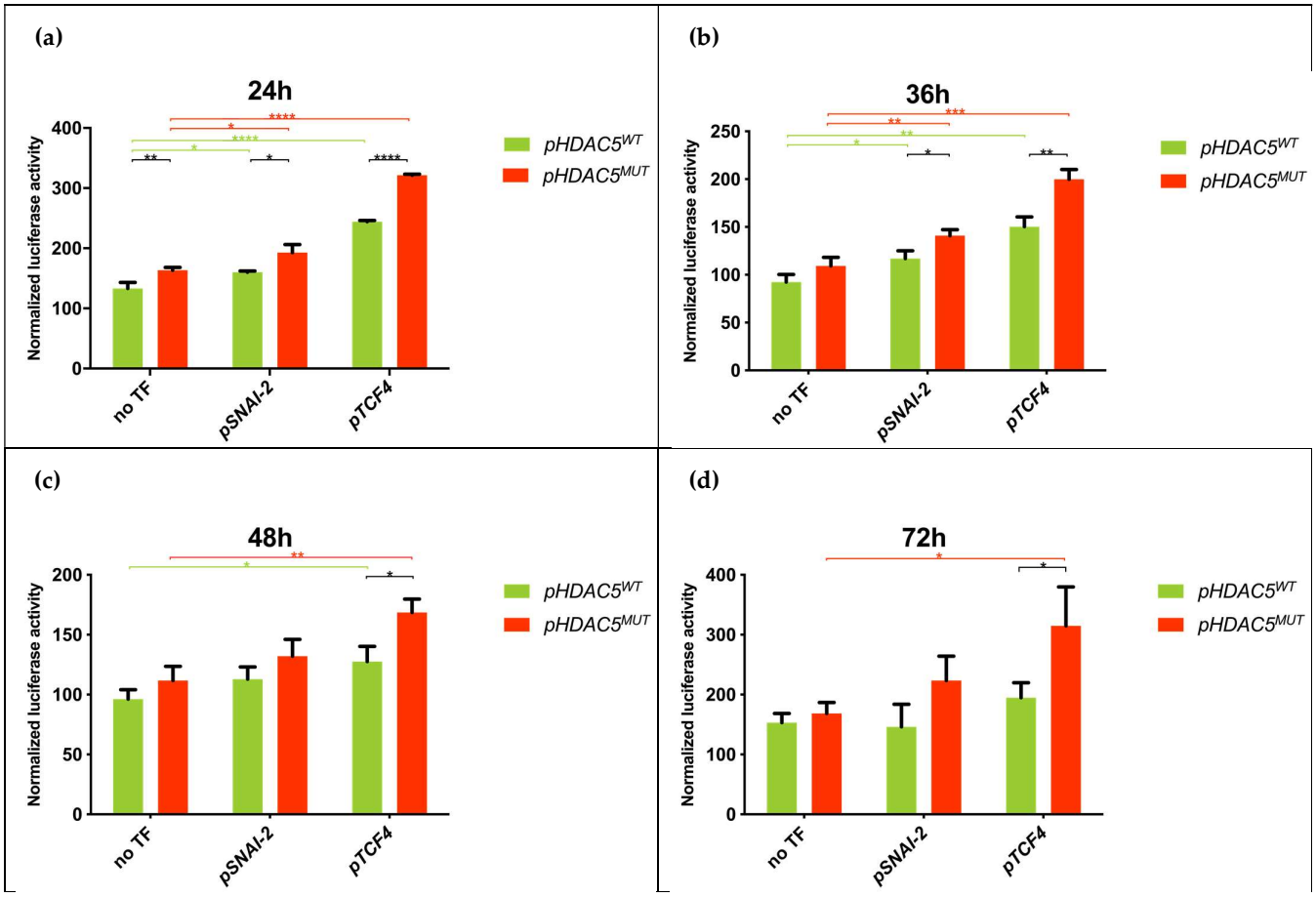

Figure 3. *pHDAC5^{MUT}* shows a significantly increased luciferase activity compared to *pHDAC5^{WT}*. Dual luciferase reporter assays performed for pGL4.10-*HDAC5^{WT}* and pGL4.10-*HDAC5^{MUT}* reporter constructs co-transfected with *pSNAI-2* or *pTCF4* or no transcription factor (TF) into HEK293T cells. Luciferase activity was measured at four different time points **(a-d)** and normalized to the empty pGL4.10 reporter vector. Each bar represents the mean of three independent experiments with standard deviation. *pHDAC5^{MUT}* shows a significantly increased luciferase activity compared to *pHDAC5^{WT}*. Co-transfection of *pTCF4* further enhanced the promoter activity of *pHDAC5^{MUT}* significantly. *- $p < 0.05$, **- $p < 0.01$, *** - $p < 0.001$, **** - $p < 0.0001$.

cells with both *pHDAC5*^{WT} and *pHDAC5*^{MUT}. The results of the reporter assays revealed a consistently higher luciferase activity ($R = L_F / L_R$) of cells transfected with *pHDAC5*^{MUT} compared to *pHDAC5*^{WT} after normalization to pGL4.10 vector ($R_E = 1$), respectively (Figure 3). Despite the overall increasing tendency, only the first time point after 24h post transfection resulted in a significant fold change in activity ($\Delta FA_{MUT/WT} = 122.64\%$, *t-test* $p < 0.01$), whereas the later time points showed no significant difference between *pHDAC5*^{WT} and *pHDAC5*^{MUT} using t-tests at a significance level of $\alpha = 0.05$. Since both plasmids, *pHDAC5*^{WT} and *pHDAC5*^{MUT}, are only differing in the variant of interest (*HDAC5*: 17_42200942_T_G), the detected increase of luminescent signal can be traced back to the variant itself causing the enhanced promoter activity.

2.6. 5'UTR variant of HDAC5 disrupts SNAI-2 and TCF4 transcription factor binding sites

Analysis of TFBSs predicted by Jaspar2020 with the default relative profile score threshold of 80%, resulted in the identification of 22 newly created TFBS (only present in *HDAC^{MUT}*) and 43 TFBS destroyed by the variant (only present in *HDAC5^{WT}*). The overall 65 identified TFBSs differing between *HDAC5^{WT}* and *HDAC^{MUT}* were found to be targeted by 51 different transcription factors. Further restriction of the relative score to 85%, which is referring to the likelihood sequence for the motif, narrowed down the number of identified transcription factors to 17, targeting 21 different TFBS (Table 4). Literature

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

mining showed an association of two specific transcription factors with colorectal carcinogenesis: SNAI-2 and TCF4 have been shown to be involved in Wnt as well as in Notch pathway and were thus considered as promising candidates for upregulation of *HDAC5* promoter activity in further luciferase reporter assays [33, 34]. According to Jaspar 2020, TCF4 was annotated to bind at 25 TFBS within the cloned *HDAC5* sequence of which 2 were disrupted in $HDAC^{MUT}$, whereas SNAI-2 was reported to bind at 13 TFBS with 1 binding site disrupted in $HDAC^{MUT}$.

2.7. Co-transfection of HDAC5 and TCF4 increases promoter activity due to 5'UTR variant of HDAC5 gene

To investigate the effect of potential regulatory transcription factors (SNAI-2 and TCF4) on the promoter activity of *HDAC5* gene, HEK293T cells that are not expressing these transcription factors endogenously were co-transfected with the respective expression vectors of *pSNAI-2* or *pTCF4* followed by luciferase reporter assays. The results showed an enhanced promoter activity for almost all measured time points after expression of the transcription factors compared to the cells only transfected with respective $pHDAC5^{WT}$/$pHDAC5^{MUT}$ vectors, respectively Figure 3.

Table 4. Summary of transcription factor binding sites (TFBS) identified with Jaspar2020 and relative profile score threshold of 85%. Matrix ID, names of targeting transcription factors, relative scores, start and end position, strand information as well as respective binding sequences are included. A relative score of 1 is representing the maximum likelihood sequence for the motif. TFBS newly created by the variant and thus exclusively present in $HDAC5^{MUT}$ sequence are annotated as *MUT*, whereas TFBS destroyed by the variant and thus exclusively present in $HDAC5^{WT}$ sequence are annotated as *WT*.

Matrix ID	Name	Relative score	Exclusive for *pHDAC5*	Start	End	Strand	Predicted sequence
MA1631.1	ASCL1(var.2)	0.859965928	WT	710	722	-	cagcacctcctcg
MA0598.1	EHF	0.869445857	WT	711	718	-	acctcctc
MA0056.1	MZF1	0.854364496	WT	710	715	+	cgagga
MA0673.1	NKX2-8	0.857521064	WT	708	716	-	ctcctcgac
MA1558.1	SNAI-1	0.875024573	WT	712	721	+	aggaggtgct
MA0745.1	**SNAI-2**	**0.891814443**	**WT**	**712**	**720**	**+**	**aggaggtgc**
MA1563.1	SOX18	0.858931866	MUT	714	721	-	agcaccGc
MA0079.3	SP1	0.871792316	MUT	710	720	-	gcaccGcctcg
MA0079.3	SP1	0.862428729	WT	710	720	-	gcacctcctcg
MA0080.2	SPI1	0.865925519	WT	712	718	+	aggaggt
MA1566.1	TBX3	0.876483102	WT	714	723	+	gaggtgctgc
MA0806.1	TBX4	0.854062458	WT	715	722	+	aggtgctg
MA1567.1	TBX6	0.867675674	WT	714	723	+	gaggtgctgc
MA1648.1	TCF12(var.2)	0.868262172	WT	711	721	-	agcacctcctc
MA0522.3	TCF3	0.875470132	WT	711	721	-	agcacctcctc
MA0522.2	TCF3	0.869530427	WT	712	721	-	agcacctcct
MA0830.2	**TCF4**	**0.851528821**	**WT**	**710**	**722**	**-**	**cagcacctcctcg**
MA0830.1	**TCF4**	**0.850775152**	**WT**	**712**	**721**	**-**	**agcacctcct**
MA0003.1	TFAP2A	0.860276583	MUT	707	715	-	Gcctcgacg
MA0815.1	TFAP2C(var.3)	0.858234123	MUT	704	716	+	agccgtcgaggCg
MA0815.1	TFAP2C(var.3)	0.851532375	MUT	704	716	-	cGcctcgacggct

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

Co-transfection of *pSNAI-2* led to a significant increase in luciferase activity of both *pHDAC5^{WT}* and *pHDAC5^{MUT}* cells compared to respective cells not expressing *pSNAI-2*, showing similar fold changes at 24 h and 36 h (*t-test $p < 0.05$*). To see if the described effect is partly mediated by the identified variant, we next compared the luciferase activity between *pHDAC5^{WT}* and *pHDAC5^{MUT}* cells co-transfected with *pSNAI-2*. This resulted in a significant fold change in promoter activity after 24h and 36h of transfection (*$\Delta FA_{MUT,S/WT,S} = 120.44\%$ (24h), 120.55% (36h), t-test $p < 0.05$*). Nevertheless, the fold change between *pHDAC5^{WT}* and *pHDAC5^{MUT}* cells not expressing *pSNAI-2* was observed to be similar (*$\Delta FA_{MUT/WT} = 122.64\%$ (24h), t-test $p < 0.01$*) which was also mirrored by the overlapping 95% confidence intervals of differences between *pHDAC5^{WT}* and *pHDAC5^{MUT}* cells respectively with or without *pSNAI-2* expression (*95%CI $_{MUT,S-WT,S} = [10.90; 54.57]$; 95%CI$_{MUT-WT} = [12.37; 47.96]$*). In summary, SNAI-2 transcription factor increased *HDAC5* promoter activity independent of the variant of interest after 24h and 36h of transfection.
Co-transfection of *pTCF4* expression vector also led to enhanced luciferase activity in both *pHDAC5^{WT}* and *pHDAC5^{MUT}* cells compared to respective cells not expressing *pTCF4*, whereas a higher effect on the cells carrying the mutated sequence was observed. This stronger enhancement of *pHDAC5^{MUT}* promoter activity by *pTCF4* expression is well-reflected in the comparison between *pHDAC5^{WT}* and *pHDAC5^{MUT}* cells: after co-transfection of *pTCF4*, *pHDAC5^{MUT}* cells showed a significantly higher promoter activity compared to *pHDAC5^{WT}* cells at all four measured time points, reaching its maximum at 72h (*$\Delta FA_{MUT,T/WT,T}$* =131.66%, 132.99%, 132.05%, 161.52%, respectively, *t-test $p < 0.05$*). In contrast to that, fold changes in promoter activity of *pHDAC5^{MUT}* compared to *pHDAC5^{WT}* without any co-transfection never exceeded the maximum value of 122.64% (24h post transfection) and were thus lower than in *pTCF4* co-transfected cells. Consistently, the difference between means of *pHDAC5^{WT}* and *pHDAC5^{MUT}* promoter activity was detected to be at least 2.5-fold higher in cells co-transfected with *pTCF4* compared to cells not expressing *pTCF4*; here again the maximum of differences between means was reached after 72h, showing an almost 8-fold value (*$\Delta R_{MUT,T-WT,T} = 119.9$; $\Delta R_{MUT-WT} = 15.29$*). Furthermore, comparison of the 95% confidence intervals of differences between *pHDAC5^{WT}* and *pHDAC5^{MUT}* revealed no overlapping for cells with or without *pTCF4* expression 24h post transfection (*95%CI $_{MUT,T-WT,T} = [73.44; 81.18]$; 95%CI$_{MUT-WT} = [12.37; 47.96]$*). Summing up the described results, the enhancing effect of *pTCF4* on *HDAC5* promoter activity can be partly traced back to a mechanism depending on the prioritized 5'UTR variant.
The dependency of TCF4 mediated promoter activity enhancement on the identified *HDAC5* variant was confirmed in a second experiment of luciferase reporter assays, focusing only on the comparison of *pTCF4* co-transfected cells at the same 4 time points: Fold changes in promoter activity of *pHDAC5^{MUT}* compared to *pHDAC5^{WT}* even reached values between 368.82% and 405.05% at all measured time points. Two-tailed t-tests resulted in extreme significance of promoter activity increase between *pHDAC5^{WT}* and *pHDAC5^{MUT}* cells (*$p < 0.0001$*).

3. Discussion

By applying our in-house developed FCVPPv2 on a CRC-affected family, we were able to prioritize two novel heterozygous germline variants, a coding variant in *APCDD1* and a non-coding variant in *HDAC5* gene. The *APCDD1* variant was identified in 8 additional familial CRC cases, 1 CRC case without family history and in 2 healthy elderly individuals without cancer family history, leading to a 4.9-fold increased CRC risk for the variant carriers (*$p = 0.04$*), while no other *HDAC5* variants were identified among the 1705 familial CRC cases. Cell proliferation assays indicated an insignificant proliferative impact for the *APCDD1* variant, luciferase reporter assay results showed an increased promoter activity by the 5'UTR variant of the *HDAC5* gene.
APCDD1 has been first identified as a direct Wnt target of the β-catenin/TCF4 transcription complex by Takahashi et al. who have further reported increased APCDD1 expres-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

sion in primary colon cancer tissue compared with corresponding healthy tissue [25]. In contrast, a study by De Sousa et al. showed that increased expression was restricted only to adenoma stages and not observed in carcinoma stages and that Wnt target genes such as *APCDD1* were epigenetically silenced by promoter methylation in different colon cancer cell lines. Since re-expression of APCDD1 was further associated with decreased Wnt signaling levels, the authors explained these observations in CRC by the known negative feedback regulation of Wnt signaling driven by several target genes including *APCDD1* [35, 36]. Showing decreased levels of β-catenin and Wnt target genes upon APCDD1 expression, Ordóñez-Morán et al. confirmed its role as a Wnt inhibitor and further proposed APCDD1 as a potential tumor suppressor in CRC [37]. Accordingly, both mentioned studies found a correlation of high APCDD1 expression with favorable prognosis for CRC patients [35, 37]. Though a cell growth promoting function was reported by Takahashi et al. for APCDD1 *in vitro* and *in vivo* using colon cancer LoVo cells [25], cell proliferation assays of this study using HEK293T and colon cancer cells HT-29 could not confirm the postulated proliferative impact of wild type APCDD1 nor of the identified *APCDD1* variant. Based on the potential tumor suppressor role of APCDD1 and the variant being relatively common among Polish CRC cases, the identified APCDD1 variant could be considered to play a possible role in colorectal carcinogenesis, though our cell proliferation experiments did not show any difference between the wild type and mutated HT-29 cells.

On the other hand, our study showed that implementation of the identified 5'UTR variant of *HDAC5* gene increased the promoter activity. Experimental confirmation of pedigree segregation as well as the established functional role of HDAC5 in colorectal carcinogenesis supported a role for the 5'UTR variant in CRC predisposition: HDAC5 plays a crucial part in epigenetic modulation of gene expression. By removing acetyl groups from the N-acetyllysine residues of histones, HDACs are able to enhance chromatin condensation, leading to transcriptional repression of genes [38]. Thus, dysregulation of HDACs induces chromatin rearrangement possibly affecting tumor suppressor genes or oncogenes which may explain the well-established association of HDACs with carcinogenesis of various malignancies such as CRC, medulloblastoma [39] or hepatocellular carcinoma [40]. In particular, upregulation of Notch ligand DLL4 and the possibly resulting activation of Notch pathway and angiogenesis could represent an important carcinogenic mechanism induced by HDAC5 in CRC. Furthermore, HDAC inhibitors such as azaindolylsulfonamides have already been investigated in CRC xenografts and have shown promising results in tumor growth suppression [41].

We showed an enhancing impact of the identified non-coding variant on *HDAC5* promoter activity with the help of luciferase reporter assays. A further upregulation of promoter activity by TCF4 expression, especially in cells carrying the mutation, could implicate TCF4 as a potential regulator of *HDAC5* expression partly depending on the inserted variant. Since implementation of the *HDAC5* 5'UTR variant leads to loss of TCF4 binding sites, the described enhancement could be traced back to the reduction of a potentially repressive function of TCF4 on *HDAC5* promoter activity at these specific sites. The enhancing effect of TCF4 as well on promoter activity of cells not carrying the mutation might be explained by further, still unexplored regulatory mechanisms as additional molecular interactions of TCF4 transcription factor with *HDAC5* promoter sequence. This may be supported by the identification of 23 additional TFBS targeted by TCF4 within the cloned *HDAC5* sequence. Anyway, the results of this work indicate the involvement of TCF4 transcription factor in regulation of *HDAC5* gene expression, resulting in HDAC5 upregulation and potentially promoting colorectal carcinogenesis in this family.

The proposed regulating impact of TCF4 on *HDAC5* gene expression in the studied family with CRC aggregation is further supported by the generally established role of TCF4 in colorectal carcinogenesis: As part of the Wnt signaling pathway, TCF4 was shown to form the β-catenin/TCF4 transcription complex in the nucleus and induce gene expression of Wnt targets such as *MYC* [33, 42-44], a known oncogene overexpressed in CRC

 doi:10.20944/preprints202102.0040.v1

[45-47]. As reported by Hatzis et al, genes upregulated by TCF4 are involved in cell proliferation, transcription, cell adhesion, negative regulation of programmed cell death, establishment and maintenance of chromatin [48]. Furthermore, TCF4 has been reported as a negative prognostic factor in CRC associated with shorter overall survival [49]. Although the molecular mechanisms leading to further *HDAC5* upregulation after the assumed TCF4 binding site loss are not yet fully understood, the described carcinogenic role of TCF4 in CRC is supported by our results and supports reciprocally the postulated CRC promoting function of *HDAC5* gene. A possible explanation approach for the underlying molecular mechanisms may consider the known dual regulatory role of TCF4 depending on the interaction with either transcriptional co-repressor (such as Groucho/transducin-like enhancer of split (Gro/TLE) family members and HDACs) or co-activator complexes (such as β-catenin and SMADs) [50-56].

The results of our functional experiments provided further evidence for the application of the FCVPPv2 to families with cancer aggregation and confirmed our pipeline's significance in the prioritization of both the coding and non-coding variants. The integration of a variety of annotation tools by the FCVPPv2 enables the identification of functionally important coding regions and regulatory elements in the non-coding sequence of genes. Regarding the prediction of TFBSs modified by the prioritized variant, CADD v1.4 identified a relatively high number of overlapping ChIP-seq TFBSs whereas SNPnexus and the intersect function of bedtools did not reveal any. The differing results for predicted TFBSs could be traced back to the different databases each tool is based on and different cell lines used in the studies. Thus, the synergetic application of all tools within the FCVPPv2 with subsequent integration of their respective predictions could be considered as a good approach for an all-encompassing analysis of TFBSs.

4. Materials and Methods

4.1. Patient samples

A family with 8 confirmed CRC cases in 3 generations was identified at the Department of Genetics and Pathology, Pomeranian Medical University in Szczecin, Poland. Blood samples were collected from 3 CRC cases, two siblings who were diagnosed with CRC at the age of 52 and 35 years and their aunt who developed CRC at the age of 83, 1 individual with polyps diagnosed at the age of 56, 59 and 71 years and 2 unaffected family members (Figure 1). The study was approved by the Bioethics Committee of the Pomeranian Medical Academy in Szczecin (No: BN-001/174/05). All participating individuals signed an informed consent.

4.2. Whole exome sequencing and variant evaluation

Genomic DNA was isolated using a modified Lahiri and Schnabel method [57] and WES was performed for 3 CRC cases (II7, III7, III8), a family member with polyps (III10) and 2 unaffected family members (III3, IV3) using Illumina-based small read sequencing. Mapping to reference human genome (assembly version Hs37d5) was performed using BWA [58] and duplicates were removed using Picard (http://broadinstitute.github.io/picard/).

4.3. Variant calling, annotation and filtering

Single nucleotide variants (SNVs) were detected by using SAM tools [59] and indels by using Platypus [60]. Variants were annotated using ANNOVAR [61], 1000 Genomes project [62], dbSNP [63] and Exome Aggregation Consortium (ExAC) [64]. Variants with a quality score of greater than 20 and a coverage of greater than 5x, SNVs that passed the strand bias filter (a minimum one read support from both forward and reverse strand) and indels that passed all the Platypus internal filters were retained. With respect to the 1000 Genomes project Phase 3, non-TCGA ExAC data [64], NHLBI-ESP6500 and local data sets variants with minor allele frequency (MAF) less than 0.1% in the European

 doi:10.20944/preprints202102.0040.v1

population were selected. A pairwise comparison of shared rare variants among the family was performed to check for sample swaps and family relatedness.

4.4. Variant filtering according to FCVPPv2

The Familial Cancer Variant Prioritization Pipeline version 2 (FCVPPv2) was applied for evaluation of identified variants as described below and summarized in Figure 2 [14].

4.4.1. Familial segregation of the cancer predisposing variants

Considering the probability of carrying the cancer predisposing variant for each analyzed family member, they were classified as *cases* and *controls* according to the presence or absence of CRC. Generally, all affected family members should carry the variant of interest, with the following exception: Since the typical age of onset in hereditary CRC patients is considered to be lower than in the general population, such as 45 years in HNPCC families [66] compared to 63 years in sporadic CRC [67], family member II7 developed CRC at a relatively high age of 83 years, untypical for familial inheritance. Therefore, she could be considered as a *phenocopy* in this family expressing the phenotypic disease, but not the underlying genotype. Whereas II7 could possibly but not definitely carry the variant of interest, family members III7 and III8 are considered as certain *cases* and thus carriers of the variant due to their young age of onset typical for familial CRC. The unaffected family members III3 and IV3 should not show the cancer predisposing variant of interest and are thus defined as *controls* in this family. On the other hand, family member III10 was diagnosed with multiple colorectal polyps at the age of 56, 59 and 71 years. Since colorectal polyps could be a preliminary stage of CRC, especially when recurrent, III10 could be considered as a possible carrier of the variant. Based on these definitions, the identified variants were filtered according to their presence in the *cases*, absence in the *controls* and presence or absence in the old-age CRC case *II7* and the *polyp carrier*.

4.4.2. Analysis of coding variants

All variants were ranked using the Combined Annotation Dependent Depletion (CADD) tool v1.3 [68]; evolutionary conservation scores: Genomic Evolutionary Rate Profiling (GERP >2.0), PhastCons (>0.3) and PhyloP (≥3.0) [69-71]; intolerance scores based on allele frequency data from our in-house datasets, from ESP [65] and ExAC supplemented by the ExAC-derived Z-score [72]; and deleteriousness scoring tools accessed from dbNSFP v3.0 (database for nonsynonymous SNVs' functional predictions) [73]. The variants should reach a PHRED-like CADD-score of ≥ 10 and fulfill at least 2 out of 3 conservational scores, 3 out of 4 intolerance scores and at least 60% of all 12 deleteriousness scores to be taken into account for further analysis. The final exonic candidates were further screened by considering the allele frequency in the non-Finnish European population in the latest version of gnomAD database (https://gnomad.broadinstitute.org/), the potential impact of amino acid substitutions with the help of Snap2 (https://rostlab.org/services/snap2web/), the prediction of cancer drivers by Cancer Genome Interpreter (https://www.cancergenomeinterpreter.org/) and recent literature for reported gene-cancer relations and potentially cancer-related protein functions [74, 75].

4.4.3. Analysis of non-coding variants

Non-coding variants were analyzed with the updated version of CADD (CADD v1.4) that provides comprehensive information about the functional importance of non-coding regions by integrating a variety of scoring tools such as transcription factor binding sites (TFBS) located in the 5'UTR and 1kb flanking region upstream of transcription start sites, mirSVR for ranking putative microRNA target sites [76], chromHmm and Segway that provide information about the biological function and active regulatory regions based on large-scale functional genomics datasets such as ChIP-seq data [26, 28]. The variants were also analyzed by SNPnexus for identification of CpG islands and TFBS and for annota-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 February 2021 doi:10.20944/preprints202102.0040.v1

tion of the functional impact of all non-coding variants [19]. Variants of the 5'UTR and 1kb upstream region as well as 3'UTR and 1kb downstream region were scanned for potential regulatory elements by means of bedtools intersect function and respective databases: the FANTOM5 consortium and the super-enhancer archive (SEA) were used for identification of promoters, enhancers or super-enhancers [77] and Targetscan 7.0 for identification of microRNA target sites [78].
Literature was checked for any gene-cancer relations and potentially cancer-related protein functions of the top non-coding candidates.

4.5. Analysis of transcription factor binding sites

By uploading the wild type sequence (*HDAC5*^{WT}) and the sequence containing the variant in the 5'UTR of *HDAC5* gene (*HDAC5*^{MUT}) to Jaspar2020, potential TFBS were predicted and compared [79].

4.6. Variant validation with IGV

Sequencing data of all prioritized variants were checked for correctness using the Integrative Genomics Viewer (IGV), a visualization tool for interactive exploration of large, integrated genomic datasets [80]. By this means, the identified variants were validated and the confidence in variant calls was increased.

4.7. Confirmation of familial segregation by Sanger sequencing

Polymerase Chain Reaction (PCR) was performed to amplify the 5'UTR of *HDAC5* gene (ENST00000225983.6) from DNA of the family members by using HotStarTaq DNA Polymerase (Qiagen, #203205) and following manufacturer´s instructions. The primers were designed with Primer3 v.0.4.0 (http://bioinfo.ut.ee/primer3-0.4.0/): *HDAC5* forward 5'-GGGGGTCTGGGTCTATTTTT-3', reverse 5'-GAAGGGGCAAATCAGACAAC-3'. PCR was run at an annealing temperature of 62 °C with 5% dimethylsulfoxide (DMSO). The amplicons were validated by gel electrophoresis and purified with ExoSAP purification kit according to the manufacturer´s instructions. Sequencing reaction was performed with BigDye Terminator v3.1 Ready Reaction Cycle Sequencing kit (Thermo Fisher Scientific, #4337455). The electrophoretic profiles of *HDAC5* sequences were analyzed manually.

4.8. Screening of large case and control cohorts

In order to determine the allele frequency of the *HDAC5* and *APCDD1* variants, 1705 familial CRC cases and 1674 healthy elderly individuals without cancer family history were checked using custom-made Taqman assays. The existence of heterozygous variants was confirmed by Sanger sequencing.

4.9.PCR based cloning of gene reporter constructs

Cloning primers were designed using Primer3 v.0.4.0 (http://bioinfo.ut.ee/primer3-0.4.0/) for the 5'UTR including the variant of interest as well as the first exon and part of the following intron of *HDAC5* gene (ENST00000225983.6). Adding specific restriction sites of Kpn I or Hind III and a 5' leader sequence of 6 bp resulted in the following primer pair: forward 5'-TAAGCAGGTAC^C gcaccaaagtcagggaagtc-3'; reverse 5'-TGCTTA^AAGCTTgaaggggcaaatcagacaac-3'. PCR was performed with an annealing temperature of 59 °C with 5% DMSO to amplify the required *HDAC5* insert with a total length of 1116 bp from human DNA. Digestion of the PCR amplicon and the promoter-less pGL4.10[*luc2*] vector purchased from Promega (#E6651) was performed using FastDigest Kpn I (Thermo Fisher Scientific, FD0524), FastDigest Hind III (Thermo Fisher Scientific, #FD0504) and FastDigest Buffer 10x (Thermo Fisher Scientific, #B64) according to the manufacturer's instructions. The digested products were validated by gel electrophoresis and extracted with Monarch® DNA Gel Extraction Kit (New England BioLabs, #T1020S). Ligation of the digested

 doi:10.20944/preprints202102.0040.v1

HDAC5 gene insert (1099bp) and pGL4.10 vector (4210bp) was done by using Quick Ligation™ Kit (New England BioLabs, #M2200S). The ligated product pGL4.10-*HDAC5* (5309bp), illustrated in Supplementary Figure S3, was again validated by gel electrophoresis and extracted with Monarch® DNA Gel Extraction Kit (New England BioLabs, #T1020S) for further use as the wild type pGL4.10-*HDAC5* (*pHDAC5*^{WT}) construct. The mutant pGL4.10-*HDAC5* (*pHDAC5*^{MUT}) construct was created by site-directed mutagenesis using QuikChange II XL Site-Directed Mutagenesis Kit (Agilent, #200521) according to the manufacturer's instructions and the following primer pair: forward 5'- gcggcagcaccgcctcgacggct -3', reverse 5'- agccgtcgaggcggtgctgccgc -3', designed based on Agilent QuikChange Primer Design (https://www.agilent.com/store/primerDesignProgram.jsp). Both plasmids, *pHDAC5*^{WT} and *pHDAC5*^{MUT}, were confirmed by Sanger sequencing.

4.10. Cloning of SNAI-2, TCF4 and APCDD1

Human pENTR223-*SNAI-2* (#172707094), pENTR223-*TCF4* (#107260711) and pENTR223-*APCDD1* clones (GPCF, #115154469) purchased from the Genomics and Proteomics Core Facility of the DKFZ Heidelberg (GPCF) were cloned into pDEST26 vector using Gateway™ LR Clonase™ II Enzyme mix (Thermo Fisher Scientific, #11791020). The identified variant was introduced into pDEST26-*APCDD1* vector (*pAPCDD1*) using QuikChange II XL SDM Kit and the following SDM primers designed with Agilent QuikChange Primer Design: forward 5'-gggtgagccagcactgtgaggtgcg-3', reverse 5'-cgcacctcacagtgctggctcaccc-3'. All sequences were confirmed by Sanger sequencing.

4.11. Plasmid amplification and extraction

Stellar chemically competent cells (Takara, #636763) were used for transformation of *pHDAC5*^{WT}, *pSNAI-2*, *pTCF4* and *pAPCDD1*^{WT}, whereas *pHDAC5*^{MUT} and *pAPCDD1*^{MUT} were transformed into XL10-Gold Ultracompetent cells (Agilent, #200314) after site-directed mutagenesis. Plasmid extraction was performed using PureLink™ HiPure Plasmid Midiprep Kit (Thermo Fisher Scientific, #K210004).

4.12. Cell line and culture conditions

Human embryonic kidney 293 (HEK293T) cells and human colon cancer cells HT-29 were a kind gift from Peter Krammer´s lab (DKFZ) and cultured in RPMI. Using Harmonizome, a database of processed datasets about genes and proteins, endogenous expression levels of proteins for HDAC5, SNAI-2 and TCF4 were checked and ruled out in HEK293T cells [81].

4.13. Cell proliferation assay – APCDD1

HEK293T and HT-29 cells were seeded in 24-well plates and 24h later transfected with either 150 ng of *pAPCDD1*^{WT}, *pAPCDD1*^{MUT} or pDEST26 vector as negative control. Merck's Cell Counting Kit – 8 (CCK-8, #96992) was used for quantitation of viable cell numbers in proliferation and cytotoxicity assays at four different time points: 0h, 24h, 48h and 72h post transfection. Briefly, 100 µl cell suspension of each well was treated with 10 µl CCK-8 solution and incubated for 1 hr at 37 °C. The absorbance was measured at 450 nm using a microplate reader and the number of viable cells was calculated based on a standard curve. Comparing the numbers of viable cells at different time points and the respective growth curves between *pAPCDD1*^{WT} and *pAPCDD1*^{MUT}, the proliferative impact of the implemented *APCDD1* variant (p.R299H) could be estimated for each cell line.

4.14. Luciferase reporter assay – HDAC5

HEK293T cells were seeded in 48-well plates and 24h later transfected with 100 ng of *pHDAC5*^{WT} or *pHDAC5*^{MUT} as a test reporter, 10 ng renilla as the control reporter and 25 µl Lipofectamine 2000 (Thermo Fisher Scientific, #11668030). Negative controls were considered by including cells transfected with promoter-less pGL4.10 vector (EMPTY, E). For

 doi:10.20944/preprints202102.0040.v1

investigating the impact of SNAI-2 and TCF4 transcription factors, cells were co-transfected with 20 ng of *pSNAI-2* or *pTCF4* expression vector and the corresponding negative controls were included: pGL4.10 vector in combination with each expression vector (*pSNAI-2* or *pTCF4*) and *pHDAC5*^{MUT} with the empty expression vector pDEST26. Luciferase assays were conducted using the dual-luciferase reporter assay system (Promega, #E1910) at four different time points: 24h, 36h, 48h and 72h post transfection. Since renilla luminescence is measured for vector normalization, the relative ratio R of firefly luminescence L_F to renilla luminescence L_R ($R = L_F / L_R$) was calculated and later referred to as *luciferase activity*. After normalizing R values to the empty promoter-less pGL4.10 vector ($R_E = 1$), the ratios were compared between *pHDAC5*^{WT} and *pHDAC5*^{MUT} for each condition. For this purpose, we calculated fold changes in promoter activity (Δ *Fold Activity* $_{MUT/WT} = \Delta FA_{MUT/WT} = R_{MUT} / R_{WT}$) as well as two-tailed t-tests at a significance level of $\alpha = 0.05$. All experiments were conducted in triplicates and repeated at least thrice.

5. Conclusions

Application of the FCVPPv2 on a CRC-affected family identified a novel missense variant in the *APCDD1* gene and a 5'UTR variant in the *HDAC5* gene as potentially cancer predisposing. While the *APCDD1* variant was relatively common among Polish CRC cases ($AF = 0.003$) and increased the risk of CRC 4.9-fold for the variant carriers, it did not seem to affect cell proliferation *in vitro*. On the other hand, the *HDAC5* variant shows a low allele frequency in any world population as well as an enhancing effect on *HDAC5* promoter activity in luciferase reporter assays and thereby on *HDAC5* gene expression. Our findings support the importance of taking into account both coding and non-coding variants in cancer predisposition, population screening and functional validation of variants.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Representative electropherograms depicting the *APCDD1* and *HDAC5* variants identified in the studied CRC family, Figure S2: Cell proliferation assays conducted for *pAPCDD1*^{WT} and *pAPCDD1*^{MUT}, Figure S3: Graphical overview of pGL4.10-*HDAC5* reporter constructs, Table S1: Summary of family members analyzed in our study.

Author Contributions: Conceptualization, Kari Hemminki, Asta Försti and Obul Reddy Bandapalli; Data curation, Diamanto Skopelitou, Abhishek Kumar, Nagarajan Paramasivam, Matthias Schlesner and Obul Reddy Bandapalli; Formal analysis, Diamanto Skopelitou, Abhishek Kumar, Nagarajan Paramasivam and Obul Reddy Bandapalli; Funding acquisition, Kari Hemminki and Asta Försti; Investigation, Diamanto Skopelitou, Beiping Miao, Aayushi Srivastava, Magdalena Kuświk, Asta Försti and Obul Reddy Bandapalli; Methodology, Diamanto Skopelitou, Beiping Miao, Aayushi Srivastava, Dagmara Dymerska, Jan Lubiński and Obul Reddy Bandapalli; Project administration, Kari Hemminki, Asta Försti and Obul Reddy Bandapalli; Resources, Dagmara Dymerska and Jan Lubiński; Software, Abhishek Kumar, Nagarajan Paramasivam, Matthias Schlesner and Obul Reddy Bandapalli; Supervision, Kari Hemminki, Asta Försti and Obul Reddy Bandapalli; Validation, Diamanto Skopelitou, Beiping Miao, Aayushi Srivastava, Magdalena Kuświk, Dagmara Dymerska, Jan Lubiński and Obul Reddy Bandapalli; Visualization, Diamanto Skopelitou, Beiping Miao and Obul Reddy Bandapalli; Writing – original draft, Diamanto Skopelitou and Obul Reddy Bandapalli; Writing – review & editing, Diamanto Skopelitou, Aayushi Srivastava, Kari Hemminki, Asta Försti and Obul Reddy Bandapalli.

Funding: K.H. was supported from the EU Horizon 2020 program, grant No. 856620.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Bioethics Committee of the Pomeranian Medical Academy in Szczecin (No: BN-001/174/05).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to conduct this research and publish the results.

Data Availability Statement: Unfortunately, for reasons of ethics and patient confidentiality, we are not able to provide the sequencing data into a public database. The data underlying the results presented in the study are available from the corresponding author or from Dr. Asta Försti (Email: a.foersti@kitz-heidelberg.de)

Acknowledgments: This article is based upon work from COST Action CA17118, supported by COST (European Cooperation in Science and Technology) and Transcan ERA-NET funding from the German Federal Ministry of Education and Research (BMBF). We thank the DKFZ Genomics and Proteomics Core Facility for Illumina Sequencing Services and Omics IT and Data Management Core Facility (ODCF), DKFZ for managing the NGS data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland**. *N Engl J Med* 2000, **343**(2):78-85.
2. Jasperson KW, Tuohy TM, Neklason DW, Burt RW: **Hereditary and familial colon cancer**. *Gastroenterology* 2010, **138**(6):2044-2058.
3. FC DAS, Wernhoff P, Dominguez-Barrera C, Dominguez-Valentin M: **Update on Hereditary Colorectal Cancer**. *Anticancer Res* 2016, **36**(9):4399-4405.
4. Wei C, Peng B, Han Y, Chen WV, Rother J, Tomlinson GE, Boland CR, Chaussabel D, Frazier ML, Amos CI: **Mutations of HNRNPA0 and WIF1 predispose members of a large family to multiple cancers**. *Fam Cancer* 2015, **14**(2):297-306.
5. Kuiper RP, Hoogerbrugge N: **NTHL1 defines novel cancer syndrome**. *Oncotarget* 2015, **6**(33):34069-34070.
6. Weren RD, Ligtenberg MJ, Kets CM, de Voer RM, Verwiel ET, Spruijt L, van Zelst-Stams WA, Jongmans MC, Gilissen C, Hehir-Kwa JY *et al*: **A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer**. *Nat Genet* 2015, **47**(6):668-671.
7. Briggs S, Tomlinson I: **Germline and somatic polymerase epsilon and delta mutations define a new class of hypermutated colorectal and endometrial cancers**. *J Pathol* 2013, **230**(2):148-153.
8. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Guarino E, Salguero I *et al*: **Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas**. *Nat Genet* 2013, **45**(2):136-144.
9. Lorans M, Dow E, Macrae FA, Winship IM, Buchanan DD: **Update on Hereditary Colorectal Cancer: Improving the Clinical Utility of Multigene Panel Testing**. *Clin Colorectal Cancer* 2018, **17**(2):e293-e305.
10. Gloss BS, Dinger ME: **Realizing the significance of noncoding functionality in clinical genomics**. *Exp Mol Med* 2018, **50**(8):97.
11. Mucaki EJ, Caminsky NG, Perri AM, Lu R, Laederach A, Halvorsen M, Knoll JH, Rogan PK: **A unified analytic framework for prioritization of non-coding variants of uncertain significance in heritable breast and ovarian cancer**. *BMC Med Genomics* 2016, **9**:19.
12. Alanazi IO, Al Shehri ZS, Ebrahimie E, Giahi H, Mohammadi-Dehcheshmeh M: **Non-coding and coding genomic variants distinguish prostate cancer, castration-resistant prostate cancer, familial prostate cancer, and metastatic castration-resistant prostate cancer from each other**. *Mol Carcinog* 2019, **58**(6):862-874.
13. Shabalina SA, Spiridonov NA: **The mammalian transcriptome and the function of non-coding DNA sequences**. *Genome Biol* 2004, **5**(4):105.

14. Kumar A, Bandapalli OR, Paramasivam N, Giangiobbe S, Diquigiovanni C, Bonora E, Eils R, Schlesner M, Hemminki K, Forsti A: **Familial Cancer Variant Prioritization Pipeline version 2 (FCVPPv2) applied to a papillary thyroid cancer family**. *Sci Rep* 2018, **8**(1):11635.
15. Bandapalli OR, Paramasivam N, Giangiobbe S, Kumar A, Benisch W, Engert A, Witzens-Harig M, Schlesner M, Hemminki K, Forsti A: **Whole genome sequencing reveals DICER1 as a candidate predisposing gene in familial Hodgkin lymphoma**. *Int J Cancer* 2018, **143**(8):2076-2078.
16. Srivastava A, Kumar A, Giangiobbe S, Bonora E, Hemminki K, Forsti A, Bandapalli OR: **Whole Genome Sequencing of Familial Non-Medullary Thyroid Cancer Identifies Germline Alterations in MAPK/ERK and PI3K/AKT Signaling Pathways**. *Biomolecules* 2019, **9**(10).
17. Srivastava A, Giangiobbe S, Kumar A, Paramasivam N, Dymerska D, Behnisch W, Witzens-Harig M, Lubinski J, Hemminki K, Försti A *et al*: **Identification of Familial Hodgkin Lymphoma Predisposing Genes Using Whole Genome Sequencing**. 2020, **8**(179).
18. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M: **CADD: predicting the deleteriousness of variants throughout the human genome**. *Nucleic Acids Res* 2019, **47**(D1):D886-d894.
19. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C: **SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine**. *Nucleic Acids Res* 2018, **46**(W1):W109-W113.
20. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y *et al*: **The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses**. *Curr Protoc Bioinformatics* 2016, **54**:1 30 31-31 30 33.
21. Wang Z, Cummins JM, Shen D, Cahill DP, Jallepalli PV, Wang TL, Parsons DW, Traverso G, Awad M, Silliman N *et al*: **Three classes of genes mutated in colorectal cancers with chromosomal instability**. *Cancer Res* 2004, **64**(9):2998-3001.
22. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI *et al*: **InterPro in 2019: improving coverage, classification and access to protein sequence annotations**. *Nucleic Acids Res* 2019, **47**(D1):D351-D360.
23. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A *et al*: **The Pfam protein families database in 2019**. *Nucleic Acids Res* 2019, **47**(D1):D427-D432.
24. Letunic I, Bork P: **20 years of the SMART protein domain annotation resource**. *Nucleic Acids Res* 2018, **46**(D1):D493-D496.
25. Takahashi M, Fujita M, Furukawa Y, Hamamoto R, Shimokawa T, Miwa N, Ogawa M, Nakamura Y: **Isolation of a novel human gene, APCDD1, as a direct target of the beta-Catenin/T-cell factor 4 complex with probable involvement in colorectal carcinogenesis**. *Cancer Res* 2002, **62**(20):5651-5656.
26. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization**. *Nat Methods* 2012, **9**(3):215-216.
27. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J *et al*: **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015, **518**(7539):317-330.
28. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS: **Unsupervised pattern discovery in human chromatin structure through genomic segmentation**. *Nat Methods* 2012, **9**(5):473-476.
29. Fre S, Pallavi SK, Huyghe M, Lae M, Janssen KP, Robine S, Artavanis-Tsakonas S, Louvard D: **Notch and Wnt signals cooperatively control cell proliferation and tumorigenesis in the intestine**. *Proc Natl Acad Sci U S A* 2009, **106**(15):6309-6314.
30. Badenes M, Trindade A, Pissarra H, Lopes-da-Costa L, Duarte A: **Delta-like 4/Notch signaling promotes Apc (Min/+) tumor initiation through angiogenic and non-angiogenic related mechanisms**. *BMC Cancer* 2017, **17**(1):50.
31. He P, Liang J, Shao T, Guo Y, Hou Y, Li Y: **HDAC5 promotes colorectal cancer cell proliferation by up-regulating DLL4 expression**. *Int J Clin Exp Med* 2015, **8**(4):6510-6516.

32. Stypula-Cyrus Y, Damania D, Kunte DP, Cruz MD, Subramanian H, Roy HK, Backman V: **HDAC up-regulation in early colon field carcinogenesis is involved in cell tumorigenicity through regulation of chromatin structure**. *PLoS One* 2013, **8**(5):e64600.
33. Shah M, Rennoll SA, Raup-Konsavage WM, Yochum GS: **A dynamic exchange of TCF3 and TCF4 transcription factors controls MYC expression in colorectal cancer cells**. *Cell Cycle* 2015, **14**(3):323-332.
34. Findlay VJ, Wang C, Nogueira LM, Hurst K, Quirk D, Ethier SP, Staveley O'Carroll KF, Watson DK, Camp ER: **SNAI2 modulates colorectal cancer 5-fluorouracil sensitivity through miR145 repression**. *Mol Cancer Ther* 2014, **13**(11):2713-2726.
35. de Sousa EMF, Colak S, Buikhuisen J, Koster J, Cameron K, de Jong JH, Tuynman JB, Prasetyanti PR, Fessler E, van den Bergh SP *et al*: **Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients**. *Cell Stem Cell* 2011, **9**(5):476-485.
36. Shimomura Y, Agalliu D, Vonica A, Luria V, Wajid M, Baumer A, Belli S, Petukhova L, Schinzel A, Brivanlou AH *et al*: **APCDD1 is a novel Wnt inhibitor mutated in hereditary hypotrichosis simplex**. *Nature* 2010, **464**(7291):1043-1047.
37. Ordonez-Moran P, Dafflon C, Imajo M, Nishida E, Huelsken J: **HOXA5 Counteracts Stem Cell Traits by Inhibiting Wnt Signaling in Colorectal Cancer**. *Cancer Cell* 2015, **28**(6):815-829.
38. Marek L, Hamacher A, Hansen FK, Kuna K, Gohlke H, Kassack MU, Kurz T: **Histone deacetylase (HDAC) inhibitors with a novel connecting unit linker region reveal a selectivity profile for HDAC4 and HDAC5 with improved activity against chemoresistant cancer cells**. *J Med Chem* 2013, **56**(2):427-436.
39. Milde T, Oehme I, Korshunov A, Kopp-Schneider A, Remke M, Northcott P, Deubzer HE, Lodrini M, Taylor MD, von Deimling A *et al*: **HDAC5 and HDAC9 in medulloblastoma: novel markers for risk stratification and role in tumor cell growth**. *Clin Cancer Res* 2010, **16**(12):3240-3252.
40. Fan J, Lou B, Chen W, Zhang J, Lin S, Lv FF, Chen Y: **Down-regulation of HDAC5 inhibits growth of human hepatocellular carcinoma by induction of apoptosis and cell cycle arrest**. *Tumour Biol* 2014, **35**(11):11523-11532.
41. Lee HY, Tsai AC, Chen MC, Shen PJ, Cheng YC, Kuo CC, Pan SL, Liu YM, Liu JF, Yeh TK *et al*: **Azaindolylsulfonamides, with a more selective inhibitory effect on histone deacetylase 6 activity, exhibit antitumor activity in colorectal cancer HCT116 cells**. *J Med Chem* 2014, **57**(10):4009-4022.
42. Hrckulak D, Janeckova L, Lanikova L, Kriz V, Horazna M, Babosova O, Vojtechova M, Galuskova K, Sloncova E, Korinek V: **Wnt Effector TCF4 Is Dispensable for Wnt Signaling in Human Cancer Cells**. *Genes (Basel)* 2018, **9**(9).
43. Korinek V, Barker N, Morin PJ, van Wichen D, de Weger R, Kinzler KW, Vogelstein B, Clevers H: **Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma**. *Science* 1997, **275**(5307):1784-1787.
44. Bienz M, Clevers H: **Linking colorectal cancer to Wnt signaling**. *Cell* 2000, **103**(2):311-320.
45. Smith DR, Myint T, Goh HS: **Over-expression of the c-myc proto-oncogene in colorectal carcinoma**. *Br J Cancer* 1993, **68**(2):407-413.
46. Erisman MD, Rothberg PG, Diehl RE, Morse CC, Spandorfer JM, Astrin SM: **Deregulation of c-myc gene expression in human colon carcinoma is not accompanied by amplification or rearrangement of the gene**. *Mol Cell Biol* 1985, **5**(8):1969-1976.
47. Rochlitz CF, Herrmann R, de Kant E: **Overexpression and amplification of c-myc during progression of human colorectal cancer**. *Oncology* 1996, **53**(6):448-454.
48. Hatzis P, van der Flier LG, van Driel MA, Guryev V, Nielsen F, Denissov S, Nijman IJ, Koster J, Santo EE, Welboren W *et al*: **Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells**. *Mol Cell Biol* 2008, **28**(8):2732-2744.
49. Kriegl L, Horst D, Reiche JA, Engel J, Kirchner T, Jung A: **LEF-1 and TCF4 expression correlate inversely with survival in colorectal cancer**. *J Transl Med* 2010, **8**:123.

50. Mao CD, Byers SW: **Cell-context dependent TCF/LEF expression and function: alternative tales of repression, de-repression and activation potentials**. *Crit Rev Eukaryot Gene Expr* 2011, **21**(3):207-236.
51. Hoverter NP, Waterman ML: **A Wnt-fall for gene regulation: repression**. *Sci Signal* 2008, **1**(39):pe43.
52. Arce L, Yokoyama NN, Waterman ML: **Diversity of LEF/TCF action in development and disease**. *Oncogene* 2006, **25**(57):7492-7504.
53. Cadigan KM: **TCFs and Wnt/beta-catenin signaling: more than one way to throw the switch**. *Curr Top Dev Biol* 2012, **98**:1-34.
54. Cadigan KM, Waterman ML: **TCF/LEFs and Wnt signaling in the nucleus**. *Cold Spring Harb Perspect Biol* 2012, **4**(11).
55. Brantjes H, Roose J, van De Wetering M, Clevers H: **All Tcf HMG box transcription factors interact with Groucho-related co-repressors**. *Nucleic Acids Res* 2001, **29**(7):1410-1419.
56. Wang H, Matise MP: **Tcf7l2/Tcf4 Transcriptional Repressor Function Requires HDAC Activity in the Developing Vertebrate CNS**. *PLoS One* 2016, **11**(9):e0163267.
57. Lahiri DK, Schnabel B: **DNA isolation by a rapid method from human blood samples: effects of MgCl2, EDTA, storage time, and temperature on DNA yield and quality**. *Biochem Genet* 1993, **31**(7-8):321-328.
58. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
59. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data**. *Bioinformatics* 2011, **27**(21):2987-2993.
60. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, Consortium WGS, Wilkie AO, McVean G, Lunter G: **Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications**. *Nat Genet* 2014, **46**(8):912-918.
61. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic Acids Res* 2010, **38**(16):e164.
62. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA *et al*: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
63. Smigielski EM, Sirotkin K, Ward M, Sherry ST: **dbSNP: a database of single nucleotide polymorphisms**. *Nucleic Acids Res* 2000, **28**(1):352-355.
64. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB *et al*: **Analysis of protein-coding genetic variation in 60,706 humans**. *Nature* 2016, **536**(7616):285-291.
65. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB: **Genic intolerance to functional variation and the interpretation of personal genomes**. *PLoS Genet* 2013, **9**(8):e1003709.
66. Steinke V, Engel C, Buttner R, Schackert HK, Schmiegel WH, Propping P: **Hereditary nonpolyposis colorectal cancer (HNPCC)/Lynch syndrome**. *Dtsch Arztebl Int* 2013, **110**(3):32-38.
67. Brandt A, Bermejo JL, Sundquist J, Hemminki K: **Age of onset in familial cancer**. *Ann Oncol* 2008, **19**(12):2084-2088.
68. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, **46**(3):310-315.
69. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A: **Distribution and intensity of constraint in mammalian genomic sequence**. *Genome Res* 2005, **15**(7):901-913.
70. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**(8):1034-1050.
71. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome Res* 2010, **20**(1):110-121.

72. Ward LD, Kellis M: **HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants**. *Nucleic Acids Res* 2012, **40**(Database issue):D930-934.
73. Liu X, Wu C, Li C, Boerwinkle E: **dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs**. *Hum Mutat* 2016, **37**(3):235-241.
74. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, Tusquets I, Albanell J, Rodon J, Tabernero J *et al*: **Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations**. *Genome Med* 2018, **10**(1):25.
75. Hecht M, Bromberg Y, Rost B: **Better prediction of functional effects for sequence variants**. *BMC Genomics* 2015, **16 Suppl 8**:S1.
76. Betel D, Koppal A, Agius P, Sander C, Leslie C: **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. *Genome Biol* 2010, **11**(8):R90.
77. Wei Y, Zhang S, Shang S, Zhang B, Li S, Wang X, Wang F, Su J, Wu Q, Liu H *et al*: **SEA: a super-enhancer archive**. *Nucleic Acids Res* 2016, **44**(D1):D172-179.
78. Agarwal V, Bell GW, Nam JW, Bartel DP: **Predicting effective microRNA target sites in mammalian mRNAs**. *Elife* 2015, **4**.
79. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranasic D *et al*: **JASPAR 2020: update of the open-access database of transcription factor binding profiles**. *Nucleic Acids Res* 2020, **48**(D1):D87-D92.
80. Robinson JT, Thorvaldsdottir H, Wenger AM, Zehir A, Mesirov JP: **Variant Review with the Integrative Genomics Viewer**. *Cancer Res* 2017, **77**(21):e31-e34.
81. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A: **The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins**. *Database (Oxford)* 2016, **2016**.