Article

Whole exome sequencing identifies novel germline variants of SLC15A4 gene as potentially cancer predisposing in familial colorectal cancer

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

[1] Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany; mando.skopelitou@yahoo.de (D.S.); srivastava.aayhushi97@gmail.com (A.S.); b.miao@kitz-heidelberg.de (B.M.); abhishek@ibioinformatics.org (A.K.); k.hemminki@dkfz.de (K.H.); a.foersti@kitz-heidelberg.de (A.F.)
[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, 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 dymerska@pum.edu.pl (D.D.); lubinski@pum.edu.pl (J.L.)
[8] Computational Oncology, Molecular Diagnostics Program, National Center for Tumor Diseases (NCT), Germany; n.paramasivam@dkfz.de
[9] Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), Heidelberg, Germany; m.schlesner@dkfz.de
[10] 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: About 15% of colorectal cancer (CRC) patients have first-degree relatives affected by the same malignancy. However, for most families the cause of familial aggregation of CRC is unknown. In order to identify novel high-to-moderate penetrant germline variants underlying CRC susceptibility, we performed whole exome sequencing (WES) on four CRC cases and two unaffected family members of a Polish family without any mutation in known CRC predisposition genes. After WES, we used our in-house developed Familial Cancer Variant Prioritization Pipeline and identified two novel variants in the solute carrier family 15 member 4 (SLC15A4) gene. The heterozygous missense variant, p. Y444C, was predicted to affect the phylogenetically conserved PTR2/POT domain and to have a deleterious effect on the function of the encoded peptide/histidine transporter. The other variant was located in the upstream region of the same gene (GRCh37.p13, 12_129308531_C_T; 43bp upstream of transcription start site, ENST00000266771.5) and it was annotated to affect the promoter region of SLC15A4 as well as binding sites of 17 different transcription factors. Our findings of two distinct variants in the same gene may indicate a synergistic up-regulation of SLC15A4 as the underlying genetic cause and implicate this gene for the first time in genetic inheritance of familial CRC.

Keywords: SLC15A4; germline variant; familial colorectal cancer; whole exome sequencing

1. Introduction

Several studies have estimated that around 15% of colorectal cancer (CRC) patients show a first-degree family history of colorectal malignancies [1-4]. Analyzing the underlying heritable and environmental factors in twins from Sweden, Denmark, and Finland, Lichtenstein et al. have estimated that genetic factors account for up to 35% of the CRC risk

[5]. Nevertheless, only a small proportion of CRC cases can be traced back to germline mutations in established CRC-predisposing genes. These include the early-identified traditional susceptibility genes APC and mismatch repair genes (MLH1, MSH2, MSH6, PMS2), MUTYH and SMAD4/BMPR1A. Later on, sequencing studies have identified novel predisposition genes for CRC, such as NTHL1, RNF43, POLE, POLD1, FAN1 and RPS20 [6-16]. Further candidate genes recently suggested by modern next generation sequencing methods include the solute carrier (SLC) family of membrane transport genes: SLC5A9 (p.G492Afs*13), SLC26A8 (p.R954C) and SLC11A1 (p.P64A) [17,18]. Additionally, germline deletions affecting the open reading frame of SLC18A1 gene have been reported to increase the risk of CRC and lower SLC18A1 protein expression has been further associated with poor clinical outcome [19].

Despite novel findings of predisposition gene candidates in CRC, there still exist 75% of unexplored familial CRC cases. This proportion of familial CRC with unknown genetic background may be accounted for by two major components: either following a monogenic inheritance model based on a single high-penetrance mutation or a polygenic inheritance model based on the combination of multiple low/moderate-penetrance risk alleles [20]. Assuming the monogenic disease model for CRC cases with strong familial clustering, the identification of rare highly penetrant germline variants within pedigree-based studies constitutes a promising approach for elucidating the remaining genetic burden of familial CRC.

For this purpose, we performed whole exome sequencing (WES) on a Polish family with CRC aggregation over three generations. Sequencing data of four CRC cases and two unaffected family members was subsequently analyzed using our in-house developed Familial Cancer Variant Prioritization Pipeline (FCVPPv2) which was used earlier in identification of variants and pathways involved in several familial cancers [21-27]. Further in silico analyses resulted in the prioritization of a novel missense variant in the solute carrier family 15 member 4 gene (SLC15A4), encoding a proton-dependent peptide/histidine transporter. By being involved in multiple signaling pathways regulating cytokine production and thus innate immune responses, SLC15A4 has been shown to promote colitis in an in vivo mouse model [28,29]. Since high expression of the encoded membrane transporter has been further reported in the feces of CRC patients as well as in early-stage CRC cell lines, an important role of SLC15A4 in initial inflammation-induced colorectal carcinogenesis has been suggested [30]. In this study, we conducted in silico analyses as well as further literature search in order to link the function of the SLC15A4 protein to a genetic basis, potentially contributing to CRC development in the studied family. By identifying and analyzing an additional variant in the upstream region of the SLC15A4 gene showing the same familial segregation, we aimed to expand the theory of high-penetrance monogenic inheritance to a synergistic model of coding and non-coding variants underlying cancer predisposition.

2. Materials and Methods

Patient samples & ethical permissions

The family with CRC history over three generations was recruited from Poland (Figure 1A). Six family members were included in our experiments: four siblings diagnosed with CRC, a child of one CRC case with colorectal polyps and a healthy cousin of the CRC cases. The family was screened for alterations in APC, the mismatch repair genes MLH1, MSH2, MSH3, large deletions in EPCAM and POLE p. Leu424Val, POLD1 p.Ser478Asn and NTHL1 p.Gln90* mutations and found to be negative. Collection of blood samples and clinical information from subjects was undertaken with informed written consent in agreement with the tenets of the declaration of Helsinki. The study was approved by the

Bioethics Committee of the Pomeranian Medical Academy in Szczecin (protocol code No: BN-001/174/05).

Whole exome sequencing, variant calling and annotation

Genomic DNA was isolated with a modified Lahiri and Schnabel method [31] and WES was performed using Illumina-based small read sequencing. After mapping to the human reference genome (assembly GRCh37 version Hs37d5) by means of BWA [32], duplicates were removed with Picard (http://broadinstitute.github.io/picard/). SAM tools [33] and Platypus [34] were used for calling single nucleotide variants (SNVs) as well as short insertions and deletions (indels), respectively. Variants were then annotated by ANNOVAR [35], 1000 Genomes Project [36], dbSNP [37] and Exome Aggregation Consortium (ExAC) [38]. In order to be further processed, variants should have a quality score of ≥20 and a coverage score of ≥5x, SNVs should pass the strand bias filter (a minimum one read support from both forward and reverse strand) and indels should pass all the Platypus internal filters. Based on minor allele frequencies (MAFs) deduced from the 1000 Genomes Project Phase 3, non-TCGA ExAC data , NHLBI-ESP6500 and local data sets, rare variants with a MAF ≤0.1% in the European population were retained for further analysis. We checked for potential sample swaps and family relatedness by pairwise comparison of the shared rare variants.

Coding variant analysis according to the FCVPPv2

The resulting variants were analyzed based on our in-house developed FCVPPv2 [21]. First, variants were filtered according to the pedigree segregation of the malignancy. Variants should be present in family members affected by CRC and absent in the healthy family member. Since colorectal polyps at a relatively young age may represent a preliminary stage of familial CRC, the respective family member could be a possible carrier and show either presence or absence of the variant of interest.

Of the coding variants fulfilling the pedigree segregation criteria, the most deleterious 10% were retained for further analysis, represented by a PHRED-like CADD score ≥10 [40]. In order to evaluate the evolutionary conservation as an indicator for functional importance of a genomic position, the following scoring tools were applied with respective cutoff values given in brackets: Genomic Evolutionary Rate Profiling (GERP; ≥2.0), PhastCons (>0.3) and PhyloP score (≥3.0) [41-43]. Next, the intolerance of genes against functional genetic variation was assessed by using three intolerance scores (<0) based on allele frequency data from our in-house datasets, from NHLBI-ESP6500 and ExAC [39]. In the course of intolerance screening, missense and loss-of-function variants were further annotated by the Z-Score (>0) and pLI score (≥0.9), respectively, which were specifically developed by the ExAC consortium for the particular type of variants [38]. Last, we evaluated the deleteriousness of non-synonymous and splice site SNVs by applying 10 different scoring tools accessed from dbNSFP v3.0 (database for nonsynonymous SNPs' functional predictions): Sorting Tolerant From Intolerant (SIFT), Polymorphism Phenotyping v2 (PolyPhen-2) HumDiv, PolyPhen-2 HumVar, Log ratio test (LRT), MutationTaster, MutationAssessor, Functional Analysis Through Hidden Markov Models (FATHMM), Reliability Index, Variant Effect Scoring Tool version 3 (VEST3) and Protein Variation Effect Analyzer (PROVEAN) [44]. Summarizing, variants with a PHRED-like CADD-score of ≥10 as well as ≥2 out of the 3 conservational tools, ≥60% of the 4 intolerance scores and ≥60% of the 10 deleteriousness scores fulfilling the selection criteria were retained as the top exonic candidates. Allele frequencies in the non-Finnish European (NFE) population were re-evaluated by means of gnomAD browser (https://gnomad.broadinstitute.org/) [45]. We further assessed the potential of the variants for being cancer drivers in CRC by checking overall somatic alteration frequencies according to cBioPortal and TCGA

PanCancer Atlas, comprising data of 594 CRC patients [46,47]. Moreover, protein expression levels in CRC tissue were accessed from The Human protein atlas (http://www.proteinatlas.org) [48].

Additional in silico analyses based on protein function and phylogenetic conservation

The potential impact of the top missense variants on protein function was assessed by means of Snap2 [49]. Based on a neural network, Snap2 calculates the likelihood of single amino acid substitutions to alter protein function, giving scores between -100 (low) and +100 (high). The predicted functional impact is represented in form of heat maps covering each possible amino acid substitution at each position.

Non-coding variant analysis according to the FCVPPv2

Since predictions of the functional impact of variants are based on evolutionary information, we further checked phylogenetic conservation of the top variants among different vertebrate species. Multiple protein sequences of the candidate genes and their orthologs were derived from the National Center for Biotechnology Information (NCBI) [50] and aligned using COBALT, a constraint-based multiple alignment tool [51]. Visualized alignments were manually checked for conservation at the mutation sites and the surrounding regions and percent identity of protein sequences was further calculated by NCBI BLAST (Basic Local Alignment Search Tool). Details of multiple sequence alignment including selected representative species and NCBI accessions of respective genes and their orthologs are summarized in Table S1.

We checked recent literature for established gene-cancer associations, postulated oncogene or tumor-suppressor roles as well as potential cancer-promoting protein functions of the top candidates. Considering the entirety of derived information and in silico analysis results, the candidates showing the most promising impact on protein function or gene regulation were prioritized as the potentially cancer-causing variants in the studied family. Familial segregation of the top-listed variants with the disease was confirmed by visually checking WES data with the help of the Integrative Genomics Viewer (IGV) [52].

Analysis of regulatory elements and prediction of transcription factor binding sites in the non-coding regions

In order to assess the biological function and to identify potentially active regulatory regions, the chromatin state of specific genomic positions was predicted by the updated version of CADD (v1.6). For this purpose, CADD v1.6 provides chromHmm and Segway data, which annotate the chromatin state based on large-scale functional genomics datasets such as ChIP-seq data [53-55]. Using the intersect function of the Bedtools as well as FANTOM5 and SEA databases, we further scanned for potentially affected regulatory elements such as promoters, enhancers and super-enhancers. Moreover, transcription factor binding sites (TFBSs) were predicted by means of Jaspar2020 with the default relative profile score threshold of 80% and compared between wild type and mutant sequence [56].

3. Results

Application of the FCVPPv2 results in the prioritization of two coding variants in PTGES and SLC15A4 genes

The studied family was diagnosed with CRC over three generations, as represented in the pedigree (Figure 1A). Four siblings affected by CRC in the second generation at the age of <60 years were considered as cases and should therefore carry the variant of interest. Similarly, the daughter (IV2) of one of the cases (III2) developed colorectal polyps at the relatively young age of 29 years, potentially representing a preliminary stage of familial CRC. Considering the option of having inherited the variant of interest, IV2 was defined as a possible carrier and may present the variant as well. In contrast, an unaffected first cousin of the four CRC cases of a similar age and with healthy parents served as a control and should thus not carry the variant.

Analysis of WES data was performed using our in-house developed FCVPPv2, as visually summarized in Figure 1B. Of the totally identified 11,076 variants with a MAF ≤0.1%, only 135 variants fulfilled the pedigree segregation criteria. Exclusion of intergenic and intronic variants resulted in 28 variants in the coding region and 43 variants located in the non-coding region near transcription start and end sites (5′ and 3′ untranslated regions, up-stream and downstream regions). Due to their less pathogenic character, synonymous variants were excluded, leaving 17 missense or nonsense variants for further analysis. 12 of the remaining coding variants reached a PHRED-like CADD score ≥10, representing the most deleterious 10% of the variants in the human genome. Application of conservational, intolerance and deleteriousness scores further narrowed down the number of variants to 9, 4 and 2, respectively. The two final missense variants were located in solute carrier family 15 member 4 gene (SLC15A4, p.Y444C) and prostaglandin E synthase gene (PTGES, p.A133T) and are summarized with respective analysis results in Table 1. PTGES encodes a glutathione-dependent synthase catalyzing the oxidoreduction to prostaglandin E2. By playing a role in inflammatory responses, fever and pain, PTGES protein has been reported to be involved in inflammatory diseases such as collagen-induced arthritis and gastritis [57,58]. Similarly, the gene product encoded by SLC15A4 regulates innate immune responses. Being a proton-dependent peptide/histidine transporter, SLC15A4 protein controls the transport of various molecules from the inside of endosomes to the cytosol and has been associated inter alia with systemic lupus erythematosus [59-62].

According to the gnomAD browser, both top-listed variants showed very low allele frequencies in the general NFE population: the PTGES variant was annotated with a frequency of around 8.4x10-5 and the SLC15A4 variant even less with 0 counts in 113,688 alleles. Moreover, only one allele of the SLC15A4 variant has been reported in the world-wide population accessed by gnomAD browser, counting in total 251,362 alleles [45].

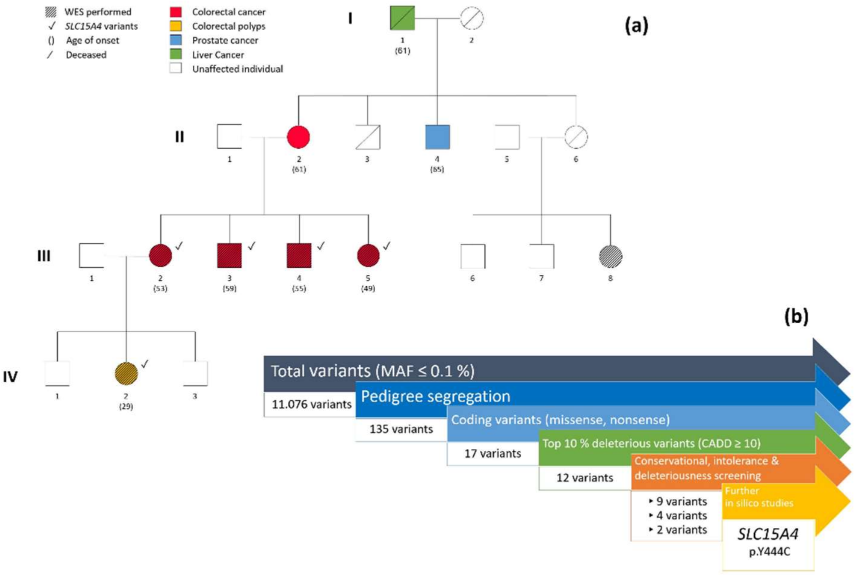

Figure 1. (a) Pedigree of the studied family with CRC aggregation over three generations and the presence of the missense and upstream variants in the *SLC15A4* gene. (**b)** Graphical overview of the filtering process according to the Familial Cancer Variant Prioritization Pipeline version 2 (FCVPPv2).

Higher alteration frequency and protein expression of SLC15A4 in CRC compared to PTGES

We next checked available CRC patient data for overall somatic gene alteration frequencies in order to assess the potential of the top candidates for being cancer drivers in CRC. cBioPortal recorded 6 somatic missense mutations in the SLC15A4 gene (frequency = 1.01%, Figure 2A) and only 2 somatic mutations in PTGES gene (frequency = 0.34%, Figure 3A) identified within 594 colorectal adenocarcinoma samples from the TCGA PanCancer Atlas. Regarding the overall somatic alteration frequency in all listed cancers, SLC15A4 showed a generally higher frequency with up to 5.48% in uterine cancer (Figure S1A), whereas the maximum alteration frequency of PTGES gene was only 1.7%, also in uterine cancer (Figure S1B) [46,47]. Besides genetic alterations documented in CRC, we checked protein expression levels in CRC samples. According to the Human Protein Atlas, 4 out of 12 investigated CRC samples showed a medium expression of SLC15A4 protein, whereas 0 out of 11 CRC samples showed a high or medium expression of PTGES protein [48].

Table 1. Overview of the top exonic variants prioritized in the studied CRC family. Chromosomal position, classification, pedigree segregation, allele frequency in the Non-Finnish European (*NFE*) population, PHRED-like CADD score, conservational score and the percentage of reached intolerance and deleteriousness scores are summarized for each variant. Snap^2 results for the predicted amino acid changes are included with calculated effect scores and accuracies given in %. Respective protein functions of the encoded gene products are derived from Genecards [63]. *non-syn SNV - non-synonymous single nucleotide variant*

Gene name	Chromosomal position	Exonic classification	Pedigree segregation	NFE allele frequency		CADD SCORE	Conservational scores			Intolerance scores (%)	Deleteriousness scores* (%)	Amino acid change	Snap^2		Protein function
				ExAC	gnomAD		GERP++	PhyloP	PhastCons				Effect score	Accuracy (%)	
PTGES	9_132501952_C_T	nonsyn SNV	III2, III3, III4, III5, IV2	2.10x10^{-4}	8.43x10^{-5}	34	4.67	7.723	1	75	80	A133T	16	59	glutathione-dependent prostaglandin E synthase, involved in inflammatory responses, fever, pain
SLC15A4	12_129285482_T_C	nonsyn SNV	III2, III3, III4, III5, IV2	0	0	23.7	5.49	5.609	1	100	90	Y444C	44	71	proton-dependent peptide/histidine transporter, regulation of innate immune responses

* 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).

In silico analyses predict functional consequences of the SLC15A4 variant on protein level

The identified SLC15A4 variant p.Y444C was predicted to affect the PTR2 (peptide transport) domain (p.104-495) of the POT (proton-dependent oligopeptide transporter) family (Figure 2A) and in particular a non-cytoplasmic loop of the SLC15A4 transporter protein, which comprises in total 12 transmembrane domains according to Interpro [64].

Analysis of the potential impact of the SLC15A4 missense variant on protein function by means of Snap2 resulted in a predicted effect score of 44 with an accuracy of 71% (Figure 2B). In contrast, the missense variant p.A133T in the PTGES gene, affecting the cytoplasmic part of the MAPEG (membrane-associated proteins in eicosanoid and glutathione metabolism) domain (p.17-146) of the PTGES protein (Figure 3A), was annotated by Snap2 with a score of 16 and an accuracy of 59% (Figure 3B) [49]. Due to a higher effect score and accuracy of the prediction in cross-validation, the functional impact of the SLC15A4 variant was expected to be of higher relevance.

Sequence alignment of the orthologs showed for both variants a universally conserved position with an overall high conservation of the surrounding region among the selected vertebrate species (Figure 2C and 3C, respectively). Focusing on the 5 directly adjacent upstream and downstream amino acid positions, multiple sequence alignment resulted in 95.86% identity of SLC15A4 and 82.64% identity of PTGES gene with their orthologs. Based on this observation, a higher phylogenetic conservation in the region surrounding the mutation site can be assumed for SLC15A4.

The entirety of the in silico analyses led to the prioritization of the missense variant in SLC15A4 gene (p.Y444C). Familial segregation of this variant was manually checked and confirmed by applying IGV on the WES data.

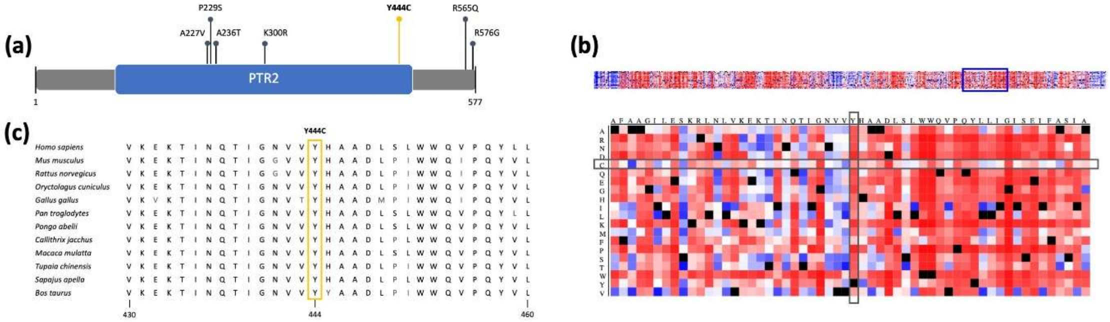
(a)
P229S
A227V
A236T
K300R
Y444C
R565Q
R576G
PTR2
1
577
(b)
A R N D C Q E G H I L K M F P S T W Y V
(c)
Y444C
Homo sapiens V K E K T I N Q T I G N V V Y H A A D L S L W W Q V P Q Y L L
Mus musculus V K E K T I N Q T I G G V V Y H A A D L P I W W Q I P Q Y V L
Rattus norvegicus V K E K T I N Q T I G G V V Y H A A D L P I W W Q I P Q Y V L
Oryctolagus cuniculus V K E K T I N Q T I G N V V Y H A A D L P I W W Q V P Q Y V L
Gallus gallus V K V K T I N Q T I G N V T Y H A A D M P I W W Q I P Q Y V L
Pan troglodytes V K E K T I N Q T I G N V V Y H A A D L S L W W Q V P Q Y L L
Pongo abelii V K E K T I N Q T I G N V V Y H A A D L S L W W Q V P Q Y V L
Callithrix jacchus V K E K T I N Q T I G N V V Y H A A D L P L W W Q V P Q Y V L
Macaca mulatta V K E K T I N Q T I G N V V Y H A A D L S L W W Q V P Q Y V L
Tupaia chinensis V K E K T I N Q T I G N V V Y H A A D L P I W W Q V P Q Y V L
Sapajus apella V K E K T I N Q T I G N V V Y H A A D L P L W W Q V P Q Y V L
Bos taurus V K E K T I N Q T I G N V V Y Y A A D L P I W W Q V P Q Y V L
430
444
460

Figure 2. *In silico* analysis results of the *SLC15A4* variant p.Y444C. **(a)** Graphical overview of the SLC15A4 protein with the PTR2 domain. Somatic mutations identified in CRC were extracted from cBioPortal (www.cbioportal.org) on 13th of December 2020 using the TCGA PanCancer data and are represented by dark pins. The germline missense variant identified in the studied CRC family is highlighted in the form of a yellow pin. **(b)** Snap2 heatmap depicting the functional impact of amino acid substitutions. The missense mutation p.Y444C is highlighted by grey boxes. **(c)** Extract of multiple sequence alignment of amino acids 430-460 of *SLC15A4* and orthologs. The mutation site is highlighted by a yellow box.

Identification of an additional variant at an active transcription start site of SLC15A4 gene

We checked the WES data of the studied family for further variants affecting the same gene of interest. Interestingly, one additional variant in the upstream region of the SLC15A4 gene showing the same familial segregation as the missense variant (present in the cases and the possible carrier) was identified (12_129308531_C_T; 43bp upstream of transcription start site, ENST00000266771.5). Functional annotation of the non-coding variant was derived from CADD v1.6 providing a PHRED-like CADD-score of 11.38 [40]. Moreover, the variant was annotated to be located at an active transcription start site according to ChromHmm (TssA, Score = 0.969) and Segway (TSS) [53,55]. CADD v1.6 further calculated 52 different overlapping ChIP TFBSs covered by the upstream variant and 115 TFBS peaks when summed over different cell types and tissue. Using the intersect function of the Bedtools, the non-coding variant was predicted to affect the promoter

(129308487..129308588) of the SLC15A4 gene. All described analysis results of the SLC15A4 upstream variant are summarized in Table 2.

In order to identify those transcription factors for which the binding may be affected the most by the variant, we used Jaspar2020 for prediction and comparison of the TFBSs for the wild type and the mutant sequence of the SLC15A4 upstream region [56]. Whereas most of the identified TFBS were shared by both sequences, 9 transcription factors were predicted to bind only to the wild type sequence, indicating a TFBS disruption by the variant, and 8 were predicted to bind only to the mutant sequence, indicating a TFBS creation by the variant (Table 3). One of the identified transcription factors, whose binding site was disrupted was STAT1 which has been established as a favorable prognostic marker in several types of cancers, including CRC [65-67]. Moreover, STAT1 has been proposed as a tumor suppressor particularly in colitis-associated CRC [68], in turn suggesting a carcinogenic potential of its disruption by the identified upstream variant.

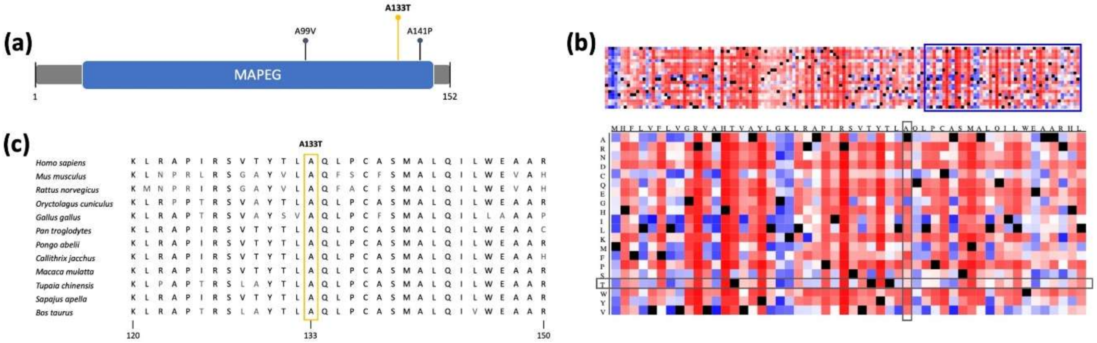
(a)
A133T
A99V
A141P
MAPEG
1
152
(b)
A R N D C Q E G H I L K M F P S T W Y V
(c)
A133T
Homo sapiens K L R A P I R S V T Y T L A Q L P C A S M A L Q I L W E A A R
Mus musculus K L N P R L R S G A Y V L A Q F S C F S M A L Q I L W E V A H
Rattus norvegicus K M N P R I R S G A Y V L A Q F A C F S M A L Q I L W E V A H
Oryctolagus cuniculus K L R P P T R S V A Y T L A Q L P C A S M A L Q I L W E A A R
Gallus gallus K L R A P T R S V A Y S V A Q L P C F S M A L Q I L L A A A P
Pan troglodytes K L R A P I R S V T Y T L A Q L P C A S M A L Q I L W E A A C
Pongo abelii K L R A P I R S V T Y T L A Q L P C A S M A L Q I L W E A A R
Callithrix jacchus K L R A P I R S V T Y T L A Q L P C A S M A L Q I L W E A A H
Macaca mulatta K L R A P I R S V T Y T L A Q L P C A S M A L Q I L W E A A R
Tupaia chinensis K L P A P T R S L A Y T L A Q L P C A S M A L Q I L W E A A R
Sapajus apella K L R A P I R S V T Y T L A Q L P C A S M A L Q I L W E A A R
Bos taurus K L R A P T R S L A Y T L A Q L P C A S M A L Q I V W E A A R
120
133
150

Figure 3. *In silico* analysis results of the *PTGES* variant p.A133T. **(a)** Graphical overview of the PTGES protein with MAPEG domain. Somatic mutations identified in CRC are extracted from cBioPortal (www.cbioportal.org) on 13th of December 2020 using the TCGA PanCancer data and are represented by dark pins. The germline missense variant identified in the studied CRC family is highlighted in the form a yellow pin. **(b)** Snap^2 heatmap depicting the functional impact of amino acid substitutions. The missense mutation p.A133T is highlighted by grey boxes. **(c)** Extract of multiple sequence alignment of amino acids 120-150 of *PTGES* and orthologs. The mutation site is highlighted by a yellow box.

Table 2 Analysis results of the *SLC15A4* upstream variant identified in the studied CRC family. Chromosomal position, variant annotation, pedigree segregation and allele frequency in the Non-Finnish European (*NFE*) population are listed. The PHRED-like CADD score, annotation of the chromatin state and location within transcription factor binding sites (*TFBS*) are derived from CADD v1.6. Affected promoter region according to Bedtools intersect function and SEA, FANTOM5 databases are included with respective start and end positions.

Gene name	Chromosomal position	Variant annotation	Pedigree segregation	NFE allele frequency		CADD v1.6						Bedtools intersect		
				ExAC	gnomAD	CADD SCORE	Chromatin state			TFBS	TFBSPeaks[III]	Promoter		
							ChromHMM[I] state	ChromHMM[I] score	Segway[II]			Start	End	Strand
SLC15A4	12_129308531_C_T	upstream	III2, III3, III4, III5, IV2	0	0.003754	11.38	TssA	0.969	TSS	52	115	129308487	129308588	-

I 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 [53,54].

II 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 [55].

IV TFBS peaks: The number of overlapping ChIP TFBS peaks summed over different cell types/tissue.

Table 3 Summary of transcription factors exclusively targeting either the wild type (*WT*) or the mutant sequence (*MUT*) of *SLC15A4* upstream region. Respective transcription factor binding sites (*TFBS*) are identified with Jaspar2020 and the default relative profile score threshold of 80%. Matrix ID, relative scores, start and end positions, strand information as well as respective binding sequences are included.

Transcription factor	Targeting	Matrix ID	Relative score*	Start	End	Strand	Predicted sequence
MEIS2	WT	MA0774.1	0.84	116	123	+	gggacAGG
NR1D2	WT	MA1532.1	0.81	108	122	+	tgggttctgggacAG
RARA::RXRG	WT	MA1149.1	0.80	109	126	+	gggttctgggacAGGTGA
RBPJ	WT	MA1116.1	0.86	113	122	+	tctgggacAG
RORC	WT	MA1151.1	0.82	110	121	+	ggttctgggacA
SREBF1	WT	MA0595.1	0.80	118	127	-	GTCACCTgtc
STAT1	WT	MA0137.2	0.84	109	123	-	CCTgtcccagaaccc
		MA0137.3	0.88	111	121	+	gttctgggacA
TGIF2LX	WT	MA1571.1	0.81	117	128	-	GGTCACCTgtcc
			0.81	117	128	+	ggacAGGTGACC
TGIF2LY	WT	MA1572.1	0.82	117	128	-	GGTCACCTgtcc
			0.82	117	128	+	ggacAGGTGACC
GRHL2	MUT	MA1105.2	0.83	116	127	+	ggaacAGGTGAC
MYF6	MUT	MA0667.1	0.82	118	127	+	aacAGGTGAC
NFATC2	MUT	MA0152.1	0.90	115	121	-	Tgttcca
PRDM4	MUT	MA1647.1	0.81	114	124	-	ACCTgttccag
SCRT1	MUT	MA0743.1	0.83	114	128	+	ctggaacAGGTGACC
		MA0743.2	0.85	113	128	+	tctggaacAGGTGACC
SCRT2	MUT	MA0744.1	0.85	114	126	+	ctggaacAGGTGA
		MA0744.2	0.85	113	128	+	tctggaacAGGTGACC
TEF	MUT	MA0843.1	0.80	110	121	-	Tgttccagaacc
ZBTB26	MUT	MA1579.1	0.92	107	121	-	Tgttccagaacccag

* A relative score of 1 is representing the maximum likelihood sequence for the motif.

4. Discussion

SLC15A4 belongs to the family of the proton-coupled oligopeptide transporters (POTs) that enable the transfer of histidine and oligopeptides derived from degradation products from inside of the endosome to the cytosol. Since proton dependency implies higher transport activity at low pH levels, endosomal acidification during the maturation to lysosomes is required for substrate uptake by the SLC15A4 transporter [69,70].

Well-established examples of SLC15A4 substrates are the NOD1 ligands L-Ala-D-Glu-meso-diaminopimelic acid (Tri-DAP) and γ-D-Glu-meso-diaminopimelic acid (iE-DAP), components of the cell wall peptidoglycan of primarily Gram-negative bacteria [28,71]. NOD1 stimulation by DAP induces the activation of nuclear factor-κB and mitogen-activated protein (MAP) kinases and thus the transcription of various genes responsible for innate and adaptive immune responses [72,73]. Knockdown of SLC15A4 in HEK293T cells has been shown to lead to decreased nuclear factor-κB activation by the NOD1 ligands [71], which was supported by in vivo experiments resulting in loss of Tri-DAP–induced cytokine production in SLC15A4-deficient mice. The same study has further reported an association of SLC15A4 with toll like receptor 9 (TLR9) functions: SLC15A4-deficient dendritic cells showed decreased TLR9-mediated cytokine production which was traced back by the authors to high lysosomal histidine concentrations in the absence of SLC15A4. By being required for TLR9- as well as NOD1-mediated cytokine production, SLC15A4 has been shown to promote Th1-dependent colitis in vivo [28].

Since chronic intestinal inflammation has been associated with increased CRC risk, potentially mediated by oxidative DNA damage and innate and adaptive immune responses [74,75], SLC15A4 may further play an important role in the initial inflammation-induced colorectal carcinogenesis (https://www.ebi.ac.uk/gwas/efotraits/EFO_0003767; accessed on March 5th, 2021). Based on these findings, we are suggesting a role in CRC susceptibility as well for genetic variation of SLC15A4.

Performing WES on a family with CRC aggregation and applying our in-house developed FCVPPv2, we were able to identify a novel heterozygous variant in the coding region of the SLC15A4 gene. By being present in all four CRC affected siblings as well as one direct descendant with colorectal polyps, the identified missense variant in SLC15A4 shows segregation with the disease and a potential for medium-to-high-penetrance susceptibility to CRC in the studied family. Considering the very low allele frequency of the variant in the NFE population of 0 counts in 113,688 alleles, the proposed association of the identified genetic variation with familial CRC is further supported. In silico analyses based on evolutionary conservation, intolerance against functional genetic alterations and deleteriousness led to the prediction of pathogenicity for the missense variant. Snap2 further predicted an effect on protein function by the missense variant leading to the amino acid substitution Y444C in SLC15A4. Considering all analyses, we propose an up-regulating mode of action for the identified missense variant on SLC15A4 protein level.

Interestingly, we identified another variant with the same familial segregation in the upstream region of the SLC15A4 gene (12_129308531_C_T; 43bp upstream of transcription start site, ENST00000266771.5). GnomAD browser reported an allele frequency of 3.754 x 10^{-3} in the NFE population. Taking this relatively high frequency into account, high penetrance and thus strong functional consequence of the upstream variant by itself may not be expected. Nevertheless, synergistic effects of both variants occurring in the same gene have to be considered: The upstream variant may have an enhancing impact on SLC15A4 protein expression, potentially of minor relevance when solely occurring but which may reinforce the postulated up-regulating mode of action of the SLC15A4 coding variant in the course of colorectal carcinogenesis. In order to confirm the proposed mode of function, we assessed the upstream variant for potentially influencing gene transcription. According to our analysis, the upstream variant was annotated to be located at an active transcription start site affecting the promoter region of the SLC15A4 gene. In particular, binding sites of 17 different transcription factors were predicted to be exclusive for either the wild type or the mutant sequence due to the identified upstream variant, representing a potential mechanism of enhancing gene transcription. Whether the variant potentially destroys TFBSs for transcriptional repressors or creates new TFBSs for transcriptional activators, remains unclear and requires further functional experiments. By providing a list of TFBSs and potential transcriptional repressors or activators, including the tumor suppressor STAT1, we aim to lay the foundation for functional validation of the regulatory impact of the upstream variant and instigate further research in this field. Thus, we hope to facilitate a better understanding of the identified upstream variant in the context of SLC15A4 gene regulation in particular and of the postulated synergistic model of coding and non-coding variants in cancer predisposition in general.

By identifying germline variants in the SLC15A4 gene in familial CRC, we implicated this gene for the first time in genetic inheritance of a malignancy, expanding its role from a potential CRC marker in quantitative fecal tests to a potential marker of CRC susceptibility in genetic testing. In doing so, we hope to facilitate genetic counseling not only of the particular family of this study but also of other individuals at risk of developing familial CRC. However, the results of this study need to be further replicated in validation cohorts and validated using experimental approaches in cell lines.

Funding: This research was funded by 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). KH was supported by 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 (protocol code 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 patients to publish this paper.

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: The authors thank the members of the families for participating in this study, Genomics and Proteomics Core Facility (GPCF) of the German Cancer Research Center (DKFZ) for providing excellent library preparation and sequencing services and the Omics IT and Data Management Core Facility (ODCF) of the DKFZ for the whole exome sequencing data management.

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. Ponz de Leon, M.; Sassatelli, R.; Sacchetti, C.; Zanghieri, G.; Scalmati, A.; Roncucci, L. Familial aggregation of tumors in the three-year experience of a population-based colorectal cancer registry. *Cancer Res* **1989**, *49*, 4344-4348.
2. Chau, R.; Jenkins, M.A.; Buchanan, D.D.; Ait Ouakrim, D.; Giles, G.G.; Casey, G.; Gallinger, S.; Haile, R.W.; Le Marchand, L.; Newcomb, P.A., *et al.* Determining the familial risk distribution of colorectal cancer: A data mining approach. *Fam Cancer* **2016**, *15*, 241-251.
3. Hemminki, K.; Sundquist, J.; Bermejo, J.L. How common is familial cancer? *Ann Oncol* **2008**, *19*, 163-167.
4. Frank, C.; Fallah, M.; Sundquist, J.; Hemminki, A.; Hemminki, K. Population landscape of familial cancer. *Sci Rep* **2015**, *5*, 12891.
5. Lichtenstein, P.; Holm, N.V.; Verkasalo, P.K.; 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*, 78-85.

6. Weren, R.D.; Ligtenberg, M.J.; Kets, C.M.; de Voer, R.M.; Verwiel, E.T.; Spruijt, L.; van Zelst-Stams, W.A.; Jongmans, M.C.; Gilissen, C.; Hehir-Kwa, J.Y., *et al.* A germline homozygous mutation in the base-excision repair gene nthl1 causes adenomatous polyposis and colorectal cancer. *Nat Genet* **2015**, *47*, 668-671.

7. Kuiper, R.P.; Hoogerbrugge, N. Nthl1 defines novel cancer syndrome. *Oncotarget* **2015**, *6*, 34069-34070.

8. Yan, H.H.N.; Lai, J.C.W.; Ho, S.L.; Leung, W.K.; Law, W.L.; Lee, J.F.Y.; Chan, A.K.W.; Tsui, W.Y.; Chan, A.S.Y.; Lee, B.C.H., *et al.* Rnf43 germline and somatic mutation in serrated neoplasia pathway and its association with braf mutation. *Gut* **2017**, *66*, 1645-1656.

9. Gala, M.K.; Mizukami, Y.; Le, L.P.; Moriichi, K.; Austin, T.; Yamamoto, M.; Lauwers, G.Y.; Bardeesy, N.; Chung, D.C. Germline mutations in oncogene-induced senescence pathways are associated with multiple sessile serrated adenomas. *Gastroenterology* **2014**, *146*, 520-529.

10. 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*, 148-153.

11. Palles, C.; Cazier, J.B.; Howarth, K.M.; Domingo, E.; Jones, A.M.; Broderick, P.; Kemp, Z.; Spain, S.L.; 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*, 136-144.

12. Valle, L.; de Voer, R.M.; Goldberg, Y.; Sjursen, W.; Forsti, A.; Ruiz-Ponte, C.; Caldes, T.; Garre, P.; Olsen, M.F.; Nordling, M., *et al.* Update on genetic predisposition to colorectal cancer and polyposis. *Molecular aspects of medicine* **2019**, *69*, 10-26.

13. Jasperson, K.W.; Tuohy, T.M.; Neklason, D.W.; Burt, R.W. Hereditary and familial colon cancer. *Gastroenterology* **2010**, *138*, 2044-2058.

14. Lorans, M.; Dow, E.; Macrae, F.A.; Winship, I.M.; Buchanan, D.D. Update on hereditary colorectal cancer: Improving the clinical utility of multigene panel testing. *Clin Colorectal Cancer* **2018**, *17*, e293-e305.

15. Nieminen, T.T.; O'Donohue, M.F.; Wu, Y.; Lohi, H.; Scherer, S.W.; Paterson, A.D.; Ellonen, P.; Abdel-Rahman, W.M.; Valo, S.; Mecklin, J.P., *et al.* Germline mutation of rps20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. *Gastroenterology* **2014**, *147*, 595-598 e595.

16. Segui, N.; Mina, L.B.; Lazaro, C.; Sanz-Pamplona, R.; Pons, T.; Navarro, M.; Bellido, F.; Lopez-Doriga, A.; Valdes-Mas, R.; Pineda, M., *et al.* Germline mutations in fan1 cause hereditary colorectal cancer by impairing DNA repair. *Gastroenterology* **2015**, *149*, 563-566.

17. Hansen, M.F.; Johansen, J.; Sylvander, A.E.; Bjornevoll, I.; Talseth-Palmer, B.A.; Lavik, L.A.S.; Xavier, A.; Engebretsen, L.F.; Scott, R.J.; Drablos, F., *et al.* Use of multigene-panel identifies pathogenic variants in several crc-predisposing genes in patients previously tested for lynch syndrome. *Clin Genet* **2017**, *92*, 405-414.

18. Yu, L.; Yin, B.; Qu, K.; Li, J.; Jin, Q.; Liu, L.; Liu, C.; Zhu, Y.; Wang, Q.; Peng, X., *et al.* Screening for susceptibility genes in hereditary non-polyposis colorectal cancer. *Oncol Lett* **2018**, *15*, 9413-9419.

19. Zhang, D.; Li, Z.; Xu, X.; Zhou, D.; Tang, S.; Yin, X.; Xu, F.; Li, H.; Zhou, Y.; Zhu, T., *et al.* Deletions at slc18a1 increased the risk of crc and lower slc18a1 expression associated with poor crc outcome. *Carcinogenesis* **2017**, *38*, 1057-1062.

20. Zetner, D.B.; Bisgaard, M.L. Familial colorectal cancer type x. *Curr Genomics* **2017**, *18*, 341-359.

21. Kumar, A.; Bandapalli, O.R.; 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*, 11635.

22. Bandapalli, O.R.; 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*, 2076-2078.

doi:10.20944/preprints202103.0257.v1

23. Srivastava, A.; Kumar, A.; Giangiobbe, S.; Bonora, E.; Hemminki, K.; Forsti, A.; Bandapalli, O.R. Whole genome sequencing of familial non-medullary thyroid cancer identifies germline alterations in mapk/erk and pi3k/akt signaling pathways. *Biomolecules* **2019**, *9*.
24. Srivastava, A.; Miao, B.; Skopelitou, D.; Kumar, V.; Kumar, A.; Paramasivam, N.; Bonora, E.; Hemminki, K.; Forsti, A.; Bandapalli, O.R. A germline mutation in the pot1 gene is a candidate for familial non-medullary thyroid cancer. *Cancers (Basel)* **2020**, *12*.
25. Srivastava, A.; Giangiobbe, S.; Kumar, A.; Paramasivam, N.; Dymerska, D.; Behnisch, W.; Witzens-Harig, M.; Lubinski, J.; Hemminki, K.; Forsti, A., *et al.* Identification of familial hodgkin lymphoma predisposing genes using whole genome sequencing. *Front Bioeng Biotechnol* **2020**, *8*, 179.
26. Srivastava, A.; Giangiobbe, S.; Skopelitou, D.; Miao, B.; Paramasivam, N.; Diquigiovanni, C.; Bonora, E.; Hemminki, K.; Försti, A.; Bandapalli, O.R. Whole genome sequencing prioritizes chek2, ewsr1, and tiam1 as possible predisposition genes for familial non-medullary thyroid cancer. **2021**, *12*.
27. Skopelitou, D.M., Beiping; Srivastava, Aayushi; Kumar, Abhishek; Kuswick, Magdalena; Dymerska, Dagmara; Paramasivam, Nagarajan; Schlesner, Matthias; Lubinski, Jan; Hemminki, Kari; Försti, Asta; Bandapalli, Obul R. Whole exome sequencing identifies apcdd1 and hdac5 genes as potentially cancer predisposing in familial colorectal cancer. *Int. J. Mol. Sci.* **2021**, *22*.
28. Sasawatari, S.; Okamura, T.; Kasumi, E.; Tanaka-Furuyama, K.; Yanobu-Takanashi, R.; Shirasawa, S.; Kato, N.; Toyama-Sorimachi, N. The solute carrier family 15a4 regulates tlr9 and nod1 functions in the innate immune system and promotes colitis in mice. *Gastroenterology* **2011**, *140*, 1513-1525.
29. Kobayashi, T.; Shimabukuro-Demoto, S.; Yoshida-Sugitani, R.; Furuyama-Tanaka, K.; Karyu, H.; Sugiura, Y.; Shimizu, Y.; Hosaka, T.; Goto, M.; Kato, N., *et al.* The histidine transporter slc15a4 coordinates mtor-dependent inflammatory responses and pathogenic antibody production. *Immunity* **2014**, *41*, 375-388.
30. Lee, C.L.; Huang, C.J.; Yang, S.H.; Chang, C.C.; Huang, C.C.; Chien, C.C.; Yang, R.N. Discovery of genes from feces correlated with colorectal cancer progression. *Oncol Lett* **2016**, *12*, 3378-3384.
31. Lahiri, D.K.; 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*, 321-328.
32. Li, H.; Durbin, R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **2009**, *25*, 1754-1760.
33. Li, H. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **2011**, *27*, 2987-2993.
34. Rimmer, A.; Phan, H.; Mathieson, I.; Iqbal, Z.; Twigg, S.R.; Consortium, W.G.S.; Wilkie, A.O.; McVean, G.; Lunter, G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **2014**, *46*, 912-918.
35. Wang, K.; Li, M.; Hakonarson, H. Annovar: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **2010**, *38*, e164.
36. Genomes Project, C.; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A., *et al.* A global reference for human genetic variation. *Nature* **2015**, *526*, 68-74.
37. Smigielski, E.M.; Sirotkin, K.; Ward, M.; Sherry, S.T. Dbsnp: A database of single nucleotide polymorphisms. *Nucleic Acids Res* **2000**, *28*, 352-355.
38. Lek, M.; Karczewski, K.J.; Minikel, E.V.; Samocha, K.E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.H.; Ware, J.S.; Hill, A.J.; Cummings, B.B., *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285-291.

doi:10.20944/preprints202103.0257.v1

39. Petrovski, S.; Wang, Q.; Heinzen, E.L.; Allen, A.S.; Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **2013**, *9*, e1003709.
40. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **2014**, *46*, 310-315.
41. Cooper, G.M.; Stone, E.A.; Asimenos, G.; Program, N.C.S.; Green, E.D.; Batzoglou, S.; Sidow, A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **2005**, *15*, 901-913.
42. Siepel, A.; Bejerano, G.; Pedersen, J.S.; Hinrichs, A.S.; Hou, M.; Rosenbloom, K.; Clawson, H.; Spieth, J.; Hillier, L.W.; Richards, S., *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **2005**, *15*, 1034-1050.
43. Pollard, K.S.; Hubisz, M.J.; Rosenbloom, K.R.; Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **2010**, *20*, 110-121.
44. 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*, 235-241.
45. Karczewski, K.; Francioli, L.; Tiao, G.; Cummings, B.; Alföldi, J.; Wang, Q.; Collins, R.; Laricchia, K.; Ganna, A.; Birnbaum, D., *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv: 2019.
46. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E., *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci Signal* **2013**, *6*, pl1.
47. Cancer Genome Atlas Research, N.; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The cancer genome atlas pan-cancer analysis project. *Nat Genet* **2013**, *45*, 1113-1120.
48. Uhlen, M.; Zhang, C.; Lee, S.; Sjostedt, E.; Fagerberg, L.; Bidkhori, G.; Benfeitas, R.; Arif, M.; Liu, Z.; Edfors, F., *et al.* A pathology atlas of the human cancer transcriptome. *Science* **2017**, *357*.
49. Hecht, M.; Bromberg, Y.; Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **2015**, *16*, S1.
50. Coordinators, N.R. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **2018**, *46*, D8-D13.
51. Papadopoulos, J.S.; Agarwala, R. Cobalt: Constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **2007**, *23*, 1073-1079.
52. Robinson, J.T.; Thorvaldsdottir, H.; Wenger, A.M.; Zehir, A.; Mesirov, J.P. Variant review with the integrative genomics viewer. *Cancer Res* **2017**, *77*, e31-e34.
53. Ernst, J.; Kellis, M. Chromhmm: Automating chromatin-state discovery and characterization. *Nat Methods* **2012**, *9*, 215-216.
54. 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*, 317-330.
55. Hoffman, M.M.; Buske, O.J.; Wang, J.; Weng, Z.; Bilmes, J.A.; Noble, W.S. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **2012**, *9*, 473-476.
56. Fornes, O.; Castro-Mondragon, J.A.; Khan, A.; van der Lee, R.; Zhang, X.; Richmond, P.A.; Modi, B.P.; 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*, D87-D92.
57. Gudis, K.; Tatsuguchi, A.; Wada, K.; Futagami, S.; Nagata, K.; Hiratsuka, T.; Shinji, Y.; Miyake, K.; Tsukui, T.; Fukuda, Y., *et al.* Microsomal prostaglandin e synthase (mpges)-1, mpges-2 and cytosolic pges expression in human gastritis and gastric ulcer tissue. *Lab Invest* **2005**, *85*, 225-236.

58. Korotkova, M.; Daha, N.A.; Seddighzadeh, M.; Ding, B.; Catrina, A.I.; Lindblad, S.; Huizinga, T.W.; Toes, R.E.; Alfredsson, L.; Klareskog, L., *et al.* Variants of gene for microsomal prostaglandin e2 synthase show association with disease and severe inflammation in rheumatoid arthritis. *Eur J Hum Genet* **2011**, *19*, 908-914.

59. Zhang, M.; Chen, F.; Zhang, D.; Zhai, Z.; Hao, F. Association study between slc15a4 polymorphisms and haplotypes and systemic lupus erythematosus in a han chinese population. *Genet Test Mol Biomarkers* **2016**, *20*, 451-458.

60. Zuo, X.B.; Sheng, Y.J.; Hu, S.J.; Gao, J.P.; Li, Y.; Tang, H.Y.; Tang, X.F.; Cheng, H.; Yin, X.Y.; Wen, L.L., *et al.* Variants in tnfsf4, tnfaip3, tnip1, blk, slc15a4 and ube2l3 interact to confer risk of systemic lupus erythematosus in chinese population. *Rheumatol Int* **2014**, *34*, 459-464.

61. Lee, H.S.; Kim, T.; Bang, S.Y.; Na, Y.J.; Kim, I.; Kim, K.; Kim, J.H.; Chung, Y.J.; Shin, H.D.; Kang, Y.M., *et al.* Ethnic specificity of lupus-associated loci identified in a genome-wide association study in korean women. *Ann Rheum Dis* **2014**, *73*, 1240-1245.

62. Wang, C.; Ahlford, A.; Jarvinen, T.M.; Nordmark, G.; Eloranta, M.L.; Gunnarsson, I.; Svenungsson, E.; Padyukov, L.; Sturfelt, G.; Jonsen, A., *et al.* Genes identified in asian sle gwass are also associated with sle in caucasian populations. *Eur J Hum Genet* **2013**, *21*, 994-999.

63. Stelzer, G.; Rosen, N.; Plaschkes, I.; Zimmerman, S.; Twik, M.; Fishilevich, S.; Stein, T.I.; 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.

64. Blum, M.; Chang, H.Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S., *et al.* The interpro protein families and domains database: 20 years on. *Nucleic Acids Res* **2020**.

65. Gordziel, C.; Bratsch, J.; Moriggl, R.; Knosel, T.; Friedrich, K. Both stat1 and stat3 are favourable prognostic determinants in colorectal carcinoma. *Br J Cancer* **2013**, *109*, 138-146.

66. Klampfer, L. The role of signal transducers and activators of transcription in colon cancer. *Front Biosci* **2008**, *13*, 2888-2899.

67. Simpson, J.A.; Al-Attar, A.; Watson, N.F.; Scholefield, J.H.; Ilyas, M.; Durrant, L.G. Intratumoral t cell infiltration, mhc class i and stat1 as biomarkers of good prognosis in colorectal cancer. *Gut* **2010**, *59*, 926-933.

68. Crncec, I.; Modak, M.; Gordziel, C.; Svinka, J.; Scharf, I.; Moritsch, S.; Pathria, P.; Schlederer, M.; Kenner, L.; Timelthaler, G., *et al.* Stat1 is a sex-specific tumor suppressor in colitis-associated colorectal cancer. *Mol Oncol* **2018**, *12*, 514-528.

69. Yamashita, T.; Shimada, S.; Guo, W.; Sato, K.; Kohmura, E.; Hayakawa, T.; Takagi, T.; Tohyama, M. Cloning and functional expression of a brain peptide/histidine transporter. *J Biol Chem* **1997**, *272*, 10205-10211.

70. Bhardwaj, R.K.; Herrera-Ruiz, D.; Eltoukhy, N.; Saad, M.; Knipp, G.T. The functional evaluation of human peptide/histidine transporter 1 (hpht1) in transiently transfected cos-7 cells. *Eur J Pharm Sci* **2006**, *27*, 533-542.

71. Lee, J.; Tattoli, I.; Wojtal, K.A.; Vavricka, S.R.; Philpott, D.J.; Girardin, S.E. Ph-dependent internalization of muramyl peptides from early endosomes enables nod1 and nod2 signaling. *J Biol Chem* **2009**, *284*, 23818-23829.

72. Franchi, L.; Warner, N.; Viani, K.; Nunez, G. Function of nod-like receptors in microbial recognition and host defense. *Immunol Rev* **2009**, *227*, 106-128.

73. Hayden, M.S.; Ghosh, S. Signaling to nf-kappab. *Genes Dev* **2004**, *18*, 2195-2224.

74. Ullman, T.A.; Itzkowitz, S.H. Intestinal inflammation and cancer. *Gastroenterology* **2011**, *140*, 1807-1816.

75. Feagins, L.A.; Souza, R.F.; Spechler, S.J. Carcinogenesis in ibd: Potential targets for the prevention of colorectal cancer. *Nat Rev Gastroenterol Hepatol* **2009**, *6*, 297-305.

76. Garrity-Park, M.M.; Loftus, E.V., Jr.; Bryant, S.C.; Sandborn, W.J.; Smyrk, T.C. Tumor necrosis factor-alpha polymorphisms in ulcerative colitis-associated colorectal cancer. *Am J Gastroenterol* **2008**, *103*, 407-415.

77. Suchy, J.; Klujszo-Grabowska, E.; Kladny, J.; Cybulski, C.; Wokolorczyk, D.; Szymanska-Pasternak, J.; Kurzawski, G.; Scott, R.J.; Lubinski, J. Inflammatory response gene polymorphisms and their relationship with colorectal cancer risk. *BMC Cancer* **2008**, *8*, 112.

78. Crivello, A.; Giacalone, A.; Vaglica, M.; Scola, L.; Forte, G.I.; Macaluso, M.C.; Raimondi, C.; Di Noto, L.; Bongiovanni, A.; Accardo, A., *et al.* Regulatory cytokine gene polymorphisms and risk of colorectal carcinoma. *Ann N Y Acad Sci* **2006**, *1089*, 98-103.