ACTA ABP BIOCHIMICA POLONICA

Vol. 66, No 3/2019
305–313
https://doi.org/10.18388/abp.2019_2775

Regular paper

Functional analyses of a low-penetrance risk variant rs6702619/1p21.2 associating with colorectal cancer in Polish population

Malgorzata Statkiewicz[1#], Natalia Maryan[2#], Maria Kulecka[2], Urszula Kuklinska[1], Jerzy Ostrowski[1,2] and Michal Mikula[1]✉

[1]Department of Genetics, Maria Sklodowska-Curie Institute and Oncology Center; Warsaw, Poland; [2]Department of Gastroenterology, Hepatology and Clinical Oncology, Medical Center for Postgraduate Education; Warsaw, Poland

Several studies employed the genome-wide association (GWA) analysis of single-nucleotide polymorphisms (SNPs) to identify susceptibility regions in colorectal cancer (CRC). However, the functional studies exploring the role of associating SNPs with cancer biology are limited. Herein, using chromatin immunoprecipitation assay (ChIP), reporter assay and chromosome conformation capture sequencing (3C-Seq) augmented with publically available genomic and epigenomic databases we aimed to define the function of rs6702619/1p21.2 region associated with CRC in the Polish population. Using ChIP we confirmed that rs6702619 region is occupied by a CTCF, a master regulator of long-range genomic interactions, and is decorated with enhancer-like histone modifications. The enhancer blocking assay revealed that rs6702619 region acts as an insulator with activity dependent on the SNP genotype. Finally, a 3C-Seq survey indicated more than a hundred *loci* in the rs6702619 *locus* interactome, including *GNAS* gene that is frequently amplified in CRC. Taken together, we showed that the CRC-associated rs6702619 region has *in vitro* and *in vivo* properties of an insulator that demonstrates long-range physical interactions with CRC-relevant *loci*.

Key words: genome-wide association study, SNP, CTCF, chromosome conformation capture, colorectal cancer, GNAS

Received: 01 February, 2019; revised: 08 August, 2019; accepted: 02 September, 2019; available on-line: 17 September, 2019

✉e-mail: mikula.michal@coi.pl
#equal contribution
Acknowledgement of Financial Support: This work was supported by grant from National Science Center [2012/05/N/NZ2/00602].
Abbreviations: 3C-Seq, chromosome conformation capture sequencing; CRC, colorectal cancer; eQTL, expression quantitative trait loci

INTRODUCTION

Cancer is a multigene disorder which develops as a consequence of a stepwise accumulation of genetic and epigenetic alterations leading to uncontrolled cell divisions. Germline mutations in DNA repair genes, proto-oncogenes or tumor suppressor genes can greatly increase the cancer risk. In case of colorectal cancer (CRC), only about 5% of cases are related to highly-penetrant mutations while twin studies estimate that 35% of all CRC cases are caused by heritable factors (Lichtenstein *et al.*, 2000). Majority of the remaining alterations are most likely less penetrant and represent single nucleotide polymorphisms (SNPs) that have additive effects (Peters *et al.*, 2012). The expression of the genome is a complex and multi-step process involving cis- and trans-acting regulatory elements (Pastinen *et al.*, 2006). It is assumed that majority of genetic variants influence the abundance of transcripts level mainly by altering transcription factors' binding at the promoter, pre-mRNAs splicing or through regions containing regulatory elements that are distal to the genes. Thus, they are known as expression quantitative trait loci (eQTLs). eQTLs can be located in the proximity of a gene of interest or in a distant genomic region and affect expression by long-range interactions (Identify regulatory sequences and eQTL-causal variants, and estimate their effects on activation of transcription in a massively parallel reporter assay | Critical Assessment of Genome Interpretation, 2015). Genome-wide association studies (GWAS) showed that over 85% of genotype-phenotype associations are non-coding single nucleotide polymorphisms (SNPs) (Buroker *et al.*, 2013). SNPs overlapping regulatory regions may play a significant role in the phenotypic variability and disease susceptibility mainly due to their effect on transcription. Functional analyses of regions indicated by those SNPs are prerequisite for understanding the molecular background of the observed association (Brown *et al.*, 2013; Fareed & Afzal, 2013).

Recently, we performed GWAS analysis identifying SNPs associating with CRC in the Polish population (Gaj *et al.*, 2012). One SNP pointed out in our study, namely rs6702619/1p21.2, seemed particularly interesting for further investigation due to the epigenetic features at its location including histone modification suggesting a presence of an enhancer and CTCF binding sites. CTCF plays an important role in the regulation of gene expression and higher-order organization of the genome (Kim *et al.*, 2015). The main role of CTCF is considered to be its contribution to the chromatin architecture *via* processes like nucleosome positioning, organization of chromatin modifications, demarcation of the boundaries of independently regulated domains, as well as of active and repressive chromatin domains (Phillips & Corces, 2009). Over 20,000 CTCF target sites (CTSs) have been reported across the genome in different cell types. They are located in introns, exons, promoters, 3' and 5'UTRs, however, almost a half correlates with intergenic regions. About 40–60% CTSs were shown to be constitutive and invariant between cell types while remaining are considered to be involved in tissue-specific gene expression (Cuddapah *et al.*,

2009). Although CTCF is able to bind numerous variant CTSs differing in length and sequence, the ~11–15 bp core consensus sequence was identified which is consistent in different cell types. The 'CTCF code' model was proposed where DNA sequence within and outside the consensus motifs determines the protein partners and, in effect, CTCF role. Alterations in CTSs can thus change the binding specificity of CTCF and influence its function (Ohlsson *et al.*, 2010).

To describe the influence of the rs6702619 SNP on the local chromatin structure and expression of nearby genes we analyzed gene expression of adjacent genes, histone modifications makeup and CTCF binding, as well as *in vitro cis*-regulatory activity of CTSs. Finally, we performed chromosome conformation capture (3C) sequencing to define local and long-range interactions of the regulatory element contacting the rs6702619 SNP.

MATERIALS AND METHODS

Cell lines. CRC cell lines were obtained from American Type Culture Collection (Rockville, MD, USA). All cell lines were cultured in media purchased from Sigma, St. Louis, Missouri, United States supplemented with fetal bovine serum (FBS; Sigma, St. Louis, Missouri, United States). The following media were used: RPMI-1640 medium, 10% and 5% FBS (Colo205 and SW480 cells respectively), Eagle's Minimum Essential Medium, 10% FBS (Caco2 cells), McCoy's 5A modified medium, supplemented with 10% FBS (HCT116 and HT-29 cells). Cell lines were maintained at 37°C in a humidified atmosphere containing 5% CO_2.

Cell lines genotyping. Cell lines were genotyped using ready-to-use TaqMan SNP Genotyping Assays (Thermo, USA), SensiMix™ II Probe Kit (Bioline Ltd, United Kingdom), and a 7900HT Real-Time PCR system (Thermo, USA) as described before (Gaj *et al.*, 2012).

Gene expression analysis. Total RNA was extracted using the TRI Reagent (Thermo, USA) and single-stranded cDNA was synthesized with the High Capacity cDNA Reverse Transcription Kit (Thermo, USA) according to the manufacturer's instructions. The expression of genes in the proximity of rs6702619 (LPPR4, PALMD) was measured by qRT-PCR in five CRC cell lines as described before (Mikula *et al.*, 2011; Maryan *et al.*, 2015). The geometric mean of YWHAZ, ALAS1, ACTB, TUBA1B and HPRT1 expression was used as the normalization factor in mRNA expression. Primers used are listed in Table S1.

ChIP assay. Cell harvesting and chromatin cross-linking were performed as described previously (Flanagin *et al.*, 2008; Naito *et al.*, 2009). Chromatin was fragmented in a Bioruptor (Diagenode, Philadelphia, PA) using the protocol 30-s on, 60-s off, 18 cycles at high intensity. Chromatin immunoprecipitation assays were performed using Matrix-ChIP platform as described before (Flanagin *et al.*, 2008; Naito *et al.*, 2009) with the following antibodies: IgG (Vector Laboratories, Inc.; I-1000), CTCF (Active motif; 61311), total H3 (Abcam; ab1791), H3K27Ac (Abcam; ab4729), H3K4m1 (Millipore; 07-436), H3K4m3 (Diagenode; pAb-003–050).

Enhancer-blocking assay. Sequence-specific insulator activity of CTSs was examined with specific reporter gene system, called enhancer-blocking assay as described by Lunyak et al. (Lunyak *et al.*, 2007). The mammalian expression vector pGL3-Control (Promega, USA) was used as a backbone for preparation of the gene constructs. Site-directed mutagenesis to introduce restriction endonuclease site (AgeI) was performed using the KOD Hot Start DNA Polymerase (Merck Millipore, USA) according to the manufacturer's instructions. β-Globin insulator (495 bp) and sequence chr1:100,045,980-100,046,473 (rs6702619, genotype –GG and –TT, amplified from HCT 116 and Colo205, respectively) were amplified from pBT268 expression vector (Addgene, Cambridge, USA) and the DNA isolated from cell lines, respectively, using primers containing overlapping ends. The sequences were inserted into pGL3-AgeI vector by digestion with a restriction enzyme (AgeI, NEB, USA) and ligation (T4 DNA ligase, NEB, USA). The mammalian expression vector pGL4.74 (Promega, USA) was used as an internal control for normalization.

HeLa cells were seeded into 96-well plate in DMEM with 10% FBS. The next day cells were co-transfected with the obtained constructs and pGL4.74 (ratio 100:1). After 24 h measurement of luminescence intensities was performed using the Dual-Luciferase Reporter Assay System (Promega, USA) according to the manufacturer's instructions. The differences in the expression levels were evaluated using the Student's t-test in GraphPad Prism 5 (GraphPad Software, Inc., USA). $P<0.05$ was considered to indicate a statistically significant difference.

Chromatin conformation capture (3C-seq) assay. 3C-seq library was obtained from HCT 116 cell line according to the protocol by Stadhouders et al. (Stadhouders *et al.*, 2013). EcoRI and HaeIII were used as the first and second restriction enzyme, respectively. The region surrounding chosen SNP was used as a bait for inverse PCR (Table S1 at https://ojs.ptbioch.edu.pl/index.php/abp/). Identification of DNA sequences interacting with the bait was performed by sequencing on Ion Torrent PGM (Thermo, USA). Two biological replicates were analyzed.

Analysis of 3C-seq assay results. The reads generated by sequencing were aligned to hg19 genome assembly with the TMAP. Peaks were defined as the regions with coverage greater than the median value for the chromosome. The intersection of peaks from biological replicates was performed with bedTools. Only intersections with length 100 or more were subjected to further analysis. The peaks' distance to the closest gene was assessed with bedTools. The frequency of mutations and copy number alterations was assessed with cBioPortal (Gao *et al.*, 2013), using TCGA Colorectoral Adenocarcinoma provisional dataset.

RESULTS

Description of rs6702619 SNP epigenetic landscape with publically available databases

The rs6702619 SNP is located within 335 kb long intergenic region at chromosome 1, between *LPPR4* (lipid phosphate phosphatase-related protein type 4) and *PALMD* (palmdelphin) genes located 270 kb upstream and 65 kb downstream, respectively. LPPR4 belongs to the lipid phosphate phosphatase family which catalyzes the dephosphorylation of lipid mediators while PALMD is a cytosolic homolog of PALM (paralemmin) which is implicated to influence plasma membrane dynamics (Pruitt *et al.*, 2012). rs6702619 lies within 110 kb long linkage disequilibrium block, confined by two sites with

recombination rate values of 30–40 cM/Mb (Fig. 1) (Johnson *et al.*, 2008). The sequence within which rs6702619 is located within a highly evolutionarily conserved DNA stretch. The histone modifications at this area, H3K4Me1 and H3K27Ac, identified in several cell lines within the frames of the ENCODE project are typical for the enhancer regulatory elements. Furthermore, CTCF binding was observed at this region by ChIP-Seq in normal human epidermal keratinocytes (NHEK) and human mammary epithelial cells (HMEC), as well as in several other cell lines. The observed signal is wide and covers about 350 bp (Rosenbloom *et al.*, 2012) (Fig. 2A). Additionally, an *in silico* analysis with InsulatorDB tool indicated that SNP is located between two predicted CTCF binding sites (CTSs) and thus may influence CTCF binding specificity (Bao *et al.*, 2008) (Fig. 2B). The evidence on the presence of both enhancer element and CTCF protein suggests that the latter one under a specific lineage context may act as an enhancer-blocking insulator.

Gene expression analyses of neighboring genes and survey of histone modifications and CTCF binding at rs6702619/1p21.2 in colorectal cancer (CRC) cell lines.

We performed genotyping of rs6702619 in five human CRC cell lines using TaqMan SNP Genotyping Assays. SW480, Caco2 and HT-29 cell lines are heterozygous for rs6702619 (GT), while HCT116 is a homozygote (GG), same as Colo205 (TT). The examination of expression levels of genes adjacent to rs6702619 (*LPPR4*, *PALMD*) showed no correlation with the genotype (Fig. 3).

Next, we determined the epigenetic makeup at the SNP site and control regions by measuring CTCF binding and histone modifications associated with enhancer regions (H3K4Me1, H3K27Ac), repressive (H3K27Me3) and permissive (H3K4Me3) chromatin using chromatin immunoprecipitation (ChIP) assay and three cell lines, namely HTC116, SW480, Colo205, representing all possible SNP genotypes. Control regions included a beta-globin (HBB) and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) promoters, representing silenced and active promoters, respectively, as well as a representative enhancer region (EGR1 -811 bp) and CTCF binding site (EGR1 -15 kb) both located upstream of the *EGR1* gene. ChIP assay confirmed CTCF binding at the control EGR1 -15 kb region and along the SNP locus probed at three sites (A, B, C) with the highest recruitment at the SNP site (B probe) in the three cell lines tested. However, the level of its recruitment did not depend on the genotype present in a given cell line (Fig. 4). At the same time, CTCF binding remained low at the promoters of HBB and GAPDH as well as at the control enhancer region EGR1 -811 bp. The H3K4me1 histone mark designates active and primed enhancers. These can be distinguished on the basis of the presence and absence of H3K27Ac mark, respectively (Creyghton *et al.*, 2010). ChIP analyses at the SNP site revealed no marked presence of H3K27Ac in HCT116 and SW480 cell lines, while in Colo205 this mark was significantly elevated (p-value <0.05) when compared to silenced HBB promoter. The H3K27Ac levels corresponded with H3K4me1 in Colo205 where this mark was relatively, but not significantly, enriched in comparison to inactive HBB promoter and the control enhancer region EGR1 -811 bp. The H3K4me3 histone modification is mostly found to be associated with promoter regions (Heintzman *et al.*, 2007), however, this mark may also decorate enhancer regions as a consequence of local RNAP2 presence (Pekowska *et al.*, 2011). ChIP assay indicated no H3K4me3 presence in HCT116 and SW480 cell lines, while in Colo205 this modification at SNP site was significantly elevated (p-value <0.05) at A site when compared to silenced HBB promoter. As expected, the highest level of H3K4me3 mark was present at active GAPDH promoter in all cell lines. In sum, ChIP measurements confirmed CTCF binding to the SNP site in tested cell lines and suggested that this *cis*-regulatory element could act as an enhancer-blocking insulator. This possibility was subsequently tested using a reporter system.

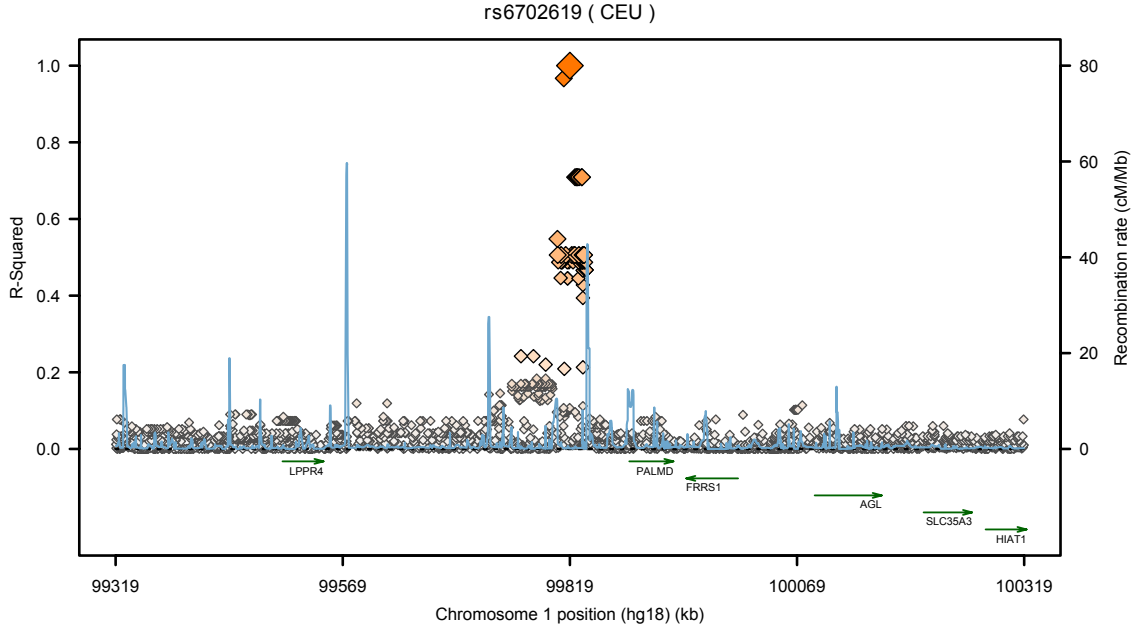

Figure 1. Linkage disequilibrium plot of 500 kb long region of chromosome 1 containing rs6702619.
On the left axis, pair-wise r^2 values between rs6702619 and SNPs within the region are shown. Plot generated with SNAP tool (Johnson *et al.*, 2008).

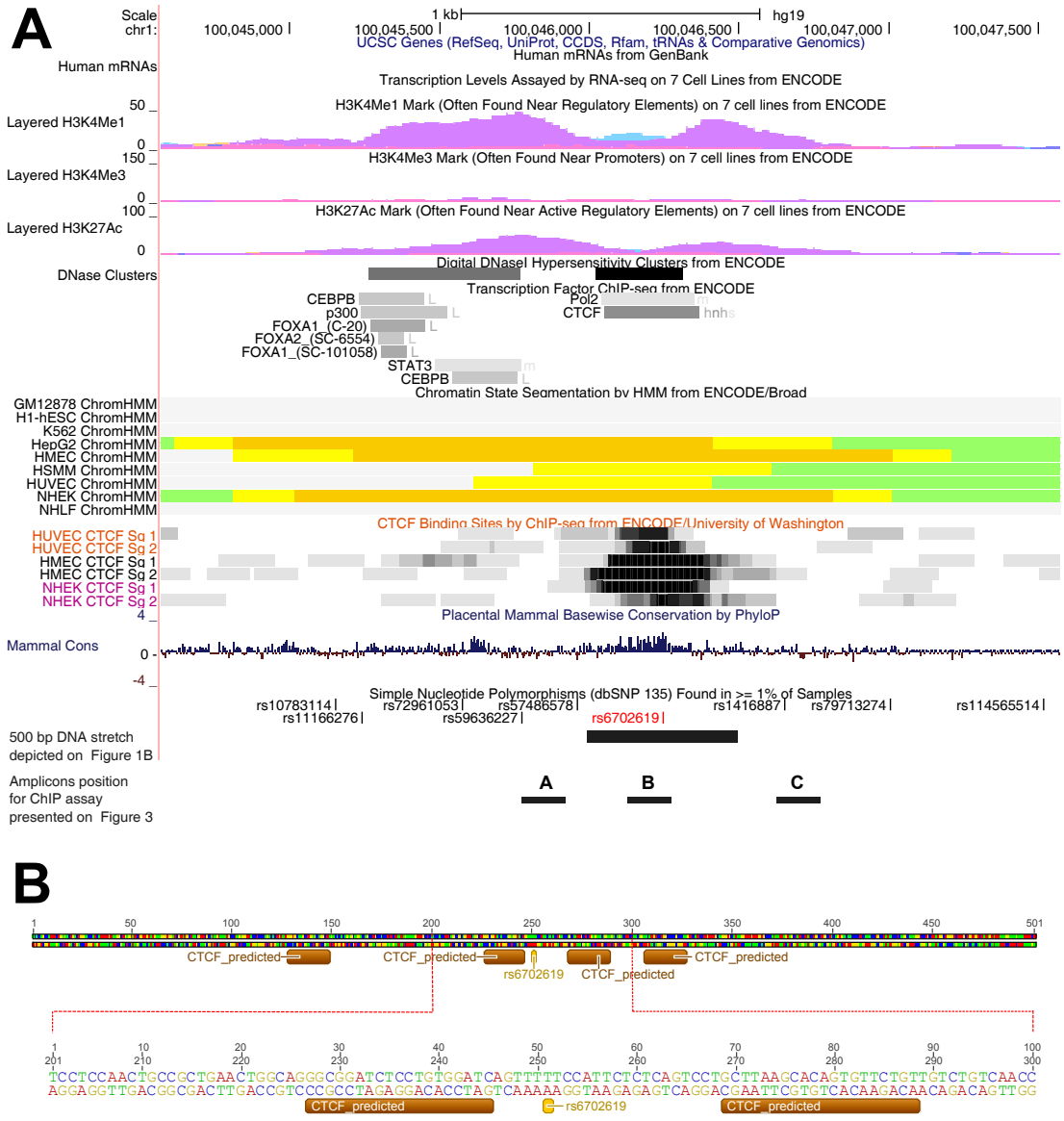

Figure 2. (A) Chromatin features in the proximity of rs6702619 suggest the presence of an enhancer-blocking insulator. Figure generated with UCSC Genome Browser (chr1:100,044,569-100,047,573; version hg19) (B) CTCF binding sites in the proximity of rs6702619 at chromosome 1 predicted with InsulatorDB tool (Bao *et al*., 2008) (chr1:100,045,996-100,046,496; version hg19).

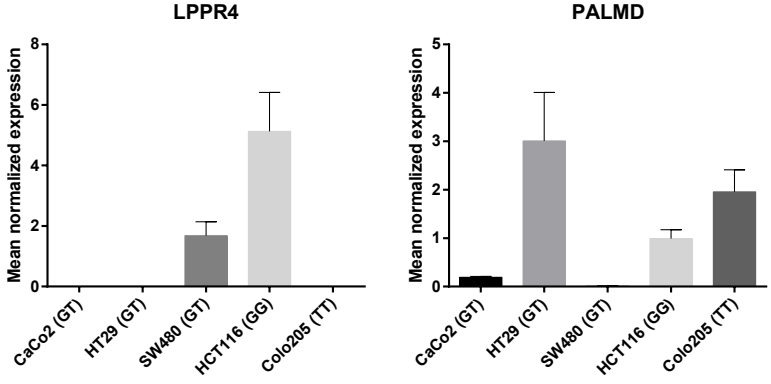

Figure 3. Expression of *LPPR4* and *PALMD* genes adjacent to the rs6702619 SNP was measured in human colon cancer cell lines with qPCR using SYBR Green chemistry.
Cells were harvested, RNA extracted with Trizol, DNAse-treated and subjected to RT-qPCR measurements. RNA expression was normalized to YWHAZ, ALAS1, ACTB, TUBA1B and HPRT1 mRNA (n=3; mean ±SD).

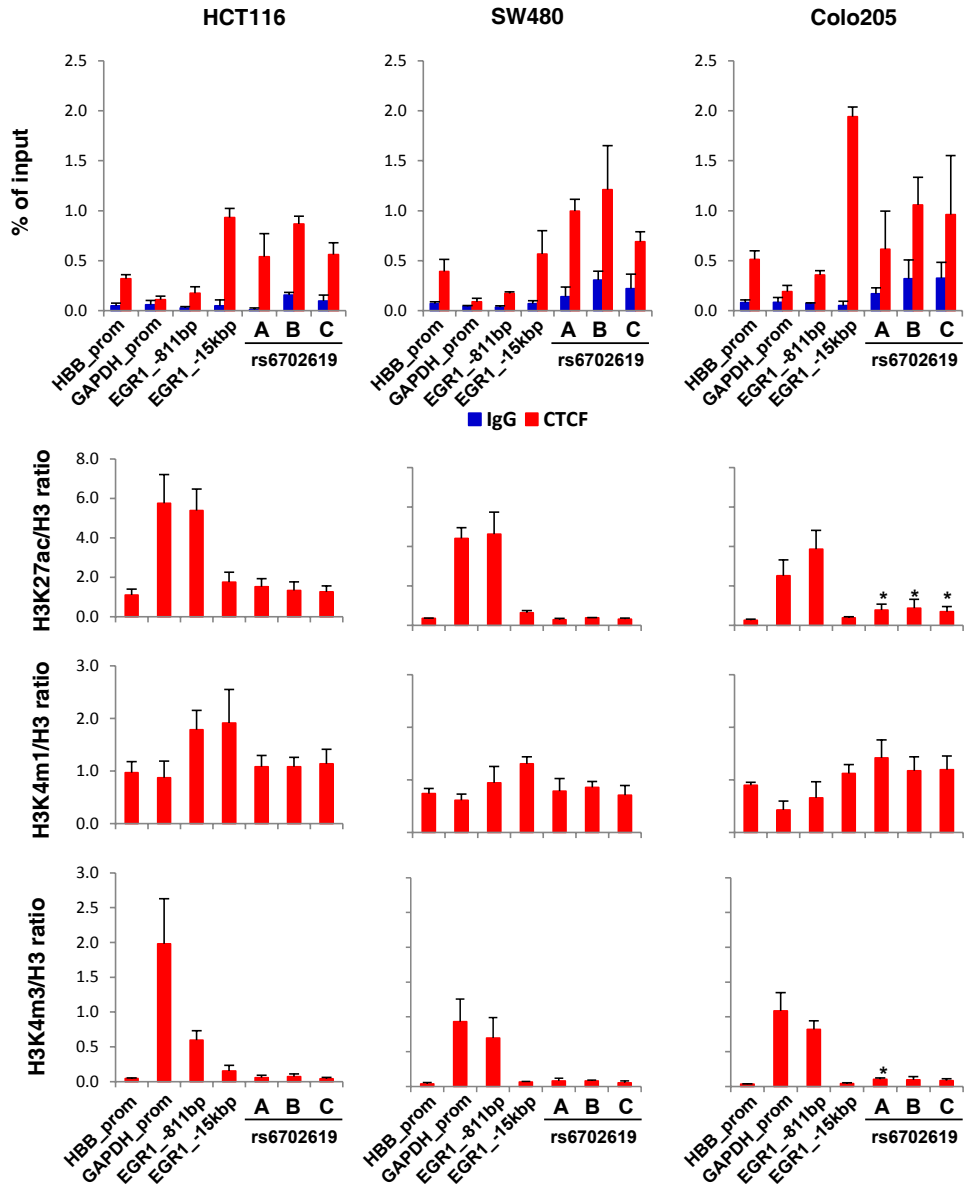

Figure 4. ChIP assay of CTCF and histone H3 modifications in the proximity of rs6702619 site and control regions.
Cross-linked chromatin was sheared and ChIP assay was performed using a Matrix-ChIP platform. ChIP DNA was analyzed in qPCR with primers amplifying rs6702619 site at region A, B, C (see Fig. 2 for amplicons location), beta-globin (*HBB*) , glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) promoters, as well as a control enhancer region (*EGR1* -811 bp) and CTCF binding site (*EGR1* -15kb) both located upstream of the *EGR1* gene. Data are presented as the means ± S.D. (n=3), expressed either as the percentage (%) of input DNA for CTCF or as the ratio of modified histone to total H3 histone. Differences in binding levels between rs6702619 A, B, C regions and *HBB* promoter were evaluated using the Student's *t*-test – *P<0.05.

The region harboring rs6702619 SNP acts as a genotype-dependent insulator in an enhancer-blocking assay

The sequence-specific insulator activity of the predicted CTCF binding site in the proximity of rs6702619 was measured with the enhancer-blocking assay. The DNA sequences containing β-globin insulator as well as the rs6702619 derived from HCT116 (G/G) and Colo205 (T/T) which differ in their sequence at the SNP site only, were inserted between the SV40 enhancer and promoter elements of the pGL3-control vector. As expected, the chicken β-globin insulator, used as a positive control, significantly inhibited the luciferase expression by 62% (Fig. 5). For the rs6702619 site with G/G and T/T genotype, the luciferase promoter activity was significantly decreased by 50% and 17%, respectively. Additionally, there was a significant difference in luciferase expression between the rs6702619 constructs containing G/G and T/T genotype. These data confirm that the rs6702619 site *in vitro* acts as an insulator and that its activity is genotype-dependent.

Portraying genomic interactions of rs6702619/1p21.2 locus with 3C-Seq

In order to define long-range chromatin interactions for rs6702619 *locus,* we performed 3C-Seq analyses using two biological replicates of HCT-116 cell line. Two replicates yielded 3,927,047 and 3,291,544 reads mapped to human genome version hg19, respectively. Overall, 122 interacting common *loci* for both

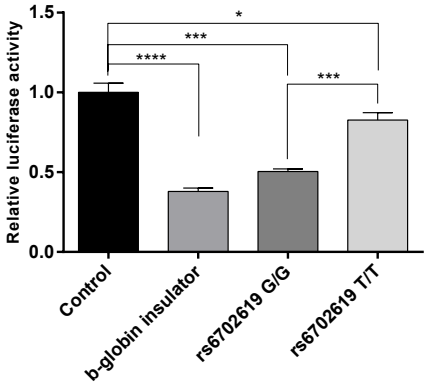

Figure 5. Enhancer-blocking assay.
Relative luciferase activity of promoter constructs in pGL3 plasmid was measured when cotransfected with pGL4.74 plasmid into HeLa cells with an empty vector, chicken β-globin insulator and the rs6702619 site with G/G or T/T genotype containing construct. All bars show the average of four independent experiments, error bars indicate standard deviation. Differences in expression levels were evaluated using the Student's *t*-test – *P<0.05, ***P<0.001 ****P<0.0001

replicates were identified, of which 65 were located at chromosome 1 as the rs6702619 bait (Fig. 6A, Table S2 at https://ojs.ptbioch.edu.pl/index.php/abp/). The majority of 3C-seq sequences originating from the bait chromosome is expected as this technical bias is common for 3C methods (Raviram *et al.*, 2014). 49 peaks were found within the gene bodies and 20 of them were not located on chromosome 1 (Table S2 at https://ojs.ptbioch.edu.pl/index.php/abp/). To further portray the connection of rs6702619 *locus* and 20 genes inter-chromosomally interacting with CRC we annotated them using TCGA Colorectal Adenocarcinoma dataset and cBioPortal (Gao *et al.*, 2013). Interestingly, 14 of these genes exhibited a genetic lesion, either mutation or copy number alteration (CNA), in 26% of 640 samples from this dataset (Fig. 6B). Based on the data from 1000 random draws of 20 genes, the probability of obtaining this result is as small as 6.9e-17. The most frequently altered gene was *GNAS* (20q13.32) for which the amplification event was present in 8.44% of patients in TCGA dataset. *GNAS* activating mutations were identified in multiple cancers including CRC (Wilson *et al.*, 2010). Inspection of the sequencing data at *GNAS* locus in UCSC browser revealed 3C-Seq peak presence in the intron sequence of the first and the third *GNAS* isoform as well as in the intron of *GNAS* isoform coding its antisense transcript (*GNAS-As*) (Figure 6C). Additionally, the 3C-Seq peak overlapped with CTCF binding sites as measured for K562 and HeLa cell lines in the ENCODE project, suggesting that this long-range interaction could be mediated by CTCF. We next measured *GNAS* and *GNAS-As* expression in five CRC cell lines using qRT-PCR. We observed no association between rs6702619 genotype and GNAS mRNA abundance. However, such a connection was visible for GNAS-As mRNA (Fig. 7) where its transcript level was decreased and elevated in HCT116 and Colo205, respectively, which is in line with the activity of the G/G and T/T allele in the reporter system. Overall, the 3C-seq data and functional analyses showed that the rs6702619 region's interchromosomal gene network is enriched with *loci* that are relevant for CRC tumorigenesis.

DISCUSSION

To date, at least 38 GWASs studies reported 310 SNPs associating with CRC for different populations (MacArthur *et al.*, 2017), however, the biological understanding and their functional contribution to CRC development remain unexplained for most of them. GWAS yields information about the statistical association between observed phenotype and tagging SNPs. To uncover the biological reason for observed association functional analyses of the whole LD block indicated by tagging SNP have to be performed. Most of the tagging SNPs are located in the intergenic or intronic regions which makes the task daunting and introduces the need for multistage analysis where the sequence and type of experiments depend on the structure of the analyzed region (Freedman *et al.*, 2011). The biochemical characterization of chromatin within the frames of the ENCODE project defined a plethora of cell-type-specific distal regulatory regions including enhancers and insulators (ENCODE Project Consortium, 2012). Further collation of GWAS SNPs to these regulatory regions has revealed that many SNPs fall within these regions having specific chromatin's biochemical makeup, including distinct histone modifications, open chromatin signatures and binding sites for transcription factors such as CTCF (Farnham, 2012). The ENCODE datasets, therefore, expanded the possibilities of rational SNP choosing for functional studies prioritizing SNP in high LD with GWAS SNPs that are located in a regulatory element. This approach combined with CRISPR/Cas technology allows studying the influence of CRC risk-associated SNPs and regulatory regions on gene expression (Yao *et al.*, 2014).

In CRC, functional analyses of SNPs were described for several *loci*, however only for one of them, rs6983267 in 8q24 region, the functional variant and the mechanism of its influence was confirmed by several independent studies (Pomerantz *et al.*, 2009; Tuupanen *et al.*, 2009; Sotelo *et al.*, 2010; Wright *et al.*, 2010). The rs6983267 is located in a 1.5 Mb long gene-desert and molecular analyses showed that SNP lies within a region highly conserved in mammals, in a consensus sequence for transcription factor TCF4. ChIP, electrophoretic mobility shift (EMSA) and reporter assays confirmed the activity of computationally identified enhancer and demonstrated TCF4 differential binding depending on SNP allele (Pomerantz *et al.*, 2009; Tuupanen *et al.*, 2009; Sotelo *et al.*, 2010). Spatial chromatin structure analyses, like 3C allowed to identify the interaction between the abovementioned enhancer and *MYC* promoter, located 335 kb apart (Wright *et al.*, 2010). Recently, through the deletion of rs6983267 bearing enhancer with CRISPR/Cas followed by gene expression analyses a set of down-regulated transcripts, including MYC, was identified in HCT116 CRC cell line (Yao *et al.*, 2014).

In this study, we used publically available epigenetic and genomic datasets together with ChIP, reporter assay and 3C-Seq experiments to characterize the structure and function of the genomic region tagged by rs6702619 SNP associated with adenoma and CRC in the Polish population, with odds ratio 0.71 and 0.73, respectively (Gaj *et al.*, 2012). However, to date, the SNP rs6702619 has not been confirmed to be associated with CRC in other populations. Interestingly, the rs6702619 was recently found to associate with cardiac structure, specifically with aortic root diameter (Wild *et al.*, 2017). Furthermore, in a study utilizing UK biobank specimens the rs6702619 significantly associated with calcific aortic

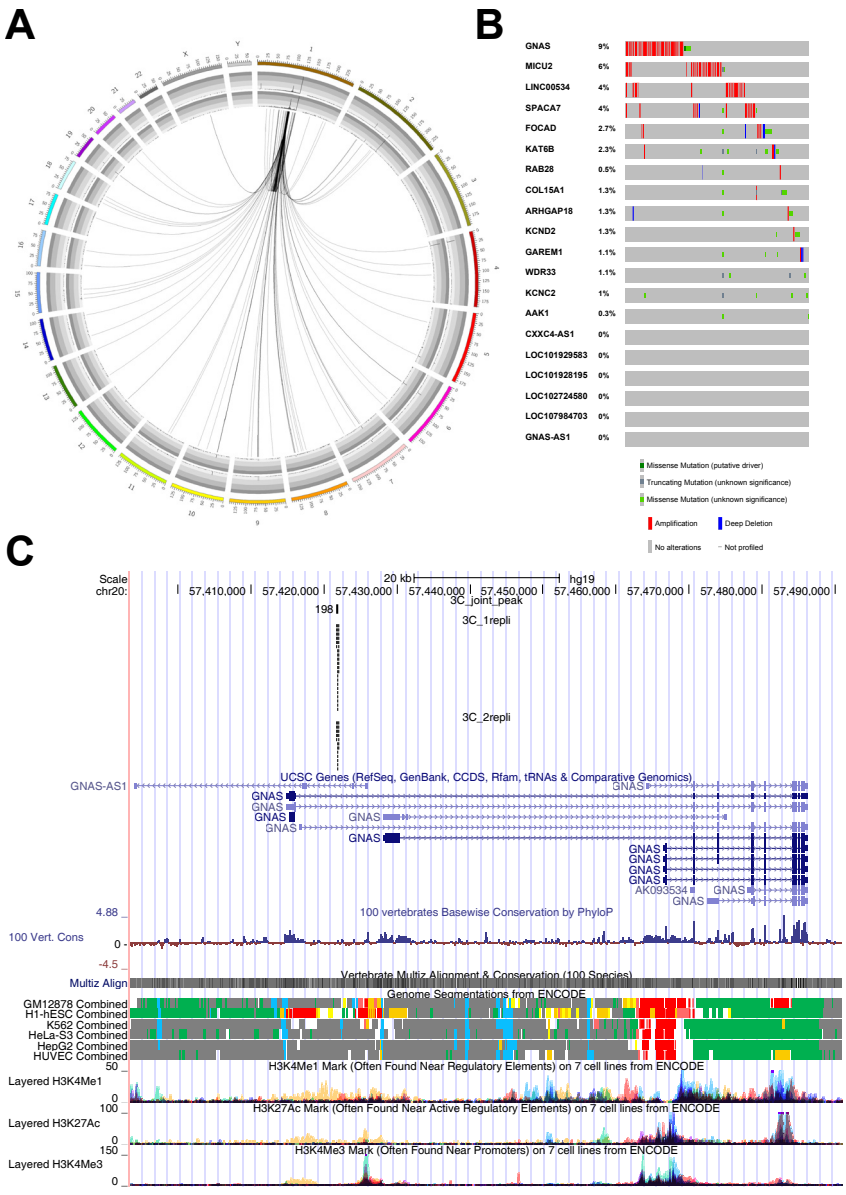

Figure 6. (A) Summary of 3C-Seq long-range interactions between rs6702619 region and distant regions. Lines represent peaks present in the two biological replicates (peaks in grey circles). (B) Genetic annotation of rs6702619 site inter-chromosomal interacting genes with TCGA Colorectal Adenocarcinoma dataset. Data fetched and visualized with cBioPortal. (C) Overview of *GNAS* locus in UCSC genome browser.

Presence of 3C-Seq reads confirms the interaction of *GNAS* locus with rs6702619 region. Genome Segmentations and Integrated Regulation ENCODE annotation tracks were included to portray the biochemical chromatin makeup in the vicinity of the peak. Bright Red, predicted promoter region; Light Red, predicted promoter flanking region; Orange, predicted enhancer; Yellow, predicted weak enhancer or open chromatin; Blue, CTCF enriched element; Dark Green, predicted transcribed region; Gray, predicted repressed or low activity region

valve stenosis (CAVS) and the G risk allele correlated with lower expression of *PALMD* (Thériault *et al.*, 2018). Chromatin features from the ENCODE project at SNP rs6702619 suggested the presence of a CTCF-occupied insulator with a chromatin makeup of enhancer regulatory element. Analysis of *LPPR4*, *PALMD* gene expression in the vicinity of this regulatory element in several CRC cell lines showed no correlation between rs6702619 genotype and transcripts abundances in the studied cell lines. Furthermore, rs6702619 loci characterization with ChIP assay showed CTCF recruitment to this regulatory element in three CRC cell lines tested and indicated the presence of enhancer-like chromatin features in one of them at this site. The enhancer blocking assay with rs6702619 region confirmed that this genomic stretch could act as an insulator with activity dependent on the rs6702619 genotype. Finally, a 3C-Seq survey indicated more than a hundred *loci* in the rs6702619 locus interactome. Of them, 20 were inter-chromosomal connections to the genes. Importantly, one of the interacting genes was *GNAS* which is regarded as an oncogene contributing to the development of several neoplasms including pituitary, thyroid glands, pancreas and colon tumors (O'Hayre *et al.*, 2013). Of note, the mRNA expression analyses in CRC indicated a putative connection of the rs6702619 with GNAS-As mRNA abundance, specifically for the homozygotic genotype. However, this result should be interpreted with caution since it was obtained for only two cell lines and has no statistical support. In CRC, the *GNAS* locus is frequently amplified which correlates with its increased mRNA expression (Ptashkin *et al.*, 2017). Importantly, we previously observed a sig-

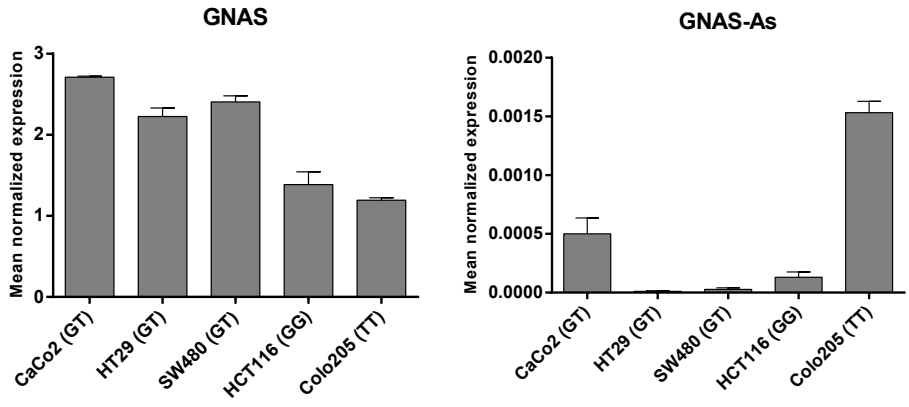

Figure 7. The expression of GNAS and GNAS-As mRNA measured in five human colorectal cancer cell lines.
Cells were harvested, RNA extracted with Trizol, DNAse-treated and subjected to RT-qPCR measurements with SYBR Green chemistry. The genotype at the rs6702619 SNP is indicated along with the cell line name. RNA expression was normalized to YWHAZ, ALAS1, ACTB, TUBA1B and HPRT1 mRNA (n=3; mean ± S.D.).

nificant 2.4 fold GNAS overexpression in CRC when compared to normal mucosa in our transcriptome study (Skrzypczak *et al.*, 2010), this result was also confirmed in a microarray data from a study by Hong et al. (Hong *et al.*, 2010). *GNAS* encodes α-subunit of the Gs stimulatory protein and it was proposed that gain of function mutations in *GNAS*, which result in constitutive activity by reducing the rate of GTP hydrolysis, or its overexpression may influence pro-inflammatory gene expression and fuel tumor development (O'Hayre *et al.*, 2013). Interestingly, the long-range interaction with *GNAS* locus could be CTCF-mediated since its ChIP-Seq binding signal, detected for several cell lines in the ENCODE project, is adjacent to 3C peaks from rs6702619 bait.

A limitation of our study is that we did not test the outcome of a given SNP variant in the same genetic background following CRISPR/Cas rs6702619 edition. Further experiments utilizing this approach are warranted as they could uncover *bona fide* molecular changes driven by the status of rs6702619. Nevertheless, we demonstrated that the CRC-associating locus containing rs6702619 has the properties of an insulator that exhibits multiple long-range physical interactions with CRC-relevant *loci* including *GNAS*.

Data Availability

Sequencing data are available for view and download in UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=prutten&hgS_otherUserSessionName=rs6702619_3C%2DSeq_2replicates

Author contributions

MM and NM conceived and designed the study, NM, MS, UK and MK acquired the data, MM and MK analyzed and interpreted the data, MS, MM and JO wrote the paper.

REFERENCES

Bao L, Zhou M, Cui Y (2008) CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res* **36**: D83–D87. https://doi.org/10.1093/nar/gkm875.

Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLOS Genet* **9**: e1003649. https://doi.org/10.1371/journal.pgen.1003649.

Buroker NE, Ning X, Zhou Z, Li K, Cen W, Wu X, Zhu W, Ronald Scott C, Chen S (2013) SNPs and TFBS associated with high altitude sickness*. *Open J Blood Dis* **03**: 85–93. https://doi.org/10.4236/ojbd.2013.33018.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931–21936. https://doi.org/10.1073/pnas.1016071107.

Cuddapah S, Jothi R, Schones DE, Roh T-Y, Cui K, Zhao K (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**: 24–32. https://doi.org/10.1101/gr.082800.108.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. https://doi.org/10.1038/nature11247.

Fareed M, Afzal M (2013) Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service. *Egypt J Med Hum Genet* **14**: 123–134. https://doi.org/10.1016/j.ejmhg.2012.08.001.

Farnham PJ (2012) Thematic minireview series on results from the ENCODE Project: Integrative global analyses of regulatory regions in the human genome. *J Biol Chem* **287**: 30885–30887. https://doi.org/10.1074/jbc.R112.365940.

Flanagin S, Nelson JD, Castner DG, Denisenko O, Bomsztyk K (2008) Microplate-based chromatin immunoprecipitation method, Matrix ChIP: a platform to study signaling of complex genomic events. *Nucleic Acids Res* **36**: e17. https://doi.org/10.1093/nar/gkn001.

Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Vikis HG, You M, Mills IG (2011) Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet* **43**: 513–518. https://doi.org/10.1038/ng.840.

Gaj P, Maryan N, Hennig EE, Ledwon JK, Paziewska A, Majewska A, Karczmarski J, Nesteruk M, Wolski J, Antoniewicz AA, Przytulski K, Rutkowski A, Teumer A, Homuth G, Starzyńska T, Regula J, Ostrowski J (2012) Pooled sample-based GWAS: a cost-effective alternative for identifying colorectal and prostate cancer risk variants in the Polish population. *PloS One* **7**: e35307. https://doi.org/10.1371/journal.pone.0035307.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**: pl1. https://doi.org/10.1126/scisignal.2004088.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**: 311–318. https://doi.org/10.1038/ng1966.

Hong Y, Downey T, Eu KW, Koh PK, Cheah PY (2010) A 'metastasis-prone' signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics. *Clin Exp Metastasis* **27**: 83–90. https://doi.org/10.1007/s10585-010-9305-4.

Identify regulatory sequences and eQTL-causal variants, and estimate their effects on activation of transcription in a massively parallel reporter assay | Critical Assessment of Genome Interpretation (2015) https://genomeinterpretation.org/content/4-eQTL-causal_SNPs (accessed: 01/06/2016).

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**: 2938–2939. https://doi.org/10.1093/bioinformatics/btn564.

Kim S, Yu N-K, Kaang B-K (2015) CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med* **47**: e166. https://doi.org/10.1038/emm.2015.33.

Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K (2000) Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**: 78–85. https://doi.org/10.1056/NEJM200007133430201.

Lunyak VV, Prefontaine GG, Núñez E, Cramer T, Ju B-G, Ohgi KA, Hutt K, Roy R, García-Díaz A, Zhu X, Yung Y, Montoliu L, Glass CK, Rosenfeld MG (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317**: 248–251. https://doi.org/10.1126/science.1140871.

MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**: D896–D901. https://doi.org/10.1093/nar/gkw1133.

Maryan N, Statkiewicz M, Mikula M, Goryca K, Paziewska A, Strzałkowska A, Dabrowska M, Bujko M, Ostrowski J (2015) Regulation of the expression of claudin 23 by the enhancer of zeste 2 polycomb group protein in colorectal cancer. *Mol Med Rep* **12**: 728–736. https://doi.org/10.3892/mmr.2015.3378.

Mikula M, Rubel T, Karczmarski J, Goryca K, Dadlez M, Ostrowski J (2011) Integrating proteomic and transcriptomic high-throughput surveys for search of new biomarkers of colon tumors. *Funct Integr Genomics* **11**: 215–224. https://doi.org/10.1007/s10142-010-0200-5.

Naito M, Zager RA, Bomsztyk K (2009) BRG1 increases transcription of proinflammatory genes in renal ischemia. *J Am Soc Nephrol JASN* **20**: 1787–1796. https://doi.org/10.1681/ASN.2009010118.

O'Hayre M, Vázquez-Prado J, Kufareva I, Stawiski EW, Handel TM, Seshagiri S, Gutkind JS (2013) The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat Rev Cancer* **13**: 412–424. https://doi.org/10.1038/nrc3521.

Ohlsson R, Lobanenkov V, Klenova E (2010) Does CTCF mediate between nuclear organization and gene expression? *BioEssays News Rev Mol Cell Dev Biol* **32**: 37–50. https://doi.org/10.1002/bies.200900118.

Pastinen T, Ge B, Hudson TJ (2006) Influence of human genome polymorphism on gene expression. *Hum Mol Genet* **15**: R9–R16. https://doi.org/10.1093/hmg/ddl044.

Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, Imbert J, Andrau J-C, Ferrier P, Spicuglia S (2011) H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J* **30**: 4198–4210. https://doi.org/10.1038/emboj.2011.295.

Peters U, Hutter CM, Hsu L, Schumacher FR, Conti DV, Carlson CS, Edlund CK, Haile RW, Gallinger S, Zanke BW, Lemire M, Rangrej J, Vijayaraghavan R, Chan AT, Hazra A, Hunter DJ, Ma J, Fuchs CS, Giovannucci EL, Kraft P, Liu Y, Chen L, Jiao S, Makar KW, Taverna D, Gruber SB, Rennert G, Moreno V, Ulrich CM, Woods MO, Green RC, Parfrey PS, Prentice RL, Kooperberg C, Jackson RD, Lacroix AZ, Caan BJ, Hayes RB, Berndt SI, Chanock SJ, Schoen RE, Chang-Claude J, Hoffmeister M, Brenner H, Frank B, Bézieau S, Küry S, Slattery ML, Hopper JL, Jenkins MA, Le Marchand L, Lindor NM, Newcomb PA, Seminara D, Hudson TJ, Duggan DJ, Potter JD, Casey G (2012) Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* **131**: 217–234. https://doi.org/10.1007/s00439-011-1055-0.

Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. *Cell* **137**: 1194–1211. https://doi.org/10.1016/j.cell.2009.06.001.

Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz H-J, Haiman CA, Yan C, Henderson BE, Frenkel B, Barretina J, Bass A, Tabernero J, Baselga J, Regan MM, Manak JR, Shivdasani R, Coetzee GA, Freedman ML (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**: 882–884. https://doi.org/10.1038/ng.403.

Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res* **40**: D130–D135. https://doi.org/10.1093/nar/gkr1079.

Ptashkin RN, Pagan C, Yaeger R, Middha S, Shia J, O'Rourke KP, Berger MF, Wang L, Cimera R, Wang J, Klimstra DS, Saltz L, Ladanyi M, Zehir A, Hechtman JF (2017) Chromosome 20q amplification defines a subtype of microsatellite stable, left-sided colon cancers with wild-type RAS/RAF and better overall survival. *Mol Cancer Res MCR* **15**: 708–713. https://doi.org/10.1158/1541-7786.MCR-16-0352.

Raviram R, Rocha PP, Bonneau R, Skok JA (2014) Interpreting 4C-Seq data: how far can we go? *Epigenomics* **6**: 455–457. https://doi.org/10.2217/epi.14.47.

Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, Cline MS, Karolchik D, Barber GP, Clawson H, Diekhans M, Fujita PA, Goldman M, Gravell RC, Harte RA, Hinrichs AS, Kirkup VM, Kuhn RM, Learned K, Maddren M, Meyer LR, Pohl A, Rhead B, Wong MC, Zweig AS, Haussler D, Kent WJ (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res* **40**: D912–D917. https://doi.org/10.1093/nar/gkr1012.

Skrzypczak M, Goryca K, Rubel T, Paziewska A, Mikula M, Jarosz D, Pachlewski J, Oledzki J, Ostrowski J, Ostrowsk J (2010) Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability. *PloS One* **5**: https://doi.org/10.1371/journal.pone.0013091.

Sotelo J, Esposito D, Duhagon MA, Banfield K, Mehalko J, Liao H, Stephens RM, Harris TJR, Munroe DJ, Wu X (2010) Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A* **107**: 3001–3005. https://doi.org/10.1073/pnas.0906067107.

Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, Kockx C, Palstra R-J, Wendt KS, Grosveld F, van Ijcken W, Soler E (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat Protoc* **8**: 509–524. https://doi.org/10.1038/nprot.2013.018.

Thériault S, Gaudreault N, Lamontagne M, Rosa M, Boulanger M-C, Messika-Zeitoun D, Clavel M-A, Capoulade R, Dagenais F, Pibarot P, Mathieu P, Bossé Y (2018) A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. *Nat Commun* **9**: 988. https://doi.org/10.1038/s41467-018-03260-6.

Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Björklund M, Wei G, Yan J, Niittymäki I, Mecklin J-P, Järvinen H, Ristimäki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* **41**: 885–890. https://doi.org/10.1038/ng.406.

Wild PS, Felix JF, Schillert A, Teumer A, Chen M-H, Leening MJG, Völker U, Großmann V, Brody JA, Irvin MR, Shah SJ, Pramana S, Lieb W, Schmidt R, Stanton AV, Malzahn D, Smith AV, Sundström J, Minelli C, Ruggiero D, Lyytikäinen L-P, Tiller D, Smith JG, Monnereau C, Di Tullio MR, Musani SK, Morrison AC, Pers TH, Morley M, Kleber ME, Aragam J, Benjamin EJ, Bis JC, Bisping E, Broeckel U, Cheng S, Deckers JW, Del Greco M F, Edelmann F, Fornage M, Franke L, Friedrich N, Harris TB, Hofer E, Hofman A, Huang J, Hughes AD, Kähönen M, Investigators K, Kruppa J, Lackner KJ, Lannfelt L, Laskowski R, Launer LJ, Leosdottir M, Lin H, Lindgren CM, Loley C, MacRae CA, Mascalzoni D, Mayet J, Medenwald D, Morris AP, Müller C, Müller-Nurasyid M, Nappo S, Nilsson PM, Nuding S, Nutile T, Peters A, Pfeufer A, Pietzner D, Pramstaller PP, Raitakari OT, Rice KM, Rivadeneira F, Rotter JI, Ruohonen ST, Sacco RL, Samdarshi TE, Schmidt H, Sharp ASP, Shields DC, Sorice R, Sotoodehnia N, Stricker BH, Surendran P, Thom S, Töglhofer AM, Uitterlinden AG, Wachter R, Völzke H, Ziegler A, Münzel T, März W, Cappola TP, Hirschhorn JN, Mitchell GF, Smith NL, Fox ER, Dueker ND, Jaddoe VWV, Melander O, Russ M, Lehtimäki T, Ciullo M, Hicks AA, Lind L, Gudnason V, Pieske B, Barron AJ, Zweiker R, Schunkert H, Ingelsson E, Liu K, Arnett DK, Psaty BM, Blankenberg S, Larson MG, Felix SB, Franco OH, Zeller T, Vasan RS, Dörr M (2017) Large-scale genome-wide analysis identifies genetic variants associated with cardiac structure and function. *J Clin Invest* **127**: 1798–1812. https://doi.org/10.1172/JCI84840.

Wilson CH, McIntyre RE, Arends MJ, Adams DJ (2010) The activating mutation R201C in *GNAS* promotes intestinal tumourigenesis in $Apc^{Min/+}$ mice through activation of Wnt and ERK1/2 MAPK pathways. *Oncogene* **29**: 4567–4575. https://doi.org/10.1038/onc.2010.202.

Wright JB, Brown SJ, Cole MD (2010) Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol* **30**: 1411–1420. https://doi.org/10.1128/MCB.01384-09.

Yao L, Tak YG, Berman BP, Farnham PJ (2014) Functional annotation of colon cancer risk SNPs. *Nat Commun* **5**: https://doi.org/10.1038/ncomms6114.