Medical Research Archives

ESMED
European Society of Medicine

OPEN ACCESS

Published: April 30, 2022

Citation: Liu W, Gonn M, et al., 2022. Linkage And Association Analysis Define Novel Regions for The Risk Of Adenomas and Colorectal Cancer, Medical Research Archives, [online] 10(4). https://doi.org/10.18103/mra.v10i4.2780

DOI: https://doi.org/10.18104/mra.v10i4.2780

ISSN: 2375-1924

RESEARCH ARTICLE

Linkage and association analysis define novel regions for the risk of adenomas and colorectal cancer

Wen Liu[1,2], Mark Gonn[3], Susanna von Holst[1], Jessada Thutkawkorapin[1], Xiang Jiao[1], Jan Björk2[4,5], Ann-Sofie Backman[4,5], Kristina Lagerstedt-Robinson[1,6], Annika Lindblom[1,6*]

1. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden
2. Department of Neuroscience, Uppsala University, Uppsala, Sweden
3. Unit of Internal Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden
4. Department of Medicine, Solna, Karolinska Institutet, Stockholm, Sweden
5. Gastroenterology section, medical unit Gastroenterology, Rheumatology and Dermatology, Karolinska University hospital, Stockholm, Sweden
6. Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden

* annika.lindblom@ki.se

ABSTRACT

Colorectal cancer (CRC) is a multifactorial disease, where both the environment and genetics play a role. It is estimated that approximately 35% of CRCs have a potentially identifiable genetic cause. Well-known and highly penetrant genetic causes make up less than 5% of all CRC, and leave many families not explained by known predisposing genes/mutations. Low penetrant alleles have also been thought to modify the risk of CRC. Linkage studies have been successful in discovering and localizing highly penetrant genes in CRC and risk loci has become possible to discover performing genome wide association studies (GWAS).

In this study we have analyzed families with CRC where individuals with CRC as well as individuals with premalignant lesions, adenomas, were codes as affected. In total 600 individuals in 121 families were included in the study.

In total three genomic regions were found with suggestive linkage located at 4p16.3, 6p24.3 and 10p14. These regions were further studied using sequencing analysis and association studies using haplotypes.

INTRODUCTION

Colorectal cancer (CRC) is a multifactorial disease, where both the environment and genetics play a role [1]. It has been shown that family history is a major risk factor. Subjects with one first degree relative with CRC diagnosed at age under 50 years, have a relative risk of developing CRC of 2-3 [2]. It is estimated that approximately 35% of CRCs have a potentially identifiable genetic cause [3]. Among the most well-known genetic causes are the monogenic syndromes Familial Adenomatous Polyposis (FAP) and Lynch´s syndrome (LS), which make up less than 5% of all CRC [1], and leave many families not explained by known predisposing genes/mutations.

Linkage studies in familial CRC have been successful in discovering and localizing highly penetrant genes such as *APC, MSH2, MLH1*, and the most recent addition of *GREM1* [4]. Low penetrant alleles have also been thought to be able to modify the risk of CRC. By performing genome wide association studies (GWAS) it has become possible to discover several risk loci using thousands of cases and controls [5, 6].

CRC is in most cases preceded by premalignant lesions known as adenomatous polyps (adenomas). There are well supported scientific theories which stepwise explain the developmental process of adenomas to cancer; "the serrated adenoma pathway" and "the adenoma carcinoma pathway" [7, 8]. Even though more than 90 % of adenomas may not progress to cancer, some features might be helpful in evaluating their potential to undergo transformation. Adenomas more than one cm in width, or with mainly villous architecture or a high grade of dysplasia are called high-risk adenomas with greater probability of transformation [9]. Even small adenomas could be considered precursors to colorectal cancer, particularly in patients known to be at increased risk due to their family history [10].

In our recent linkage study, both patients with CRC and high-risk adenomas were coded as affected [11]. Since family members in high-risk families have been included in surveillance for many years, several of those family members have been diagnosed with adenomas over time [12]. The present study aims to investigate the power of also considering family members with small adenomas as putative gene carriers. The same genotyping data as in the previous study was used to reanalyze 121 families with familial CRC. Family members with low-risk adenomas were recoded and considered as affected. The hypothesis was that loci harboring risk genes for adenomas and CRC would generate higher LOD scores and subsequently possible disease-causing loci would become more distinct when more individuals were coded as affected. In total, 66 individuals were recoded and added as affected in comparison with the previous study.

All individuals were first included in linkage analysis followed by exome sequencing to search for high-risk gene mutations located in the suggested chromosomal regions, as seen in monogenic diseases. Next, association studies of 484 CRC cases and 1642 controls were performed to also test the hypothesis of low-risk gene candidates in the same suggested regions to explain the familiarity as seen in complex inheritance.

MATERIALS AND METHODS

Ethics statement

The study was undertaken in agreement with the Swedish legislation of ethical permission and according to the decision in the Stockholm regional ethical committee (2008/125-31.2 and 2002/489). All participants gave written informed consent to participate in the study.

Patients and healthy controls used for the studies

Linkage studies: In total 121 Swedish families with 600 individuals with an increased risk of developing CRC were included in the linkage study [11]. Information about the families were retrieved from the Department of clinical genetics at the Karolinska University Hospital in Stockholm, Sweden between years 1990 -

2005. Families were included in the study if there were at least two affected relatives informative for linkage analysis. Family members underwent colonoscopy during this time period and the findings of either polyps or CRC were documented. FAP was excluded using medical records from affected individuals and LS was excluded using our current clinical protocol [13]. High-risk families were defined as families with more than three affected individuals in more than one generation. Moderate risk families were defined as families with two or more sibs affected in one generation.

Exome studies: Genomic DNA isolated from 140 CRC cases from in total 65 CRC families were used for exome sequencing initially for a separate study (published PMID: 33729574). The samples were from the 121 families in the linkage study, plus a few other families not used in linkage analysis.

Association studies: Data from a previous CRC GWAS [6], 484 familial cases vs 1642 controls, were used as a secondary test for the four detected loci. The samples were consecutive cases from the Stockholm-Uppsala region and healthy controls were blood donors from the same region.

Haplotype studies: CRC families with one affected case and one first degree relative were used for haplotype studies. 62 familial CRC cases (34 from the families used in the linkage study and 28 from other families) and one child each were used for genome wide genotyping for testing candidate haplotypes from association studies.

Genotyping for linkage and haplotype analysis

Genotyping of 548 family members from CRC families with 6090 markers for linkage analysis was performed as described [11]. For the association study, CRC cases and controls were genotyped at the Center for Inherited Disease Research at Johns Hopkins University, US using the Illumina Infinium® OncoArray-500K BeadChips (6). To test suggested haplotypes from the association study, 62 familial cases and one child, were genotyped using the same procedure as described for the linkage analysis [11].

Linkage analysis

In the linkage analysis all individuals with CRC or any adenomas were coded as affected. All other family members were coded as unknown. The families were divided into two different groups; high risk (more than 3 affected individuals in more than one generation), and moderate risk (two or more siblings affected in one generation).

In total 121 families were used with 7256 markers spread along the genome. Since four families were too large to run through the MERLIN software they had to be split - finally adding up to 126 families. The four families were split so that each sub-family used one common ancestor and fitted into the limit as defined while running the program [11].

Pedcheck was used for the initial control of Mendelian inheritance analysis among families [14]. A parametric linkage analysis was used for all chromosomes. As a supplement non-parametric analysis using Whittemore and Halpern NPL statistics was made [15]. MERLIN was then used to detect any genetic marker inconsistencies and to compute LOD- and heterogeneity LOD (HLOD) scores. Analyses were done for both recessive and dominant mode of inheritance, with the disease allele frequency set to 0.0001. The penetrance rates for the dominant and recessive mode of inheritance for homozygous normal, heterozygous, and homozygous affected were set to 0.05, 0.80, 0.80 and 0.001, 0.001, 1.0 respectively. Since presence of linkage disequilibrium (LD) may inflate multipoint linkage statistics, a threshold of $r^2 = 0.1$ were used to avoid false positive results inflating the statistics [16]. LD among SNPs with $r^2>0.1$ was accounted for by MERLIN organizing the markers into clusters. MERLIN makes use of the population haplotype frequencies to assume LD within each cluster. To maintain uniformity in our study subsets, the same clusters were continuously used in all analysis.

Exome sequencing of CRC samples

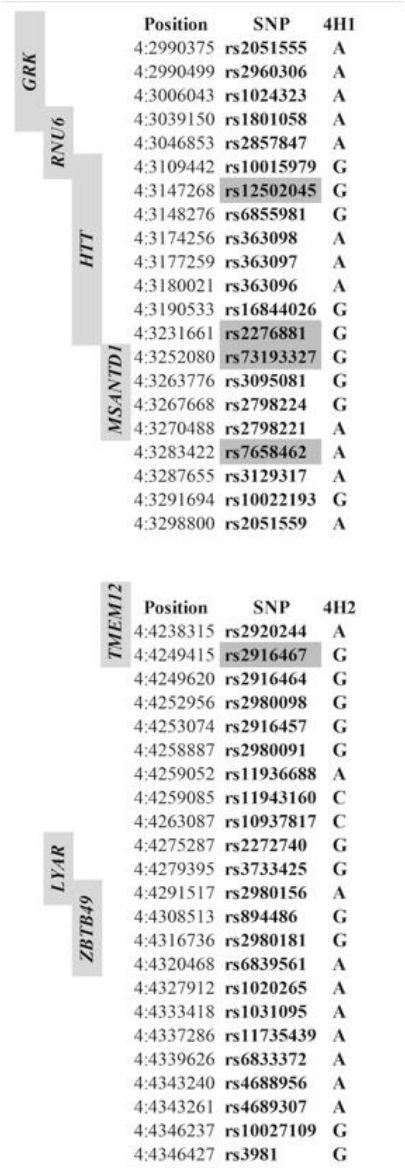
Genomic DNA was prepared from peripheral blood using standard protocols and quantified using a Qubit Fluorometer (Life Technologies, US). Sequencing libraries were prepared according to the TruSeq DNA Sample Preparation Kit EUC 15005180 or EUC 15026489 (Illumina, US). Briefly, 1-1.5 ug of genomic DNA was fragmented using the Covaris 400 bp protocol (Covaris, Inc., US). After fragmentation, all samples were subjected to end-repair, A-tailing, and adaptor ligation of Illumina Multiplexing PE adaptors. An additional gel-based size selection step was performed, and the adapter-ligated fragments were subsequently enriched by PCR followed by purification using Agencourt AMPure Beads (Beckman Coulter, Sweden). Exome capture was performed by pre-pooling equimolar amounts and performing enrichment in 5- or 6-plex reactions according to the TruSeq Exome Enrichment Kit Protocol (EUC 15013230). Library size was checked on a Bioanalyzer High Sensitivity DNA chip (Agilent Technologies, Sweden) while concentration was calculated by quantitative PCR. The pooled DNA libraries were clustered on a cBot instrument (Illumina) using the TruSeq PE Cluster Kit v3. Paired-end sequencing was performed for 100 cycles using a HiSeq 2000 instrument (Illumina) with TruSeq SBS Chemistry v3, according to the manufacturer's protocol. Base calling was performed with RTA (1.12.4.2 or 1.13.48) and the resulting BCL files were filtered, de-multiplexed, and converted to FASTQ format using CASAVA 1.7 or 1.8 (Illumina). The sequencing was performed at an average coverage of 100x.

Bioinformatics workflow

Sequencing reads were aligned to the reference genome GRCh37 using BWA [17]. Aligned reads were sorted and PCR-duplicated reads were removed using Picard (http://broadinstitute.github.io/picard/). The calculation of mapping and enrichment statistics were done with Picard and GATK. Variants were called using GATK by following the best practice procedure implemented at the Broad Institute. Variant Quality Score Recalibration from GATK were used for quality control of the variants. Variant annotation was done by ANNOVAR (version 2016-Feb-01). The annotated information includes RefSeq gene annotation (version 73) and dbSNP rs number (version 138). Background allele frequencies were from 1000 Genomes Project allele frequencies. Sequence variant analysis was performed for the regions 4p16.3, 6p24.3 and 10p14.

Quality control of the data for association study

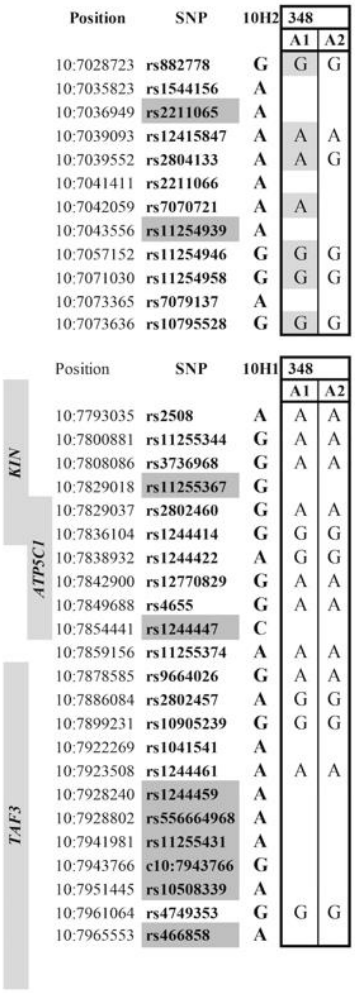
In total 4,381 individuals (2,709 cases and 1,672 controls) and 516,258 markers were included in the analysis [6]. Haploid genotypes, genotypes with gender inconsistency or genotypes with same position variants were excluded resulting in 344,234 SNPs. In the next step SNPs with <98% call rate, <5% minor allele frequency (MAF) and those inconsistent with Hardy–Weinberg equilibrium in controls were removed (markers were excluded that failed the Hardy-Weinberg test at a specified significance threshold: hwe 0.001), thus 342,359 SNPs remained. In the final step, a multidimensional scaling (MDS) analysis was conducted on all the remaining markers for the purpose of population stratification and to identifying ethnic outliers among samples. These outliers were excluded from the dataset while the remaining were plotted in an MDS plot. In total, 342,359 SNPs and 4,305 individuals remained (2,663 cases and 1,642 controls) to be used in the analyses. In this paper we only used a subset of 484 familial cases and all 1,642 controls.



	Position	SNP	4H1
GRK	4:2990375	rs2051555	A
	4:2990499	rs2960306	A
	4:3006043	rs1024323	A
RNU6	4:3039150	rs1801058	A
	4:3046853	rs2857847	A
	4:3109442	rs10015979	G
HTT	4:3147268	rs12502045	G
	4:3148276	rs6855981	G
	4:3174256	rs363098	A
	4:3177259	rs363097	A
	4:3180021	rs363096	A
	4:3190533	rs16844026	G
	4:3231661	rs2276881	G
MSANTD1	4:3252080	rs73193327	G
	4:3263776	rs3095081	G
	4:3267668	rs2798224	G
	4:3270488	rs2798221	A
	4:3283422	rs7658462	A
	4:3287655	rs3129317	A
	4:3291694	rs10022193	G
	4:3298800	rs2051559	A

	Position	SNP	4H2
TMEM12	4:4238315	rs2920244	A
	4:4249415	rs2916467	G
	4:4249620	rs2916464	G
	4:4252956	rs2980098	G
	4:4253074	rs2916457	G
	4:4258887	rs2980091	G
	4:4259052	rs11936688	A
	4:4259085	rs11943160	C
	4:4263087	rs10937817	C
LYAR	4:4275287	rs2272740	G
	4:4279395	rs3733425	G
	4:4291517	rs2980156	A
ZBTB49	4:4308513	rs894486	G
	4:4316736	rs2980181	G
	4:4320468	rs6839561	A
	4:4327912	rs1020265	A
	4:4333418	rs1031095	A
	4:4337286	rs11735439	A
	4:4339626	rs6833372	A
	4:4343240	rs4688956	A
	4:4343261	rs4689307	A
	4:4346237	rs10027109	G
	4:4346427	rs3981	G

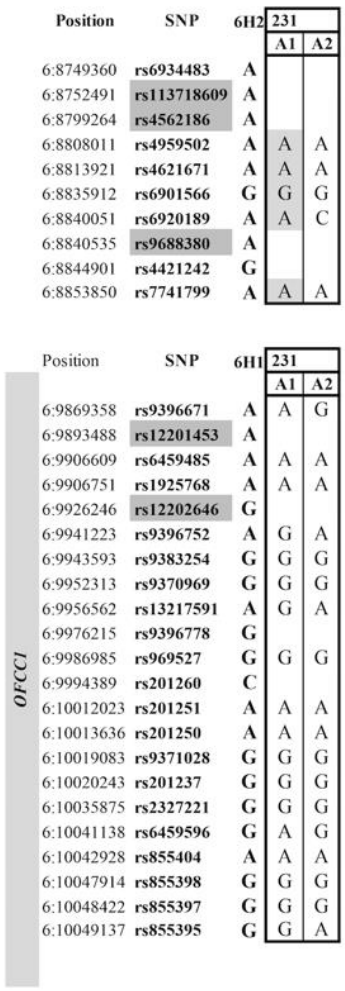
Figure 1a



	Position	SNP	10H2	348	
				A1	A2
	10:7028723	rs882778	G	G	G
	10:7035823	rs1544156	A		
	10:7036949	rs2211065	A		
	10:7039093	rs12415847	A	A	A
	10:7039552	rs2804133	A	A	G
	10:7041411	rs2211066	A		
	10:7042059	rs7070721	A	A	
	10:7043556	rs11254939	A		
	10:7057152	rs11254946	G	G	G
	10:7071030	rs11254958	G	G	G
	10:7073365	rs7079137	A		
	10:7073636	rs10795528	G	G	G

	Position	SNP	10H1	348	
				A1	A2
KIN	10:7793035	rs2508	A	A	A
	10:7800881	rs11255344	G	A	A
	10:7808086	rs3736968	G	A	A
	10:7829018	rs11255367	G		
ATP5C1	10:7829037	rs2802460	G	A	A
	10:7836104	rs1244414	G	G	G
	10:7838932	rs1244422	A	G	G
	10:7842900	rs12770829	G	A	A
	10:7849688	rs4655	G	A	A
	10:7854441	rs1244447	C		
	10:7859156	rs11255374	A	A	A
	10:7878585	rs9664026	G	A	A
	10:7886084	rs2802457	A	G	G
	10:7899231	rs10905239	G	G	G
	10:7922269	rs1041541	A		
	10:7923508	rs1244461	A	A	A
	10:7928240	rs1244459	A		
TAF3	10:7928802	rs556664968	A		
	10:7941981	rs11255431	A		
	10:7943766	c10:7943766	G		
	10:7951445	rs10508339	A		
	10:7961064	rs4749353	G	G	G
	10:7965553	rs466858	A		

Figure 1b



	Position	SNP	6H2	231	
				A1	A2
	6:8749360	rs6934483	A		
	6:8752491	rs113718609	A		
	6:8799264	rs4562186	A		
	6:8808011	rs4959502	A	A	A
	6:8813921	rs4621671	A	A	A
	6:8835912	rs6901566	G	G	G
	6:8840051	rs6920189	A	A	C
	6:8840535	rs9688380	A		
	6:8844901	rs4421242	G		
	6:8853850	rs7741799	A	A	A

	Position	SNP	6H1	231	
				A1	A2
	6:9869358	rs9396671	A	A	G
	6:9893488	rs12201453	A		
	6:9906609	rs6459485	A	A	A
	6:9906751	rs1925768	A	A	A
	6:9926246	rs12202646	G		
	6:9941223	rs9396752	A	G	A
	6:9943593	rs9383254	G	G	G
	6:9952313	rs9370969	G	G	G
	6:9956562	rs13217591	A	G	A
	6:9976215	rs9396778	G		
OFCC1	6:9986985	rs969527	G	G	G
	6:9994389	rs201260	C		
	6:10012023	rs201251	A	A	A
	6:10013636	rs201250	A	A	A
	6:10019083	rs9371028	G	G	G
	6:10020243	rs201237	G	G	G
	6:10035875	rs2327221	G	G	G
	6:10041138	rs6459596	G	A	G
	6:10042928	rs855404	A	A	A
	6:10047914	rs855398	G	G	G
	6:10048422	rs855397	G	G	G
	6:10049137	rs855395	G	G	A

Figure 1c

Figure 1: Matching haplotypes on the different chromosomes and the location of known genes in this region – chromosome 4 (H1 and H2) (Figure 1A), chromosome 10 (H1 and H2) (Figure 1B) and chromosome 6 (H1 and H2) (Figure 1C). All positions are annotated according to GRCh37. SNP, single nucleotide polymprphism.

Statistical Analysis

A logistic regression model was employed to examine the association between one single SNP or haplotype and cancer risk. A sliding window design was used. Corresponding odds ratio (OR), standard errors, 95% confidence intervals and P-values were subsequently calculated. Statistical analysis and plot generation were conducted using PLINK v1.07 (18). P-values were modified to correct for multiple testing used generated p-values divided by the number of SNPs in each haplotype in Table 3.

RESULTS

Linkage analysis

Linkage analysis was carried out using three different subsets of families and both recessive

and dominant inheritance were tested. Linkage analysis was performed in 27 high risk families (more than three affected individuals in more than one generation), 49 moderate risk families (two or more sibs affected in one generation) and finally in all 121 families.

The analyses of the 121 Swedish CRC families, where all individuals with adenomas were set to have an affected status did not generate any statistically significant (>3) LOD/HLOD score. However, there were positive LOD/HLOD scores above 2 in the sub-studies, which are suggestive of linkage (Table 1). The same table also show those families contributing most to each locus. One region on 4p16.3 was found using recessive analysis (max HLOD for the marker rs736455) for the high-risk families (Table 1). Another locus was found on 10p14 in the moderate risk families using an autosomal dominant model. All loci with LOD/HLOD above 1 are listed in Supplemental Table 1.

Study group	Max SNP	Locus	LOD	HLOD	Model	Families
High risk	rs736455	4p16.3	0.9	2.3	AR	8,110, 478, 740_2
Moderate risk	rs942434	10p14	2.1	2.1	AD	106, 231, 324-1,348-1,663, 849
Family 231	rs761116	6p24.3	2.2	1.5	AD	231

Table 1: Best LOD / HLOD from linkage analysis.

Besides, one family (no 231) had a LOD of 2.2 (between markers rs767022 - rs561332) of its own as the separate best score for a third locus on 6p24.3. No other family in this study had the power to generate a similar LOD score. For the locus on 6p24.3 a few other families also had positive LOD of 0.2-0.5 but the overall max LOD and HLOD for this locus was only 0.3 and 1.5 respectively. Family number 231 had eight individuals with low-risk adenomas in the linkage study, explaining why this family was given special attention.

Sequencing analysis in the three detected regions using high-risk hypothesis

Linkage analysis was used in families with the hypothesis that the families segregated with dominant or recessive high-penetrant disease, and we used sequencing to search for pathogenic mutations in genes located in the three regions with suggestive linkage.

Exome sequencing was performed in up to four members, in a total of 194 members from 62 families, including seven of the linked families (8, 110, 231, 348, 478, 740 and 849) (Table 1). The three regions of interest were analyzed. In short, exonic and splice variants in familial CRC family members were selected, synonymous and unknown variants were removed. Variants were also eliminated if the frequency was higher in the ExAC database (using the European population as well as all individuals in this database) compared to our familial CRC cases. As a final step the sequence variant had to be more than twice as common in our CRC families compared to the European population in the ExAC database. Any filtered sequence variant had to segregate in at least one of the families contributing to the LOD score (Table 1). Since only exonic and splice mutations were searched for we did not require two hits in any family. The results are presented in Table 2.

Locus	Position	SNP	Variant	Gene	AF	MAF_Eur	Model	Family
4p16.3	4322624	rs34623124	C>G	*ZBTB49*	0.0342	0.0149	AR	110
10p14	7774317	rs41290291	T>C	*ITIH2*	0.0089	0.0040	AD	849
10p14	8006519	rs17366712	G>C	*TAF3*	0.0804	0.0258	AD	231

Table 2: Sequence variants suggested from the high-risk analysis.

Association studies for the three regions using a low-risk hypothesis

The hypothesis of low-risk mutations as in complex disease was tested with haplotype-

association analysis using 484 familial CRC cases and 1642 healthy controls. The results did not show any statistically significant results, although borderline significant results were observed. Since a sliding window haplotype analysis was used, many haplotypes with different sizes represented the same target (five best haplotypes are shown in Supplemental Table 2). The two best different haplotypes for each locus was chosen for testing in 62 familial colorectal cancer cases with available haplotype-data (Table 3, Supplementary Table 3). Since the 62 cases were genotyped with a different SNP assay all six haplotypes had incomplete information. The number of individuals who could possibly have the candidate haplotypes were used to calculate haplotype frequency (FA62 in Table 3, Figure 1 a-c). Depending on the many incomplete alleles no conclusion could be drawn from the comparison of the allele frequency among familial cases in the association study and the 62 other familial cases. None of the linked families had any of the two best haplotypes for region on chromosome 4p16.3 (4:H1 and 4:H2 Table 3, Figure 1a). For the region on chromosome 10p.14 family 348 could possibly have the 10:H2 haplotype (Table 3, Figure 1b). Regarding the region on chromosome 6p24.3, chosen because of family 231, this family could possibly have the 6:H2 haplotype (Table 3, Figure 1c).

The genes involved as low-risk genes within these haplotypes were for 4p16.3 – (4:H1); *GRK4, RNU6, HTT, MSANTD1*, (4:H2); *TMEM128, ZBTB49,* and *LYAR* and for 10p14 – (10:H1); *KIN, ATP5C1* and *TAF3*, and finally, for 6p24.1 – (6:H1); *OFCC1* (Figure 1a-c). A search for mutations within exons of these genes was already performed and exonic variants were found in the genes *ZBTB49* and *TAF3* on chromosomes 4p16.3 and 10p14 (Table 2).

Chr	Position	Haplotype number	Haplotype	FA	FU	FA62	OR	P-value (req p-value)
4p16.3	2990375-3298800	4:H1	AAAAAGGGAAAGGGGGAAAGA	0.05	0.02	0.06	2.33	1.00E-05 (5.88E-06)
	4238315-4346427	4:H2	AGGGGGACCGGAGGAAAAAAAGG	0.05	0.02	0.05	2.84	1.20E-05 (5.88E-06)
10p14	7793035-7965553	10:H1	AGGGGGAGGCAGAGAAAAAGAGA	0.06	0.03	0.11	1.94	1.56E-05 (7.17E-06)
	7028723-7073636	10:H2	GAAAAAAAGGAG	0.08	0.05	0.04	1.71	3.37E-05 (7.17E-06)
6p24.3	9869358-10049137	6:H1	AAAAGAGGAGGCAAGGGGAGGG	0.04	0.01	0.11	3.36	1.57E-06 (3.37E-06)
	8749360-8853850	6:H2	AAAAAGAAGA	0.03	0.01	0.11	2.63	2.18E-06 (3.37E-06)

Table 3: Haplotype frequency and odds ratio for two best different haplotypes per chromosomal region with suggestive linkage.
Chr - chromosomal localization; Position – genomic position in Hg19: FA - haplotype frequency in affected: FU – haplotype frequency in unaffected: FA62 - potential haplotype frequency in 62 familial CRC cases: OR - odds ratio: Req p-value – calculated required p-value for significant result

DISCUSSION

Linkage analysis in 121 families was used in this study, in conformity with our previous study:" Linkage analysis in Familial Non-Lynch Syndrome Colorectal Cancer Families from Sweden" [11]. The difference between the two studies is the affected status criteria. In this study we considered individuals as affected when presenting with both high- and low risk adenomas, whilst in our previous linkage study, only patients with colorectal cancer or high-risk adenomas were considered as affected. The rationale behind the study was still that the families included had a family history suggesting high-risk colorectal cancer. Since the cohort used for the first study was small, we wanted to increase the power. It was not possible to include more families, but more family members could be coded as affected by changing criteria for affected status to also apply to those with adenomas less than 10mm. The hypothesis was that even early adenomas

could predict a gene carrier. Still, no statistically significant results were found. Two loci (4p16.3 and 10p14) had LODs and/or HLODs above 2, suggestive of linkage. One single family with many family members with small adenomas (family 231) had a max LOD >2 for one locus at 6p24.3.

The locus on chromosome 4p16.3 was found also in the previous study where small adenomas were not considered as risk factors [11]. The HLOD improved in this analysis from 2,1 to 2,3. The locus on chromosome 10p14, was not observed in our recent linkage analysis. The LODs for family 231 increased by the fact that additional eight persons with small adenomas were used as affected in analysis. However, this locus was not supported by other families and the overall LOD/HLOD for the locus was <1 and therefor considered of less interest.

Since high-risk disease was the first hypothesis tested for this cohort of families, exome sequencing of 23 affected family members from seven linked families were analyzed. No strong candidate gene was suggested by sequencing analysis among our linked families. However, there was one gene possibly involved in at least one (family 110) of the four families mostly contributing to the LOD score at 4p16.3, *ZBTB49* and two genes, *ITIH2* and *TAF3*, on 10p14 suggested for families 845 and 231. No gene was suggested in family 231 for the chromosome 6p24.3 region. A limitation was that only three of the six families contributing most to the LOD on chromosome 10p14 were sequenced, which means that we could have missed genes and mutations. Moreover, since exome sequencing was used, mutations outside the exons were not studied but could still be of importance and thus missed. The locus on 4p16.3 has not been described in previous GWAS but 10p14 has been suggested to harbor low-risk colorectal cancer predisposing genes and our study supports this [19, 20]. However, the chromosomal 4p16.3 region has been implicated and suggested as a tumor suppressor gene in colorectal cancer because a high frequency of loss of heterozygosity in a Chinese study [21].

The genes in the haplotypes suggested by association studies were on 4p16.3; *GRK4, RNU6, HTT, MSANTD1, ZBTB49, LYAR* and *TMEM128,* and on 10p14; *KIN, ATP5C1* and *TAF3*, and finally the *OFCC1* gene in the 6p24.3 region. None of those genes was suggested to be a high-risk gene but it is possible that some of them could contribute to CRC risk as low-risk alleles. In fact, the genes *ZBTB49* and *TAF3* were also suggested by mutations in the sequenced families and also by the haplotype analysis and are therefore the best candidate genes from this study. The *ZBTB49* different isoforms have been suggested to be induced by TP53 or induce RB1 and has been suggested having a tumor suppressor function [22]. The *TAF3* gene was suggested from the sequencing in some of the family members in family 231, as well as a possible target in one of the two risk-haplotypes on 10p14. The *CTCF* gene directly recruits *TAF3* to promoter distal sites and a role for *TAF3* in pluripotency has been suggested [23]. One gene of interest was the *HTT* gene, and in this study, no variant in this gene was segregating in the linked families. However, there were several frameshift or missense nonsynonymous variants among the sequenced CRC cases with very rare or unique mutations in this gene. None of those mutations were expected to cause Huntington´s disease (HD). The *HTT* gene has been suggested to influence CRC risk and cancer is less common than expected in the HD population. However, this does not appear to be related to glutamine-length in *HTT* [24].

In spite using high-risk families, the lack of high-risk gene in our study was not too surprising. We previously spent many years searching for high-risk genes in loci suggested by linkage analysis in one breast- and one colorectal cancer family, until we finally could conclude that the linkage results was explained by low penetrant mutations as in a complex inherited disease [25, 26]. Instead we set up to test for low-

risk genes using association analysis in 484 familial CRC cases and controls. Haplotype analysis had been shown to be more powerful than single SNP analysis [25-28], why haplotype analysis was used to test all three regions. The numbers of familial cases and controls were small and only borderline statistically significant results were found (Supplemental Table 2, Table 3). However, the genes implicated from this study could still be contributing to CRC risk.

The lack of statistically significant results in the study is obviously an issue but very difficult to solve since increasing the number of families for linkage and association will not necessarily improve the LODs/HLODs or ORs. One reason for this could be the heterogeneity in the nature of cancer disease. Increasing the number of cases and controls might help only if you add samples with same genetic background – such as the same genes involved in high-risk disease or similar allelic distribution for the low-risk studies. In fact, we have noticed that when studying subgroups selected for various phenotypes, we always get higher LOD scores compared with the analysis from the whole group [11].

CONCLUSION

In conclusion, we started to search for high-penetrant disease in familial colorectal cancer and ended up with a hypothesis of complex disease also in families where high-penetrant disease was suggested from family history. We have had similar results before [25-26,28]. Further studies need to be done to find high-, moderate- or low-risk genes acting together in familial colorectal cancer.

Conflict: Authors declared no conflict of interest

Supplementary Tables

Please find supplementary tables here.

REFERENCES

1. Haggar FA, Boushey RP. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg*. 2009;22(4):191-7.
2. Vasen HF, van der Meulen-de Jong AE, de Vos Tot Nederveen Cappel WH, Oliveira J, Group EGW. Familial colorectal cancer risk: ESMO clinical recommendations. *Annals of oncology : official journal of the European Society for Medical Oncology*. 2009;20 Suppl 4:51-3.
3. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England Journal of Medicine*. 2000;343(2):78-85.
4. Jaeger EE, Woodford-Richens KL, Lockett M, Rowan AJ, Sawyer EJ, Heinimann K, et al. An ancestral Ashkenazi haplotype at the HMPS/CRAC1 locus on 15q13-q14 is associated with hereditary mixed polyposis syndrome. *American Journal of Human Genetics*. 2003;72(5):1261-7.
5. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun*. 2019;10(1):2154.
6. Schmit SL, Edlund CK, Schumacher FR, Gong J, Harrison TA, Huyghe JR, et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *Journal of the National Cancer Institute*. 2019;111(2):146-57.
7. Jass JR. Serrated adenoma of the colorectum: a lesion with teeth. *Am J Pathol*. 2003;162(3):705-8.
8. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell*. 1990;61(5):759-67.
9. O'Brien MJ, Winawer SJ, Zauber AG, Gottlieb LS, Sternberg SS, Diaz B, et al. The National Polyp Study. Patient and polyp characteristics associated with high-grade dysplasia in colorectal adenomas. *Gastroenterology*. 1990;98(2):371-9.
10. Sillars-Hardebol AH, Carvalho B, van Engeland M, Fijneman RJ, Meijer GA. The adenoma hunt in colorectal cancer screening: defining the target. *The Journal of Pathology*. 2012;226(1):1-6.
11. Kontham V, von Holst S, Lindblom A. Linkage analysis in familial non-Lynch syndrome colorectal cancer families from Sweden. *PLoS One*. 2013;8(12):e83936.
12. Forsberg A, Kjellstrom L, Andreasson A, Jaramillo E, Rubio CA, Bjorck E, et al. Colonoscopy findings in high-risk individuals compared to an average-risk control population. *Scand J Gastroenterol*. 2015;50(7):866-74.
13. Lagerstedt-Robinson K, Rohlin A, Aravidis C, Melin B, Nordling M, Stenmark-Askmalm M, et al. Mismatch repair gene mutation spectrum in the Swedish Lynch syndrome population. *Oncology Reports*. 2016;36(5):2823-35.
14. O'Connell JR, Weeks DE. PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*. 1998;63(1):259-66.
15. Whittemore AS, Halpern J. A class of tests for linkage using affected pedigree members. *Biometrics*. 1994;50(1):118-27.
16. Goode EL, Badzioch MD, Jarvik GP. Bias of allele-sharing linkage statistics in the presence of intermarker linkage disequilibrium. *BMC Genetics*. 2005;6 Suppl 1:S82.
17. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-95.
18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 2007;81(3):559-75.
19. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature Genetics*. 2008;40(5):623-30.
20. Whiffin N, Hosking FJ, Farrington SM, Palles C, Dobbins SE, Zgaga L, et al.

Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum Mol Genet*. 2014;23(17):4729-37.

21. Zheng HT, Jinag LX, Ly ZC, Li DP, Zhou CZ and Gao JJ et al. Are there tumor suppressor genes on chromosome 4p in sporadic colorectal carcinoma. *World J Gastroenterol*. 7;14(1):90-94.

22. Jeon BN, Kim MK, Yoon JH, Kim MY, An H, Noh HJ, et al. Two ZNF509 (ZBTB49) isoforms induce cell-cycle arrest by activating transcription of p21/CDKN1A and RB upon exposure to genotoxic stress. *Nucleic Acids Research*. 2014;42(18):11447-61.

23. Liu Z, Scannell DR, Eisen MB, Tjian R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell*. 2011;146(5):720-31.

24. McNulty P, Pilcher R, Ramesh R, Neucuiniate R, Hughes A, Farewell D et al. Reduced cancer incidence in Huntingtons´s disease: Analysis in the registry study. *J Huntingtons Dis*. 2018;7(3):209-222.

25. Jiao X, Aravidis C, Marikkannu R, Rantala J, Picelli S, Adamovic T, et al. PHIP - a novel candidate breast cancer susceptibility locus on 6q14.1. *Oncotarget*. 2017;8(61):102769-82.

26. Thutkawkorapin J, Mahdessian H, Barber T, Picelli S, von Holst S, Lundin J, et al. Two novel colorectal cancer risk loci in the region on chromosome 9q22.32. *Oncotarget*. 2018;9(13):11170-9.

27. Liu W, Jiao X, Thutkawkorapin J, Mahdessian H, Lindblom A. Cancer risk susceptibility loci in a Swedish population. *Oncotarget*. 2017;8(66):110300-10.

28. von Holst S, Jiao X, Liu W, Kontham V, Thutkawkorapin J, Ringdahl J, et al. Linkage analysis revealed risk loci on 6p21 and 18p11.2-q11.2 in familial colon and rectal cancer, respectively. *European Journal of Human Genetics*. 2019;27(8):1286-95.

29. Jiao X, Liu W, Mahdessian H, Bryant P, Ringdahl J, Timofeeva M, et al. Recurrent, low-frequency coding variants contributing to colorectal cancer in the Swedish population. *PLoS One*. 2018;13(3):e0193547.

30. Thutkawkorapin J, Picelli S, Kontham V, Liu T, Nilsson D, Lindblom A. Exome sequencing in one family with gastric- and rectal cancer. *BMC Genetics*. 2016;17:41.