Differential genetic influences over colorectal cancer risk and gene expression in large bowel mucosa

Timofeeva Maria (Orcid ID: 0000-0002-2503-4253)
Vaughan-Shaw Peter (Orcid ID: 0000-0002-9790-6882)

Differential genetic influences over colorectal cancer risk and gene expression in large bowel mucosa

Authors: *Peter. G. Vaughan-Shaw[1,2], *Maria Timofeeva[1,2,4], Li-Yin Ooi[3], Victoria Svinti[1,2], Graeme Grimes[1], Claire Smillie[1,2], James P. Blackmur[1,2], Kevin Donnelly[1,2], Evi Theodoratou[2,5], Harry Campbell[2,5], Lina Zgaga[6], Farhat V.N. Din[1,2], ≠Susan M. Farrington[1,2], ≠Malcolm G. Dunlop[1,2]

*≠Joint authors at these positions

Short Title: *Topographical differences in colorectal cancer risk loci eQTLs*

Affiliations:

[1] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom.

[2] Cancer Research UK Edinburgh Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, United Kingdom.

[3] Department of Pathology, National University Hospital, Singapore.

[4] D-IAS, Danish Institute for Advanced Study, Department of Public Health, University of Southern Denmark, Odense, Denmark.

[5] Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, United Kingdom.

[6] Department of Public Health and Primary Care, Trinity College Dublin, Dublin 24, Republic of Ireland.

Corresponding Author:

Professor MG Dunlop

Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital Edinburgh, Crewe Road, Edinburgh, EH4 2XU, United Kingdom

T: +44 (0)131 651 8500

E: malcolm.dunlop@igmm.ed.ac.uk

Twitter: @ccgg_edinburgh

Article Type Research article

Keywords single-nucleotide polymorphism; colorectal cancer; eQTL

Abbreviations

CRC - colorectal cancer

eQTL - Expression quantitative trait loci

Gorilla - Gene Ontology enRIchment anaLysis and visuaLizAtion tool

GTEx – Genotype-Tissue Expression (GTEx) project

NM - Normal colorectal mucosa

OR – odds ratio

PEER - probabilistic estimation of expression residuals

QC – quality control

qRT-PCR – quantitative reverse transcriptase polymerase chain reaction

RNAseq – RNA sequencing

SCOVIDS - Scottish Vitamin D study

SNP – Single nucleotide polymorphism

SOCCS - Study of Colorectal Cancer in Scotland

Novelty and Impact

We explored whether common genetic variants influencing colorectal cancer (CRC) risk exhibit topographical differences on risk through regional differences in effects on gene expression.

Genotype at the chr11q23.1 CRC risk locus (rs3087967) imparts site-specific risk of CRC, while site-specific trans-eQTL effects are seen with this locus and expression of four genes.

Findings provide novel insight into topographical differences in genomic control over gene expression relevant to CRC risk. These results may inform individualised CRC screening programmes.

Abstract

Site-specific variation in colorectal cancer (CRC) incidence, biology and prognosis are poorly understood. We sought to determine whether common genetic variants influencing CRC risk might exhibit topographical differences on CRC risk through regional differences in effects on gene expression in the large bowel mucosa.

We conducted a site-specific genetic association study (10,630 cases, 31,331 controls) to identify whether established risk variants exert differential effects on risk of proximal, compared to distal CRC. We collected normal colorectal mucosa and blood from 481 subjects and assessed mucosal gene expression using Illumina HumanHT-12v4 arrays in relation to germline genotype. Expression quantitative trait loci (eQTLs) were explored by anatomical location of sampling.

The rs3087967 genotype (chr11q23.1 risk variant) exhibited significant site-specific effects - risk of distal CRC (OR=1.20, $P=8.20\times10^{-20}$) with negligible effects on proximal CRC risk (OR=1.05, P=0.10). Expression of 1261 genes differed between proximal and distal colonic mucosa (top hit *PRAC* gene, fold-difference=10, $P=3.48\times10^{-57}$). In eQTL studies, rs3087967 genotype was associated with expression of 8 cis- and 21 trans-genes. Four of these (*AKAP14*, *ADH5P4*, *ASGR2*, *RP11-342M1.7*) showed differential effects by site, with strongest trans-eQTL signals in proximal colonic mucosa (e.g. *AKAP14*, beta=0.61, $P=5.02\times10^{-5}$) and opposite signals in distal mucosa (*AKAP14*, beta=-0.17, P=0.04).

In summary, genetic variation at the chr11q23.1 risk locus imparts greater risk of distal rather than proximal CRC and exhibits site-specific differences in eQTL effects in normal mucosa. Topographical differences in genomic control over gene expression relevant to CRC risk may underlie site-specific variation in CRC. Results may inform individualised CRC screening programmes.

Introduction

Genome-wide association studies (GWAS) on large, well-characterised case-control series of colorectal cancer (CRC) have identified numerous common genetic variants associated with individually modest effects on CRC risk. Identifying the underlying causal mechanism may provide novel targets for cancer prevention or therapy. However, as these variants frequently lie in inter-genic positions or non-coding regions of the gene, mechanisms responsible for risk modification are not readily identifiable. The advent of high throughput genotyping and transcriptomic profiling has enabled identification of associations between CRC risk variants and gene expression levels within the normal colorectal mucosa, known as expression quantitative trait loci (eQTLs) [1-4].

Site-specific variation in CRC incidence [5], biology [6, 7], response to adjuvant therapy [8], and prognosis [9-12] are well recognised, yet incompletely understood. Differences may reflect different embryological origin or different exposure to faecal stream, microbiome and carcinogens [13]. However, topographical differences in eQTL effects may explain part of the observed differences in CRC risk. We previously reported differential effects of rs3802842 and rs4939827 on cancer risk in the rectum and colon [14]. Furthermore, differences in gene expression have been observed in normal mucosa [15] and cancers [16-18] originating from left and right colon. One explanation for differences in gene expression and CRC risk is differential genomic control through site-specific eQTL effects. Differences in eQTL effects for known CRC risk SNPs previously reported [3] using GTEx RNAseq data from harvested transverse (full thickness) and sigmoid colon (smooth muscle only) samples are likely confounded tissue of origin effects since the sigmoid samples contain no colonic mucosa.

We hypothesised that a subset of risk loci might impart site-specific effects on colorectal cancer risk through topographical differences in genomic control over gene expression. To investigate this, we first tested CRC risk loci for differential site-specific effects. We then sought to identify differences in gene expression across the colorectum and explored association between CRC risk loci and gene expression through site-specific eQTL analysis.

Methods

Site-specific association study

We searched relevant GWAS to identify all putative loci impacting CRC risk, with 160 loci identified, including those from the most recently reported GWAS [19, 20]. The list was constrained to include new and those replicated at $p<1x10^{-6}$ level and excluding variants identified previously in GWAS studies of Asian and African subjects. We also excluded SNPs identified by conditional analysis at known CRC risk loci, instead using the lead SNP (n=87) (Supplementary Table 1). We then ran association analyses for the risk of distal and proximal cancers in previously described cases-control studies of colorectal cancer in Scotland and UK Biobank (Supplementary Methods; [19]). Tumours located proximal or at the splenic flexure (ICD codes C18.0, C18.2, C18.3, C18.4, C18.5), were defined as proximal and cancer cases with tumours located distal to the splenic flexure were counted as distal CRC (ICD codes C18.7, C19, C20), reflecting the embryological origin of midgut and hindgut respectively. The associations between cancer sites and genetic variants were tested using a multinomial logistic regression likelihood as implemented in SNPTEST v2.5.2 [21-23] and assuming additive model of inheritance. Meta-analyses of four case-control studies across proximal and colorectal cancer were performed using the fixed-effects inverse-variance method using META v1.7 [24]. Cochran's Q-statistic to test for heterogeneity and the I2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated. Finally, we performed case-only analysis to study effects of genetic variants on the risk of developing distal compare to proximal cancers. Results of individual case-only analyses were combined in the fixed effects inverse-variance meta-analysis as implemented in META v.1.7.

Study population eQTL analysis

We included CRC cases from the Study of Colorectal Cancer in Scotland (SOCCS), a population-based case-control study designed to identify genetic and environmental factors impacting on CRC risk and survival outcomes [25]. We also included pre-treatment samples from participants in the Scottish Vitamin D study (SCOVIDS), who comprised patients with previous history of CRC and healthy volunteers. Clinical variables were collected from clinical records systems and pathology records, entered into a prospective study database and extracted for analysis.

Mucosa sampling and storage

Normal colorectal mucosa (NM) was sampled from a single site from freshly resected surgical specimens or rectal biopsy. Samples were immersed in the stabilization solution RNAlater (Invitrogen). All samples were kept in RNAlater for 24-72 hours prior to RNA extraction or storage at -80°C. Assessment of gene expression was performed using Illumina HumanHT-12v4 BeadChip with validation of top genes performed using standard qRT-PCR (see Supplementary Methods).

Differential expression analysis

All statistical analysis was undertaken in R [26]. Investigation of differential gene expression was undertaken using the *lmFit* and *eBayes* functions within the limma package with a total of 42,184 probes assessed. PEER factors [27] were estimated on the processed expression matrix, and were used as covariates in the model together with age and gender. Adjustment for multiple testing was undertaken and FDR p-values derived [28]. Linear regression modelling was used to adjust for relevant demographic variables (PEER factors, age and gender), with adjusted fold-difference in expression between samples taken from distal and proximal sites calculated as the antilog_2 of the model beta value.

Functional Pathway analysis

Gene ontology and enrichment analysis was undertaken using the 'GOrilla', Gene Ontology enRIchment anaLysis and visuaLizAtion tool through the Gorilla web page[29]. Process, Function and and Component ontologies were investigated using gene lists from differential gene expression analysis by site ranked by unadjusted p-value from smallest to largest. Terms enriched at FDR<0.05 were considered to be significantly enriched.

eQTL analysis

We performed genotyping of study subjects for the list of candidate risk SNPs. Genotyping of SOCCS subjects was conducted as previously described[14, 19, 30-33], with SCOVIDS subjects genotyped using the OmniExpressExome BeadChip 8v1.3 or 8v1.4 (Illumina Inc., San Diego, CA). Where necessary, imputed genotypes were used with imputation and related quality control procedures performed as previously described (Supplementary methods; [34, 35]).

For eQTL analysis, 462 samples were retained that passed quality control and for which we had genotyping data. The analysis was carried out for all samples together, and also separately by site of mucosa sample (proximal=113, distal=349). eQTL discovery was carried out with matrix eQTL [36] using linear model adjusted for age, gender and 15 PEER factors. Only the

additive genetic model was used with genotypes considered as a quantitative variable. To explore the influence of anatomical location on eQTL signals, an analysis stratified by site was performed and charted for those probes with putative eQTL signals (nominal $p<0.05$ at either both sites combined or in proximal samples or in distal samples). Interaction analysis with the site was performed using "modelLINEAR_CROSS" model specification as implemented in Matrix eQTL.

Results

Site-specific genome-wide association study

We conducted site-specific meta-analyses of case-control association studies (3089 proximal colon cancer cases, 7541 distal colorectal cancer cases, 31,331 controls) to identify whether established risk variants (n=87) influence risk of proximal and distal CRC differently (Supplementary Table 2). Using the collated cases, we then performed case-only analysis to study effects of genetic variants on the risk of developing distal compare to proximal cancers across each of the included studies. We meta-analysed results of individual studies with just one locus (rs3087967, $P3.32x10^{-5}$, FDR 0.003) showing evidence for differential risk of proximal versus distal cancer after FDR correction. Hits with a nominal p value<0.01 in the case only analysis are given in Table 1 (allele frequencies given in Supplementary Table 3, case-only site-specific results given in Supplementary Table 2).

Topographical gene expression analysis

Next, we performed gene expression profiling to assess for topographical variation in normal mucosa gene expression between the proximal colon (proximal to splenic flexure) and distal colorectum (colon distal to splenic flexure and rectum). NM samples from 481 unique subjects were analysed (Table 2).

Transcriptomic analysis provided expression data for 42436 probes and 28707 unique named genes after QC. We identified differential expression between the proximal and distal colorectum for 1430 probes, accounting for 1261 genes (Table 3, Supplementary Table 4; Figure 1) with 486 genes more highly expressed in the proximal colonic mucosa. *PRAC* was the top differentially expressed gene between proximal and distal samples in subgroup analyses of NM from patients with CRC and those subjects without CRC. 619 differentially expressed genes by site were seen in the 329 samples from patients with CRC and 255 differentially

expressed genes by site were seen in the 152 samples from subjects without CRC (Supplementary Table 5).

Gene ontology analysis demonstrated enrichment of numerous processes in relation to mucosal sampling site, with many hits relevant to carcinogenesis including cell cycle checkpoint, cell division and DNA repair (Supplementary Table 6).

qRT-PCR replication (n=116 subjects) of the top two differentially expressed genes between proximal and distal sites, confirmed significantly greater *PRAC* expression in distal colon ($P<2.2X10^{-16}$, Supplementary Figure 1) and greater expression of *PITX2* in the proximal colon ($P<2.2X10^{-16}$, Supplementary Figure 2). Good correlation between HT12 expression and qRT-PCR expression values were observed for both genes, R=0.79 and R=0.84 respectively, $P<2.2X10^{-16}$, Supplementary Figures 1 and 2).

eQTL analysis

To explore whether the site-specific effect on risk with rs3087967 genotype could be associated with topographical differences in genomic control over gene expression we performed cis- and trans-eQTL analysis. First, we sought association between genotype at chr11q23.1 and expression of genes within a 1MB distance up and downstream of the transcription start site. Of the 34 probes assessed, 8 showed putative eQTL signals (nominal $p<0.05$ at either all sites combined or in proximal samples or in distal samples, Supplementary Table 7), including strong cis-eQTL effects associated with the expression of *COLCA2* (FDR $4.52x10^{-70}$), *COLCA1* (FDR $2.48X10^{-40}$) and *C11orf53* (FDR $4.74x10^{-7}$). Cis-eQTL effects associated with *PPP2R1B* expression (FDR 0.004) were seen in proximal colonic mucosa samples only and not seen when all sites were combined, and these site-specific differences maintained in a subgroup analysis including only CRC cases (Supplementary Table 8). Cis-eQTL effects associated with *PIH1D2* expression (FDR 0.01) were seen in distal colorectal mucosa samples only. *COLCA2* eQTL effects were stronger in proximal colonic samples (beta 0.98 *vs.* 0.84, Supplementary Figure 3), yet on formal interaction testing no probe showed significant differential eQTL effects between the proximal and distal colorectum (Supplementary Table 7). There were no baseline differences in expression in these 8 probes between proximal and distal sample sites in the complete sample set (Supplementary Table 9).

To corroborate these data, COLONOMICS [37] and GTEx [38] were interrogated for data on eQTL effects at rs3087967 and expression of the 8 genes reported in Supplementary Table 6 (Supplementary Tables 10 and 11). Both COLONOMICS and GTEx eQTL data correlated

with the current findings showing stronger eQTL in proximal mucosa for *COLCA2*, *FDX1* and *C11orf53* (COLONOMICS only) yet in contrast to our data showed stronger effects for *PPP2R1B* in distal samples. We also tested for differential expression in COLONOMICS data between mucosa from healthy controls and adjacent normal mucosa from CRC patients. Significantly lower expression was seen in mucosa from CRC patients for *PPP2R1B, PIH1D2, COLCA2*, *C11orf53, FDX1 and COLCA1* (Supplementary Table 12).

As cis-eQTLs explain only a small fraction of total transcript-level heritability [39], we next performed trans-eQTL analysis for association between rs3087967 genotype and expression of 35,375 probes. Trans-eQTL effects were found to be associated with expression of 23 probes accounting for 21 genes (FDR<0.05, top hit *LRMP* FDR $6.01X10^{-12}$, Supplementary Table 13). Probes with a putative trans-eQTL signal (nominal p<0.05 at all sites combined or in proximal samples or distal samples, n=3798) were tested for differential trans-eQTL effects using a site interaction model. This revealed 4 probes with differential trans-eQTL effects dependent on site, all driven by eQTL signals in proximal colonic samples (top hit *AKAP14, interaction* FDR 0.006; Table 4; Figure 2). Site-specific trans-eQTL effects were maintained in a subgroup analysis including only CRC cases (Supplementary Table 8). There were no differences in expression in these 4 probes between proximal and distal sample sites in the complete sample set (Supplementary Table 14), or in COLONOMICS data (normal tissue).

Finally, we tested for differential expression in COLONOMICS data between mucosa from healthy controls and adjacent normal mucosa from CRC patients. No differences in *AKAP14* or *ASGR2* were seen. Of potential interest, the expression of *ADH5* was less in adjacent mucosa from CRC patients, but only when comparing proximal samples (proximal expression 7.89 *vs.* 7.55, p=0.008; distal expression 7.89 *vs.* 7.72, p=0.3).

Discussion

We report a comprehensive analysis of site-specific difference in genetic risk for CRC and explore transcriptomic data for variation in gene expression, and eQTL effects dependent on mucosa sampling site. We show that genetic variation at the chr11q23.1 CRC risk locus imparts significantly greater risk of distal rather than proximal CRC. Trans-eQTL analysis demonstrates significant differential eQTL effects for the chr11q23.1 locus between the proximal and distal colorectum, suggesting that site-specific effects on risk may be attributable to topographical differences in genomic control over gene expression in large bowel mucosa. These findings shed further light on differential genomic control effects on gene expression relevant to CRC risk.

These findings establish that, at least in the case of rs3087967 as a paradigm, there are site-specific differential effects of CRC risk loci and that these might be mediated through changes in eQTLs. The degree of differential expression provided strong rationale to then test for site-specific effects on risk. Of the established risk loci tested, only the chr11q23.1 locus imparted significantly different risk between the proximal and distal colon. Previous data had demonstrated a greater risk of rectal cancer for both rs3802842 and rs4939827 [14].

For the purposes of this analysis, we partitioned the large bowel by the embryological interface at the splenic flexure. Due to statistical power, it was not appropriate to provide further breakdown of sites, yet given data suggesting linearity in tumour characteristics beyond the simple proximal-distal divide [11], we acknowledge that there may be further risk SNPs that exert differential effects on risk and gene expression in an anatomically biased manner.

Genotype at the rs3087967 SNP imparted risk on the distal colorectum (OR=1.20, $P=1.28X10^{-20}$), with no significant impact on proximal colonic cancer risk. This locus is in LD ($r^2=1$, D'=1) with the previously reported SNP rs3802842 which also shows association between genotype and distal CRC but not proximal cancer risk [40-42].

We demonstrate a large number of genes with differential expression between the proximal and distal colorectum, validating previous reports [15-18] and supporting our downstream site-specific eQTL analysis. Functional annotation indicate differences in processes relevant to carcinogenesis including regulation of cell cycle checkpoint, cell division and DNA damage responses which may underlie site-specific variation in CRC molecular pathogenesis and increased sensitivity to certain chemotherapy regimens [8].

Trans-eQTL analysis here validate previously reported findings [1, 37]. The TT genotype at rs3087967 is associated with higher expression of *AKAP14 (*A-kinase anchor protein 14), *ASGR2* and *ADH5P4 (*Alcohol dehydrogenase 5) in proximal colonic samples (FDR<0.05),

but no effect on expression of these genes in distal colorectal samples (FDR>0.05). *ASGR2* (Asialoglycoprotein Receptor 2) is upregulated in metastatic colon cancer [43], while altered expression of alcohol dehydrogenases in CRC is also reported [44-46]. Further investigation is required to define the relevance of *AKAP14, ASGR2* and *ADH5P4* to the observed site-specific risk associated with the rs3087967 TT genotype.

We acknowledge several limitations within the current study. First, we only identified one locus with site-specific risk and we acknowledge increased sample size may uncover further relevant hits. Co-linearity in sample site and cancer status (94% samples from non-cancer patients were from rectum) precluded a robust case-control analysis within the mucosa dataset, and eQTLs may differ between cases and controls thus impacting our analysis and downstream conclusions. The significance of this sampling co-linearity may introduce a bias if current CRC differentially influences expression in distal and proximal tissue samples. To address this, we performed case-control gene expression analysis, stratified by sample site and identified no differences in gene expression between CRC-cases and controls (*data not shown*). We also performed eQTL analysis in CRC-cases only with site-specific eQTL effects reported in our overall cohort maintained in this sub-group (Supplementary Table 8). It was not appropriate to perform this analysis in 'non-CRC' cases, given the low number of proximal samples in subjects without CRC (N=8), thus further studies should consider how best to reliably sample the proximal colon outwith the operating theatre (e.g. at colonoscopy). Finally, we recognise that that while the rs3087967 locus is associated with distal CRC risk, the observed eQTL effects for this locus are strongest in proximal mucosal samples. It is unclear why this might be, but given that distal and proximal cancers are known to have different risk factors, both genetic and environmental [47, 48], it is reasonable to propose that the mechanism by which a locus imparts risk for proximal and distal cancer may be different, with possible interplay with environmental factors. Such factors might include stool make-up, the microbiome, obesity, physical activity, smoking or aspirin exposure, which could underlie the absence of relevant eQTL effects in distal mucosa samples.

Despite these limitations this is, to our knowledge, the first study to perform site-specific genetic association analysis and carry candidate loci forward to explore whether differences in genomic control over gene expression might underlie site-specific risk. We report a single locus with site specific risk and a number of trans-eQTLs which might account for this. We also identify numerous strong trans-eQTLs which provide important candidates for future functional characterisation. Finally, our findings emphasise the importance of considering site-

specific risk or eQTL effects, as subtle effects at proximal sites might be masked by a nil/opposite effect in the distal colorectum or vice-versa.

Conclusions

Genetic variation at the chr11q23.1 CRC risk locus imparts significantly greater risk of distal rather than proximal CRC and this analysis is consistent with site-specific differences in eQTL effects in normal mucosa. These findings shed further light on differential genomic control effects on gene expression relevant to CRC risk. While current CRC screening programmes consider highly penetrant rare variants, future individualised screening programmes may be informed by risk imparted by common genetic variation. Insight into site-specific CRC risk imparted by such variants will help define individualised screening programmes with screening frequency, modality and focus tailored to that specific individual's risk.

Acknowledgements

We acknowledge Fanny Roth who undertook qRT-PCR validation of the top differentially expressed genes seen HT12 analysis as part of an Erasmus internship. We acknowledge the excellent technical support from Marion Walker and Stuart Reid. We are grateful to Donna Markie and Fiona McIntosh, and all those who continue to contribute to recruitment, data collection, and data curation for the Study of Colorectal Cancer in Scotland studies. We acknowledge that these studies would not be possible without the patients and surgeons who take part and the NHS Lothian Bioresource team which contributed to the collection and storage of NM samples for this study. We acknowledge the expert support on sample preparation from the Genetics Core of the Edinburgh Wellcome Trust Clinical Research Facility.

Conflicts of interest

The authors declare no potential conflicts of interest.

Data Availability Statement

This work has been conducted using the UK Biobank Resource under Application number 7441. The UK Biobank is an open access resource and bona fide researchers can apply to use the UK Biobank dataset by registering and applying at http://ukbiobank.ac.uk/register-apply/ The HT12 gene expression data and phenotype data generated in this study are available in GEO under accession number GSE161023. Other data that support the findings of this study are available from the corresponding author upon request.

Ethics statement

All participants provided informed written consent, and research was approved by local research ethics committees (SOCCS 11/SS/0109 and 01/0/05; SCOVIDS 13/SS/0248) and National Health Service management (SOCCS 2013/0014, 2003/W/GEN/05; SCOVIDS 2014/0058).

Funding

This work was supported by funding for the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre; CRUK programme grant C348/A18927 (MGD). PVS was supported by MRC Clinical Research Training Fellowship (MR/M004007/1), a Research Fellowship from the Harold Bridges Bequest and by the Melville Trust for the Care and Cure of Cancer. The work received support from COST Action BM1206. ET is supported by a CRUK Career Development Fellowship (C31250/A22804). LYO is supported by a Cancer Research UK Research Training Fellowship (C10195/A12996). This work was also funded by a grant to MGD as Project Leader with the MRC Human Genetics Unit Centre Grant (U127527202 and U127527198 from 1/4/18). JB is supported by an Edinburgh Clinical Academic Track (ECAT) linked Cancer Research UK Clinical Research Fellowship (C157/A23218).

Role of the Funding Source

The funder had no role in design, undertaking, analysis or writing of the above study.

References

1. Closa A, Cordero D, Sanz-Pamplona R, Sole X, Crous-Bou M, Pare-Brunet L, Berenguer A, Guino E, Lopez-Doriga A, Guardiola J, Biondo S, Salazar R, et al. Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis* 2014;**35**: 2039-46.

2. Loo LW, Cheng I, Tiirikainen M, Lum-Jones A, Seifried A, Dunklee LM, Church JM, Gryfe R, Weisenberger DJ, Haile RW, Gallinger S, Duggan DJ, et al. cis-Expression QTL analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue. *PLoS One* 2012;**7**: e30477.

3. Loo LWM, Lemire M, Le Marchand L. In silico pathway analysis and tissue specific cis-eQTL for colorectal cancer GWAS risk variants. *BMC Genomics* 2017;**18**: 381.

4. Hulur I, Gamazon ER, Skol AD, Xicola RM, Llor X, Onel K, Ellis NA, Kupfer SS. Enrichment of inflammatory bowel disease and colorectal cancer risk variants in colon expression quantitative trait loci. *BMC Genomics* 2015;**16**: 138.

5. Meza R, Jeon J, Renehan AG, Luebeck EG. Colorectal cancer incidence trends in the United States and United kingdom: evidence of right- to left-sided biological gradients with implications for screening. *Cancer Res* 2010;**70**: 5419-29.

6. Lee MS, Menter DG, Kopetz S. Right Versus Left Colon Cancer Biology: Integrating the Consensus Molecular Subtypes. *J Natl Compr Canc Netw* 2017;**15**: 411-9.

7. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015;**21**: 1350-6.

8. Elsaleh H, Joseph D, Grieu F, Zeps N, Spry N, Iacopetta B. Association of tumour site and sex with survival benefit from adjuvant chemotherapy in colorectal cancer. *Lancet* 2000;**355**: 1745-50.

9. Petrelli F, Tomasello G, Borgonovo K, Ghidini M, Turati L, Dallera P, Passalacqua R, Sgroi G, Barni S. Prognostic Survival Associated With Left-Sided vs Right-Sided Colon Cancer: A Systematic Review and Meta-analysis. *JAMA Oncol* 2016.

10. Iacopetta B. Are there two sides to colorectal cancer? *Int J Cancer* 2002;**101**: 403-8.

11. Yamauchi M, Lochhead P, Morikawa T, Huttenhower C, Chan AT, Giovannucci E, Fuchs C, Ogino S. Colorectal cancer: a tale of two sides or a continuum? *Gut* 2012;**61**: 794-7.

12. Lee GH, Malietzis G, Askari A, Bernardo D, Al-Hassi HO, Clark SK. Is right-sided colon cancer different to left-sided colorectal cancer? - a systematic review. *Eur J Surg Oncol* 2015;**41**: 300-8.

13. Lewin MR, Ferulano GP, Cruse JP, Clark CG. Experimental colon carcinogenesis is facilitated by endogenous factors in the intestinal contents. *Carcinogenesis* 1981;**2**: 1363-6.

14. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N, Semple C, Clark AJ, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 2008;**40**: 631-7.

15. Glebov OK, Rodriguez LM, Nakahara K, Jenkins J, Cliatt J, Humbyrd CJ, DeNobile J, Soballe P, Simon R, Wright G, Lynch P, Patterson S, et al. Distinguishing right from left colon by the pattern of gene expression. *Cancer Epidemiol Biomarkers Prev* 2003;**12**: 755-62.

16. Sun Y, Mironova V, Chen Y, Lundh EPF, Zhang Q, Cai Y, Vasiliou V, Zhang Y, Garcia-Milian R, Khan SA, Johnson CH. Molecular Pathway Analysis Indicates a Distinct Metabolic Phenotype in Women With Right-Sided Colon Cancer. *Transl Oncol* 2020;**13**: 42-56.

17. Pira G, Uva P, Scanu AM, Rocca PC, Murgia L, Uleri E, Piu C, Porcu A, Carru C, Manca A, Persico I, Muroni MR, et al. Landscape of transcriptome variations uncovering known and novel driver events in colorectal carcinoma. *Sci Rep* 2020;**10**: 432.

18. Birkenkamp-Demtroder K, Olesen SH, Sorensen FB, Laurberg S, Laiho P, Aaltonen LA, Orntoft TF. Differential gene expression in colon cancer of the caecum versus the sigmoid and rectosigmoid. *Gut* 2005;**54**: 374-84.

19. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, Farrington S, Svinti V, Palles C, Orlando G, Sud A, Holroyd A, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun* 2019;**10**: 2154.

20. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, Conti DV, Qu C, Jeon J, Edlund CK, Greenside P, Wainberg M, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* 2019;**51**: 76-87.

21. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;**39**: 906-13.

22. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;**447**: 661-78.

23. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010;**11**: 499-511.

24. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Berrettini W, Knouff CW, Yuan X, Waeber G, Vollenweider P, Preisig M, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010;**42**: 436-40.

25. Theodoratou E, Kyle J, Cetnarskyj R, Farrington SM, Tenesa A, Barnetson R, Porteous M, Dunlop M, Campbell H. Dietary flavonoids and the risk of colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 2007;**16**: 684-93.

26. R Development Core Team. R: A language and environment for statistical computing: R Foundation for Statistical Computing, Vienna, Austria, 2013.

27. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 2012;**7**: 500-7.

28. Benjamini Y, Y H. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* 1995;**57**.

29. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 2009;**10**: 48.

30. Vaughan-Shaw PG, Zgaga L, Ooi LY, Theodoratou E, Timofeeva M, Svinti V, Walker M, O'Sullivan F, Ewing A, Johnston S, Din FVN, Campbell H, et al. Low plasma vitamin D is associated with adverse colorectal cancer survival after surgical resection, independent of systemic inflammatory response. *Gut* 2020;**69**: 103-11.

31. Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, Whiffin N, Tenesa A, Spain S, Broderick P, Ooi LY, Domingo E, Smillie C, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* 2012;**44**: 770-6.

32. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, Howarth K, Spain SL, Broderick P, Domingo E, Farrington S, Prendergast JG, Pittman AM, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* 2010;**42**: 973-7.

33. Timofeeva MN, Kinnersley B, Farrington SM, Whiffin N, Palles C, Svinti V, Lloyd A, Gorman M, Ooi LY, Hosking F, Barclay E, Zgaga L, et al. Recurrent Coding Sequence Variation Explains Only A Small Fraction of the Genetic Architecture of Colorectal Cancer. *Sci Rep* 2015;**5**: 16286.

34. Orlando G, Law PJ, Palin K, Tuupanen S, Gylfe A, Hanninen UA, Cajuso T, Tanskanen T, Kondelin J, Kaasinen E, Sarin AP, Kaprio J, et al. Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Human molecular genetics* 2016;**25**: 2349-59.

35. Rodriguez-Broadbent H, Law PJ, Sud A, Palin K, Tuupanen S, Gylfe A, Hanninen UA, Cajuso T, Tanskanen T, Kondelin J, Kaasinen E, Sarin AP, et al. Mendelian randomisation implicates hyperlipidaemia as a risk factor for colorectal cancer. *Int J Cancer* 2017;**140**: 2701-8.

36. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 2012;**28**: 1353-8.

37. Moreno V, Alonso MH, Closa A, Valles X, Diez-Villanueva A, Valle L, Castellvi-Bel S, Sanz-Pamplona R, Lopez-Doriga A, Cordero D, Sole X. Colon-specific eQTL analysis to inform on functional SNPs. *Br J Cancer* 2018;**119**: 971-7.

38. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**: 580-5.

39. Cheng TH, Thompson D, Painter J, O'Mara T, Gorman M, Martin L, Palles C, Jones A, Buchanan DD, Ko Win A, Hopper J, Jenkins M, et al. Meta-analysis of genome-wide association studies identifies common susceptibility polymorphisms for colorectal and endometrial cancer near SH2B3 and TSHZ1. *Sci Rep* 2015;**5**: 17369.

40. Mates IN, Jinga V, Csiki IE, Mates D, Dinu D, Constantin A, Jinga M. Single nucleotide polymorphisms in colorectal cancer: associations with tumor site and TNM stage. *J Gastrointestin Liver Dis* 2012;**21**: 45-52.

41. Lubbe SJ, Whiffin N, Chandler I, Broderick P, Houlston RS. Relationship between 16 susceptibility loci and colorectal cancer phenotype in 3146 patients. *Carcinogenesis* 2012;**33**: 108-12.

42. Pittman AM, Webb E, Carvajal-Carmona L, Howarth K, Di Bernardo MC, Broderick P, Spain S, Walther A, Price A, Sullivan K, Twiss P, Fielding S, et al. Refinement of the basis and impact of common 11q23.1 variation to the risk of developing colorectal cancer. *Hum Mol Genet* 2008;**17**: 3720-7.

43. Hartung F, Wang Y, Aronow B, Weber GF. A core program of gene expression characterizes cancer metastases. *Oncotarget* 2017;**8**: 102161-75.

44. Druesne-Pecollo N, Tehard B, Mallet Y, Gerber M, Norat T, Hercberg S, Latino-Martel P. Alcohol and genetic polymorphisms: effect on risk of alcohol-related cancer. *Lancet Oncol* 2009;**10**: 173-80.

45. Jelski W, Szmitkowski M. Alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH) in the cancer diseases. *Clin Chim Acta* 2008;**395**: 1-5.

46. Tiemersma EW, Wark PA, Ocke MC, Bunschoten A, Otten MH, Kok FJ, Kampman E. Alcohol consumption, alcohol dehydrogenase 3 polymorphism, and colorectal adenomas. *Cancer Epidemiol Biomarkers Prev* 2003;**12**: 419-25.

47. Demb J, Earles A, Martinez ME, Bustamante R, Bryant AK, Murphy JD, Liu L, Gupta S. Risk factors for colorectal cancer significantly vary by anatomic site. *BMJ Open Gastroenterol* 2019;**6**: e000313.

48. Carethers JM. Risk factors for colon location of cancer. *Transl Gastroenterol Hepatol* 2018;**3**: 76.

Tables

Table 1. Top SNPS with evidence for differential risk of proximal versus distal cancer

		Case-only analysis			Proximal colon case-control meta-analysis		Distal CRC case-control meta-analysis		Side with greatest risk ↑
RSID	**Effect allele**	**OR**	**Interaction**		**OR (95%CI)**	**P value**	**OR (95%CI)**	**P value**	
			P value	**FDR**					
rs3087967	T	1.14	$3.32x10^{-5}$	0.003	1.05 (0.99-1.11)	0.10	1.20 (1.15-1.25)	$8.20x10^{-20}$	Distal
rs1330889	C	1.14	$5.87x10^{-3}$	0.14	0.99 (0.92-1.08)	0.85	1.12 (1.06-1.19)	$6.53x10^{-5}$	Distal
rs35470271	G	1.12	$7.25x10^{-3}$	0.14	1.01 (0.93-1.09)	0.85	1.12 (1.06-1.18)	$1.39x10^{-5}$	Distal
rs6055286	A	0.90	$9.37x10^{-3}$	0.14	1.19 (1.10-1.28)	$5.96x10^{-6}$	1.06 (1.01-1.12)	$2.44x10^{-2}$	Proximal
rs7593422	T	1.09	$4.34x10^{-3}$	0.14	1.02 (0.97-1.08)	0.40	1.12 (1.08-1.17)	$7.18x10^{-10}$	Distal

Case-control meta-analysis results given for case-control association study of 3089 proximal colon cancer cases and 31,331 controls and between 7541 distal colorectal cancer cases and 31,331 controls, nominal P value for association with risk given. Significant association (FDR<0.05) confirmed at 70 SNPs for distal CRC and 42 SNPs for proximal CRC. Case-only meta-analysis OR and interaction P value indicates association between SNP and risk of distal CRC compared to risk of proximal colon cancer. Interaction FDR adjusted for 87 SNPs tested.

new

Table 2. Baseline characteristics in participants included in gene expression analysis

Age	Median 69 (range 17-91)
Gender	230 (48% female)
Diagnosis	
Healthy	66 (14%)
Colorectal adenocarcinoma	329 (68%)
Haemorrhoids	20 (4%)
Previous CRC	15 (3%)
Colorectal adenoma	10 (2%)
Fistula-in-ano	7 (1%)
Diverticular disease	6 (1%)
Anal intra-epithelial neoplasia	4 (1%)
Fissure-in-ano	4 (1%)
Pilonidal disease	3 (1%)
Malignancy (other)	8 (2%)
Other benign anorectal condition	5 (1%)
Other miscellaneous	6 (1%)
Sample site	
Caecum	15 (3%)
Transverse colon	96 (20%)
Not specified (proximal)	5 (1%)
Not specified (distal)	15 (3%)
Descending colon	119 (25%)
Sigmoid colon	17 (4%)

Rectum	214 (44%)

116 (24%) samples were from proximal colon. Samples taken in patients with CRC comprised 108 from proximal colon and 221 from distal colorectum. Samples taken in patients without CRC comprised 8 from proximal colon and 143 from distal colorectum.

Table 3. Top genes with differences in expression between distal and proximal colorectum mucosa samples

ILMN Probe ID	**Gene**	**Fold-difference**	**FDR p-value**	**Alternative ILMN Probe ID**	**Fold-difference**	**FDR p-value**
ILMN_3248384	*PRAC*	9.92	$1.38X10^{-57}$	ILMN_1801832	10.06	$1.38X10^{-57}$
ILMN_2391400	*PITX2*	0.41	$1.08X10^{-24}$	ILMN_1796847	0.62	$1.26X10^{-17}$
ILMN_1742677	*HOXB13*	1.97	$3.70X10^{-20}$	-	-	-
ILMN_3233239	*LOC731789*	1.42	$2.66X10^{-17}$	ILMN_3230024	0.98	0.61
ILMN_2072568	*CLDN8*	2.10	$1.72X10^{-16}$	ILMN_1746676	1.93	$1.53X10^{-14}$
ILMN_1696028	*ETNK1*	0.67	$6.01X10^{-14}$	ILMN_2316778	0.75	$2.70x10^{-8}$
ILMN_2364864	*MB*	0.67	$1.00X10^{-13}$	ILMN_1666109	0.61	$1.00X10^{-13}$
ILMN_3236709	*C17orf93*	1.47	$1.09X10^{-13}$	-	-	-
ILMN_3248309	*LOC732215*	1.32	$2.72X10^{-13}$	-	-	-
ILMN_1769839	*L1TD1*	0.56	$3.60X10^{-13}$	-	-	-

Top genes with differential proximal/distal expression, with fold-difference adjusted for age, gender and PEER factors given. Analysis not adjusted for CRC status given collinearity between CRC status and sample site. Alternative probes for top genes given where available.

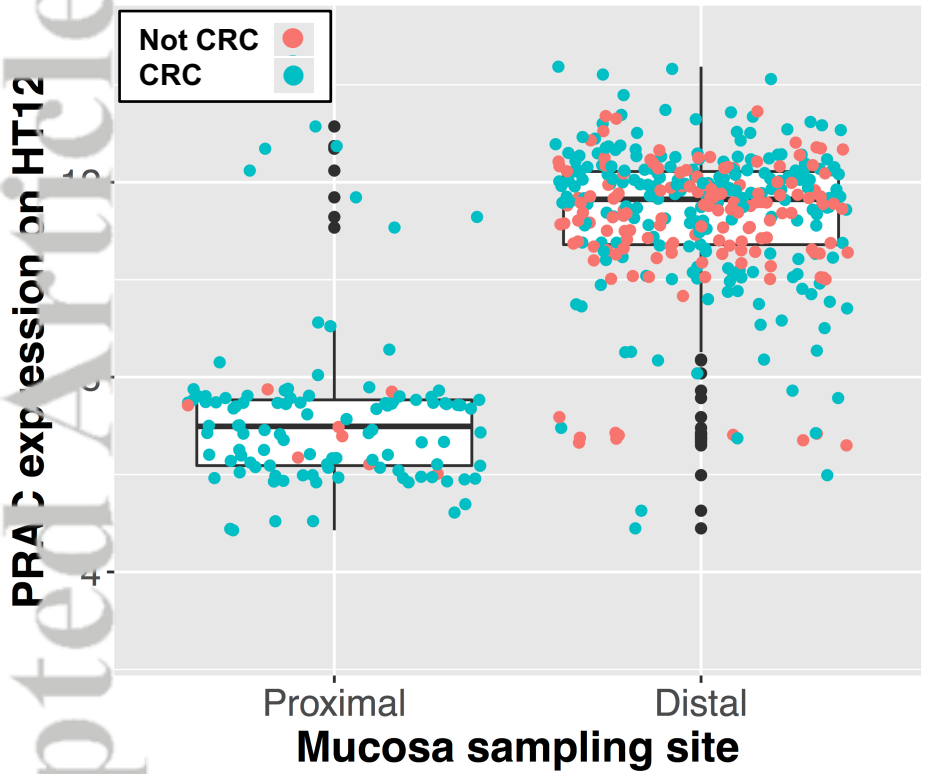
Table 4 Site interaction analysis for trans-eQTL signals for rs3087967 (11:111156836:C:T)

Gene	eQTL in all sample sites combined			eQTL in proximal colonic mucosa samples			eQTL in distal colorectal mucosa samples			Interaction analysis		
	Beta	**P value**	**FDR**	**Beta**	**P value**	**FDR**	**Beta**	**P**	**FDR**	**Beta**	**P value**	**FDR**
AKAP14	0.02	0.80	0.99	0.61	$5.02x10^{-5}$	**0.02**	-0.17	0.037	0.12	0.79	$1.61x10^{-6}$	**0.006**
ADH5P4	0.05	0.34	0.96	0.48	$2.06x10^{-4}$	**0.04**	-0.08	0.22	0.36	0.54	$1.72x10^{-5}$	**0.032**
ASGR2	0.07	0.30	0.95	0.48	$2.70x10^{-4}$	**0.04**	-0.10	0.22	0.36	0.67	$2.49x10^{-5}$	**0.032**
RP11-342M1.7	-0.02	0.65	0.99	0.35	$1.70x10^{-3}$	**0.07**	-0.14	0.02	0.11	0.46	$4.61x10^{-5}$	**0.044**
ACCS	-0.05	0.34	0.97	-0.38	$3.39x10^{-3}$	0.09	0.07	0.27	0.42	-0.51	0.0001	0.069

Overall eQTL FDR adjusted for all probes within trans region (n=35,375). Site-specific eQTL and interaction FDR adjusted for 3798 probes with putative evidence of eQTL (nominal p<0.05) at any site (all sites combined, proximal or distal). Table shows top 5 hits for interaction analysis between SNP and sample site.

Novelty and Impact:

Common genetic variants are known to influence colorectal cancer (CRC) risk. In this study, the authors asked whether sequences that influence gene-expression levels, known as “expression quantitative trait loci” (eQTLs), might lead to different transcription patterns depending on where in the mucosa the cells occur (e.g., proximal vs distal mucosa). They found site-specific, trans-eQTL effects for four genes that affect a known CRC-risk locus. These topographical differences in genomic control of gene expression may lead to more highly-individualised tools for CRC screening programs.



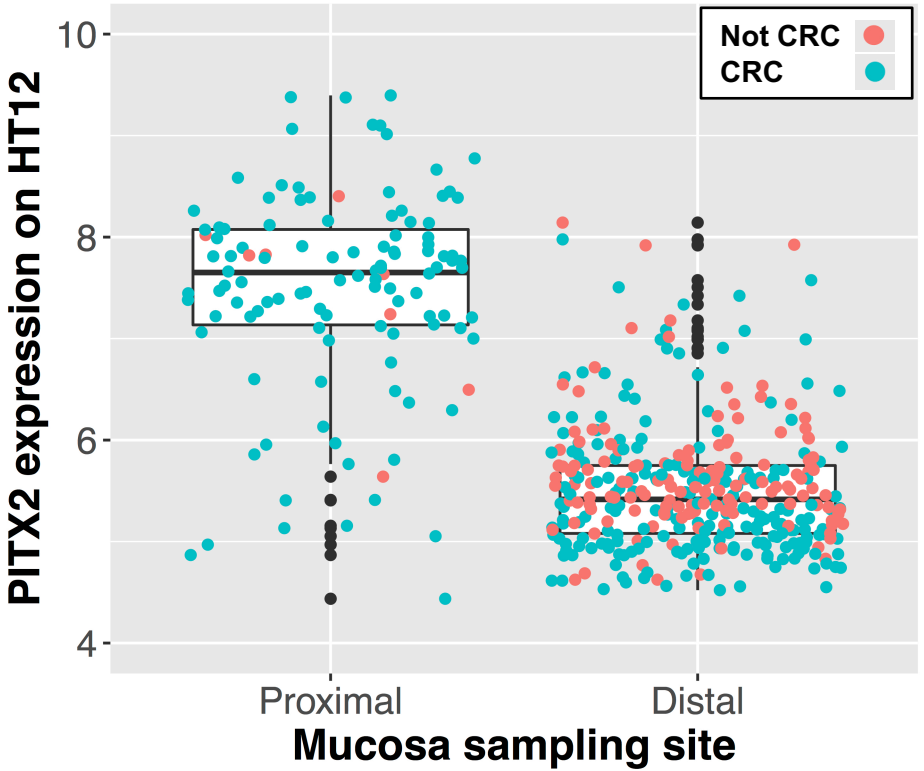


Figure 1. Log_2 expression of *PRAC* and *PITX2* - top two differentially expressed genes between proximal and distal colorectum

Normal mucosa was sampled from resected colorectal specimens or by rectal biopsy. RNA was extracted and gene expression assessed using HT12 microarrays. Expression of PRAC and PITX2 expression, which were found to be significantly associated with sample site is charted. The lower and upper hinges correspond to the first and third quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge. CRC classification at time of sampling indicated by dot colour.

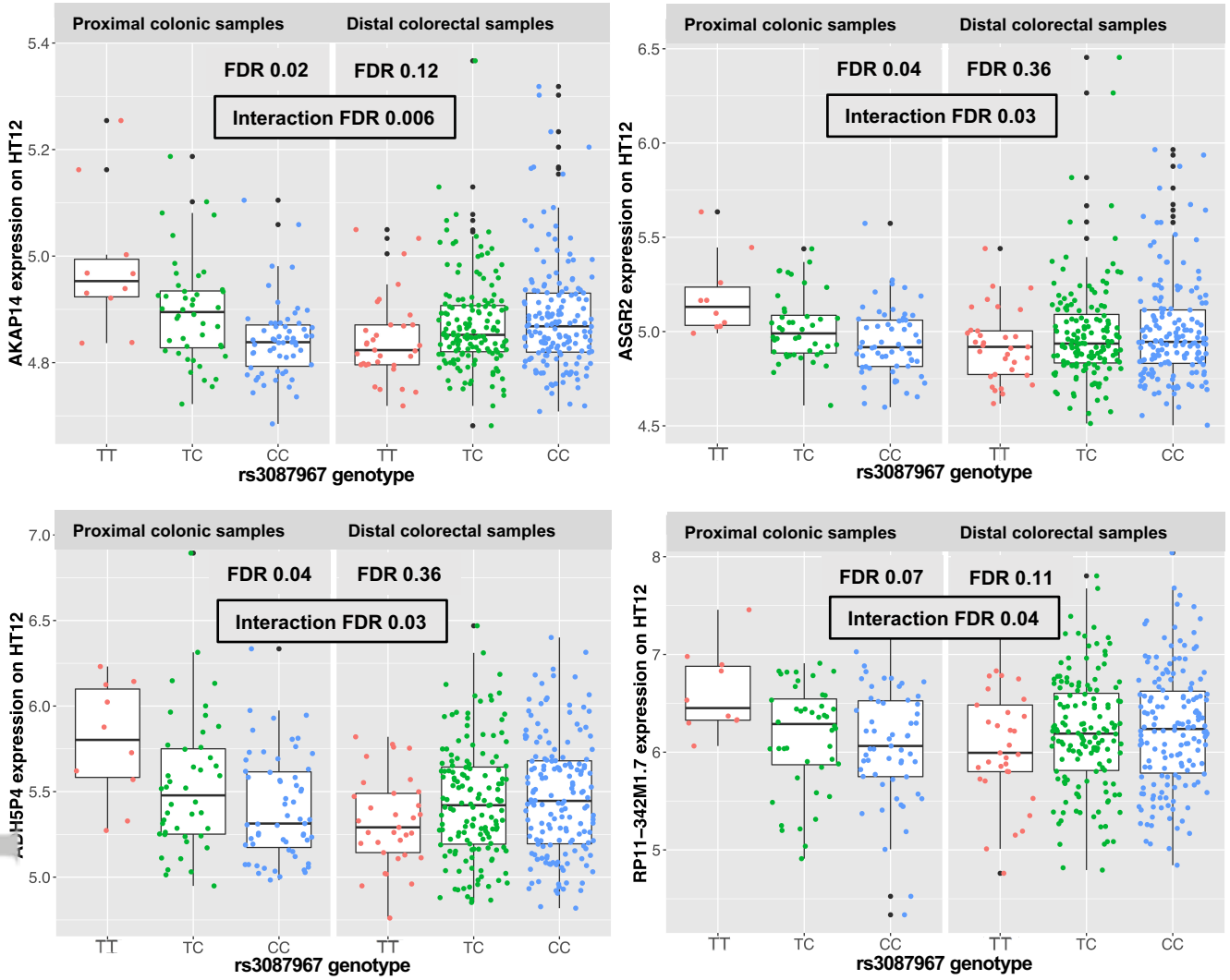


Figure 2. Differential trans-eQTL effects with rs3087967 in proximal and distal colorectal mucosa

Normal mucosa was sampled from resected colorectal specimens or by rectal biopsy. RNA was extracted and gene expression assessed using HT12 microarrays. Expression of genes with differential eQTL effects with rs3087967 are charted by site and genotype. The lower and upper hinges correspond to the first and third quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge